



**HAL**  
open science

# pKa Calculations with the Polarizable Drude Force Field and Poisson–Boltzmann Solvation Model

Alexey Aleksandrov, Benoît Roux, Alexander D. Mackerell

## ► To cite this version:

Alexey Aleksandrov, Benoît Roux, Alexander D. Mackerell. pKa Calculations with the Polarizable Drude Force Field and Poisson–Boltzmann Solvation Model. *Journal of Chemical Theory and Computation*, 2020, 16 (7), pp.4655-4668. 10.1021/acs.jctc.0c00111 . hal-02869378

**HAL Id: hal-02869378**

**<https://hal.science/hal-02869378>**

Submitted on 23 Dec 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# **pK<sub>a</sub> Calculations with the Polarizable Drude Force Field and Poisson-Boltzmann Solvation Model**

**Alexey Aleksandrov<sup>1\*</sup>, Benoît Roux<sup>3</sup>, and Alexander D. MacKerell, Jr.<sup>2\*</sup>**

<sup>1</sup>Laboratoire d'Optique et Biosciences, Ecole Polytechnique, IP Paris, F-91128 Palaiseau, France

<sup>2</sup>Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, 20 Penn Street, Baltimore, Maryland 21201, USA

<sup>3</sup>Department of Biochemistry and Molecular Biology, Gordon Center for Integrative Science, 929 E57th Street, University of Chicago, Chicago, Illinois 60637, United States.

\*Corresponding authors: [alexey.aleksandrov@polytechnique.edu](mailto:alexey.aleksandrov@polytechnique.edu), [alex@outerbanks.umaryland.edu](mailto:alex@outerbanks.umaryland.edu)

*Running title: pK<sub>a</sub> Calculations with Drude Force Field and Poisson-Boltzmann Solvation Model*

*Keywords: pK<sub>a</sub> calculations, Drude force field, implicit solvent model, Poisson-Boltzmann continuum solvation model, Monte-Carlo simulation, electronic polarization, CHARMM*

1 **ABSTRACT**

2 Electronic polarization effects have been suggested to play an important role in proton binding to titratable  
3 residues in proteins. In this work, we describe a new computational method for  $pK_a$  calculations, using  
4 Monte Carlo (MC) simulations to sample protein protonation states with the Drude polarizable force field  
5 and Poisson-Boltzmann (PB) continuum electrostatic solvent model. While the most populated protonation  
6 states at the selected pH, corresponding to residues that are half-protonated at that pH, are sampled using  
7 the exact relative free energies computed with Drude particles optimized in the field of the PB implicit  
8 solvation model, we introduce an approximation for the protein polarization of low-populated protonation  
9 states to reduce the computational cost. The highly populated protonation states used to compute the  
10 polarization and  $pK_a$ 's are then iteratively improved until convergence. It is shown that for lysozyme, when  
11 considering 9 of the 18 titratable residues, the new method converged within two iterations with computed  
12  $pK_a$ 's differing only by 0.02 pH units from  $pK_a$ 's estimated with the exact approach. Application of the  
13 method to predict  $pK_a$ 's of 94 titratable sidechains in 8 proteins shows the Drude-PB model to produce  
14 physically more correct results as compared to the additive CHARMM36 (C36) force field (FF). With a  
15 dielectric constant of two assigned to the protein interior the Root Mean Square (RMS) deviation between  
16 computed and experimental  $pK_a$ 's is 2.07 and 3.19 pH units with the Drude and C36 models, respectively,  
17 and the RMS deviation using the Drude-PB model is relatively insensitive to the choice of the internal  
18 dielectric constant in contrast to the additive C36 model. At the higher internal dielectric constant of 20,  
19  $pK_a$ 's computed with the additive C36 model converge to the results obtained with the Drude polarizable  
20 force field, indicating the need to artificially overestimate electrostatic screening in a nonphysical way with  
21 the additive FF. In addition, inclusion of both *syn* and *anti* orientations of the proton in the neutral state of  
22 acidic groups is shown to yield improved agreement with experiment. The present work, which is the first  
23 example of the use of a polarizable model for the prediction of  $pK_a$ 's in proteins, shows that the use of a  
24 polarizable model represents a more physically correct model for the treatment of electrostatic contributions  
25 to  $pK_a$  shifts in proteins.

26

27

28

29

30

## 1 INTRODUCTION

2 Titratable sites are abundant in proteins<sup>1</sup> and play an essential role in the structure, function and  
3 stability.<sup>2</sup> Thus, it is essential to reliably predict proton dissociation constants,  $pK_a$ 's, and to understand  
4 factors that modulate them.<sup>3</sup> A large multitude of methods to predict proton binding affinities in proteins  
5 have been developed over the last decades.<sup>4</sup> However, the accurate prediction of  $pK_a$ 's of protein titratable  
6 sites is still a major challenge and an active area of research.<sup>4b</sup> Accurate  $pK_a$  prediction faces several  
7 challenges including the need to consider protein conformational changes associated with the changes in  
8 protonation states, solvent contributions and interactions between titratable sites, which depend on each  
9 particular configuration of bound protons. Also contributing is the complex electronic response of the  
10 heterogeneous protein/solvent environment to changes in protonation states.<sup>5</sup>

11 A number of  $pK_a$  prediction methods rely on continuum dielectric models to describe the solvent  
12 degrees of freedom.<sup>2a, 6</sup> In these methods, frequently the protein in solution is treated using the continuum  
13 dielectric approximation based on the Poisson or Poisson-Boltzmann (PB) model<sup>7</sup> or generalized Born (GB)  
14 model in the context of an additive force field, with the GB model having the advantage of being more  
15 computationally efficient.<sup>8</sup> Bashford and Karplus were first to develop and apply the PB model using  
16 detailed 3D structural information for  $pK_a$  calculations and taking into account interactions between  
17 titratable sites as defined by a particular arrangement of bound protons.<sup>9</sup>

18 The number of possible protonation states of the protein grows exponentially with the number of  
19 titratable sites. The exact calculation of all accessible protonation states is not feasible for proteins  
20 containing a large number of titratable residues and different approximations have been introduced to  
21 overcome this challenge.<sup>7b, 9-10</sup> The early method of Tanford & Roxby introduced an approximation in the  
22 energy function which effectively reduces an ensemble of protonation micro-states to one.<sup>10b</sup> In this method  
23 a titratable residue interacts with protonated and deprotonated forms of all other residues weighted based  
24 on their  $pK_a$ 's and the targeted pH value. However, it was shown that this approximation is inaccurate for  
25 strongly interacting sites.<sup>10b, 10c</sup> Later methods include different site-reduction methods<sup>9-10, 10c</sup> and hybrid  
26 methods.<sup>11</sup> With site-reduction methods, most of configurations of bound protons are eliminated, for  
27 example based on precalculated occupancies or distances between titratable sites.<sup>10a</sup> Arguably, a more  
28 precise method is to perform Monte-Carlo (MC) simulations since, in principle, all protonation states can  
29 be sampled.<sup>7b</sup> With additional approximations, MC methods can be used together with a limited protein  
30 flexibility, for example, allowing for discrete side-chain conformational sampling with a rigid protein  
31 backbone.<sup>7c, 8a</sup>

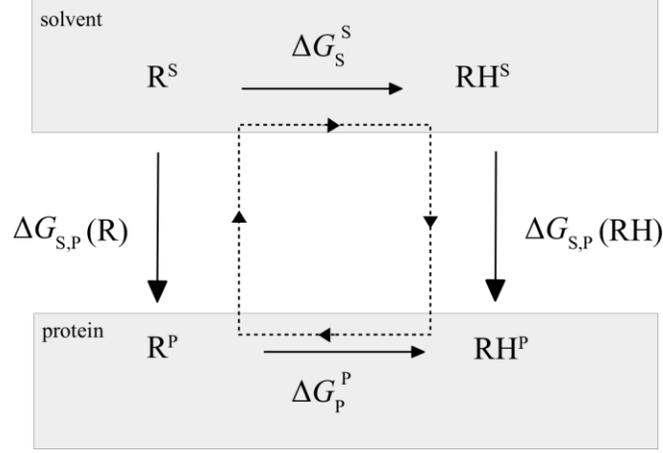
1 For computational efficiency, all these methods normally rely on the ability to decompose the free  
2 energy of the protein in a particular protonation state into energy contributions that depend only on the  
3 protonation states of individual residues or pairs of residues.<sup>10c</sup> This is possible as the field or potential  
4 determined by the Poisson equation is additive.<sup>6a</sup> The energy components can be precomputed and stored  
5 for subsequent free energy calculations performed during sampling of protonation states. However, with  
6 polarizable force fields the free energy cannot be represented in the pair-wise form, since the electronic  
7 state of the protein and, therefore, the free energy is defined by the protonation state of all titratable sites.  
8 To overcome this an effective approximation is needed to implement a polarizable model, such as the  
9 Drude-PB model, in constant-pH Monte Carlo simulations.

10 In this work, we present a new computational method to resolve the need to explicitly treat the  
11 polarization of a protein during  $pK_a$  calculations. While the calculation of  $pK_a$ 's for small molecules with a  
12 polarizable force field has been performed previously,<sup>12</sup> the present study represents their first application  
13 towards the estimation of  $pK_a$ 's in proteins. The approach is based on our previous study where we  
14 implemented and parametrized an implicit PB solvent model in conjunction with the Drude force field;  
15 similar work has been done with the AMOEBA polarizable force field.<sup>13</sup> In the new method, the most  
16 populated protonation states at the target pH, as defined by those residues that titrate in the region of the  
17 target pH, are sampled using the relative free energies that include a self-consistent field (SCF) calculation  
18 of the Drude particles in the field of the PB implicit solvation model. The states used to compute the  
19 electronic polarization and  $pK_a$ 's are iteratively improved until convergence. In addition, to facilitate the  
20 calculations, the interactions between titrating groups are calculated for a single electronic structure for  
21 each ionization state of each residue, with that approximation explicitly validated. The model was tested to  
22 predict the  $pK_a$ 's of 94 titratable sidechains in 8 proteins for which experimental  $pK_a$ 's are available.

## 24 METHODS

### 25 Classical electrostatic $pK_a$ calculations with additive force fields

26 The classical theory of  $pK_a$  calculations of a titratable residue group in the protein environment  
27 using the  $pK_a$  of the model compound in solvent is based on the thermodynamic cycle shown in Figure 1.



1

2 **Figure 1.** Thermodynamic cycle for proton binding. RH and R represent protonated and deprotonated  
 3 forms of the residue, respectively, in the solvent environment as a model compound (upper) or in the  
 4 protein environment (lower). The superscripts are used to highlight that the polarization of residue R/RH  
 5 is different in the protein and solvent. With the additive force fields these polarizations are the same.

6 It is assumed that the proton binding affinity difference of a titratable residue in the protein and a model  
 7 compound in solvent is only due to the electrostatic interactions. For a protein containing one titratable  
 8 residue:

9 
$$\text{p}K_a^{\text{protein}} = \text{p}K_a^{\text{model}} + \Delta\Delta G / \ln(10) / RT, \quad [\text{Eq 1}]$$

10 where  $\text{p}K_a^{\text{model}}$  is the  $\text{p}K_a$  of a model compound in solvent;  $R$  is the gas constant;  $T$  the temperature and  
 11  $\Delta\Delta G$  is a double difference of the electrostatic free energy associated with the residue being in the protein  
 12 environment. It is further assumed that the electrostatic field is governed by the macroscopic Poisson (or  
 13 Poisson-Boltzmann) equation:

14 
$$\nabla\epsilon(\vec{r})\nabla\varphi(\vec{r}) = -4\pi\rho(\vec{r}), \quad [\text{Eq 2}]$$

15 where  $\varphi$  is the electrostatic potential,  $\rho$  is the charge density and  $\epsilon$  is the dielectric constant. This equation  
 16 can be numerically solved, for example on a cubic lattice by finite difference methods, to give the charging  
 17 free energy,  $W$ , of a set of protein atomic charges:

18 
$$W = \frac{1}{2} \sum_i Q_i^P \varphi(\vec{r}_i), \quad [\text{Eq 3}]$$

19 where the summation is done over the protein atomic charges,  $Q_i^P$ ;  $\varphi(\vec{r}_i)$  is the electrostatic potential that  
 20 satisfies Equation 2 and computed at the position  $\vec{r}_i$  of the atomic charge  $Q_i^P$ .

21 For a macromolecule containing more than one titratable site, the protonation state of a residue,  $\mu$ ,  
 22 is affected by the charge state of all other titratable residues. In this case, the fraction of molecules,  $\theta_\mu$ ,

1 protonated at site  $\mu$  at a particular pH value is given by the Boltzmann average of all microstates where this  
 2 residue is protonated:

$$3 \quad \langle \theta_\mu \rangle = (\sum_{\{\bar{x}\}} x_{i,\mu} \exp(-\Delta G(\bar{x}_i, \text{pH})/RT)) / (\sum_{\{\bar{x}\}} \exp(-\Delta G(\bar{x}_i, \text{pH})/RT)), \quad [\text{Eq 4}]$$

4 where the summation is done over all possible protonation microstates  $\{\bar{x}\}$ ;  $\bar{x}_i$  is a vector that defines  
 5 protonation microstate  $i$ ;  $x_{i,\mu}$  is a  $\mu$ -th element of the vector  $\bar{x}_i$  and is 1 or 0 if residue  $\mu$  is protonated or  
 6 deprotonated, respectively, in the microstate  $i$ ;  $\Delta G(\bar{x}_i, \text{pH})$  is the relative free energy of protonation of  
 7 microstate  $\bar{x}_i$ , and within the context of additive force fields can be expressed as follows:

$$8 \quad \Delta G(\bar{x}_i, \text{pH}) = E(\bar{x}_i, \text{pH}) + \sum_\mu (\Delta G_{\text{Born},\mu}(x_{i,\mu}) + \Delta G_{\text{back},\mu}(x_{i,\mu})) + \frac{1}{2} \sum_{\mu \neq \nu} W_{\mu\nu}(x_{i,\mu}, x_{i,\nu}), [\text{Eq 5}]$$

9 where  $\Delta G_{\text{Born},\mu}$  is the relative Born energy of a titratable residue located in the protein environment and  
 10 related to its desolvation electrostatic free energy;  $\Delta G_{\text{back},\mu}$  is due to interactions with the background  
 11 charges on non-titratable residues;  $W_{\mu\nu}(x_{i,\mu}, x_{i,\nu})$  is electrostatic interaction energy between two titratable  
 12 residues  $\mu$  and  $\nu$  being in protonation states  $x_{i,\mu}$  and  $x_{i,\nu}$  respectively.  $E(\bar{x}_i, \text{pH})$  is a contribution from  
 13 solvent pH and reference model compounds:

$$14 \quad E(\bar{x}_i, \text{pH}) = \sum_\mu E(x_{i,\mu}, \text{pH}) = \sum_\mu (x_{i,\mu} RT \ln(10) (\text{pH} - \text{pK}_{a,\mu}^{\text{model}}) - \langle E_\mu^{\text{model}}(x_{i,\mu}) \rangle), [\text{Eq 6}]$$

15 where  $\langle E_\mu^{\text{model}}(x_{i,\mu}) \rangle$  is the average electrostatic free energy of the reference model compound for residue  
 16  $\mu$  being in protonation form  $x_{i,\mu}$  in solvent computed using the same force field model. For the convention,  
 17 in summations we will use letters from the Latin alphabet to designate protein particles (atoms, Drudes,  
 18 lone-pairs) and protein microstates, while Greek letters to denote residues in the protein. The  $\langle \theta_\mu \rangle$  are  
 19 evaluated at a discrete number of pH values to obtain a titration curve for site  $\mu$ .  $\text{pK}_{a,\mu}$  of a titratable residue  
 20  $\mu$  in the protein is then defined as the pH value where the titratable residue is half-protonated.

21 In practice calculations of titration curves directly using Equation 4 are limited to macromolecules  
 22 containing only a few titratable residues since it requires sampling of a large number of protonation  
 23 microstates that grows exponentially ( $2^N$ ) with the number of titratable residues. To solve this problem MC  
 24 simulations are performed to sample only relevant protonation states, while high-energy states that do not  
 25 contribute significantly in Equation 4 are not visited. To perform MC simulations, energies appearing in  
 26 Equation 5 must be precomputed and stored in the first step. Relative free energy of the protein in a  
 27 particular protonation state is then recovered from the energy matrices as a simple sum of energy terms in  
 28 the MC simulations.

## 1 **The Poisson-Boltzmann method for $pK_a$ calculations with a polarizable force field and multiple** 2 **titratable sites**

3 In the case of polarizable force fields,  $\Delta G_{\text{Born},\mu}$ ,  $\Delta G_{\text{back},\mu}$  and  $W_{\mu\nu}$  in Equation 5 depend on the  
4 electronic state, or polarization, of all protein atoms. In particular, with the Drude force field  $\Delta G_{\text{Born},\mu}$ ,  
5  $\Delta G_{\text{back},\mu}$  and  $W_{\mu\nu}$  are functions of the position of the Drudes on all atoms including titratable residues. In  
6 turn, the positions of all Drudes, including on protein backbone atoms, depend on the protonation states of  
7 all residues. In the case of polarizable force fields the relative free energy  $\Delta G(\bar{x}, \text{pH})$  contains additional  
8 contributions. In the context of the additive force field, these contributions do not depend on the protein  
9 protonation state  $\bar{x}$ , and thus do not contribute in Equation 5. These energy terms include (i) a contribution  
10 from interactions between background charges with background charges, since polarization of background  
11 atoms depends on the protonation state; (ii) the Born energy of background atoms, which now depends on  
12 the polarization affected by the protonation state of all residues; and (iii) the polarization work needed to  
13 polarize titratable and non-titratable groups of atoms from the polarization in solvent to the polarization in  
14 a protein. We will use  $G_{\text{BB}}(\bar{x})$  to denote the sum of the first two terms (i) and (ii), and the term (iii) will be  
15 included in  $G_{\text{BB}}(\bar{x})$ ,  $G_{\text{Born},\mu}(\bar{x})$ , and  $G_{\text{back},\mu}(\bar{x})$ . The term (iii) is computed within the Drude force field as  
16 the bond energy contributed by the atomic core-Drude particle bonds (i.e. self-polarization energy term or  
17 polarization work), which is different due to the different polarization in solvent and protein as well as  
18 being coupled to the protein protonation state. Thus, the total relative free energy of a microstate within the  
19 Drude polarizable force field is calculated using the following formula:

$$20 \Delta G(\bar{x}, \text{pH}) = E(\bar{x}, \text{pH}) + \Delta G_{\text{BB}}(\bar{x}) + \sum_{\mu} (\Delta G_{\text{Born},\mu}(x_{\mu}, \bar{x}) + \Delta G_{\text{back},\mu}(x_{\mu}, \bar{x})) + \frac{1}{2} \sum_{\mu \neq \nu} W_{\mu\nu}(x_{\mu}, x_{\nu}, \bar{x}),$$

21 [Eq 7]

22 where  $\bar{x}$  is, as above, a vector with element  $x_{\mu}$  defining the protonation state of residue  $\mu$ ; and the argument  
23  $\bar{x}$  in functions  $G_{\text{Born},\mu}(x_{\mu}, \bar{x})$ ,  $G_{\text{back},\mu}(x_{\mu}, \bar{x})$  and  $W_{\mu\nu}(x_{\mu}, x_{\nu}, \bar{x})$  is repeated to emphasize that in contrast  
24 to Equation 5, these terms depend on the protonation state of all residues including titratable residues  $\mu$  and  
25  $\nu$ .

26 In contrast to additive force fields,  $G(\bar{x}, \text{pH})$  given by Equation 7 is not a residue-pairwise function.  
27 This means that the free energy of all protein protonation microstates cannot readily be recovered in MC  
28 simulations. Accordingly, in what follows, we present an approximate MC method suitable for the  
29 polarizable Drude force field in the context of a constant pH formalism. We first note that to define  $pK_{a,1/2}$   
30 of a titratable residue only the point on the titration curve where  $\text{pH} = pK_{a,1/2}$  needs to be identified. Thus,  
31 the approach just needs to reproduce exactly the free energies of microstates highly populated at

1  $\text{pH} \sim \text{p}K_{a,1/2}$  that contribute significantly in Equation 4. In the method presented later in this section, free  
2 energies of the most populated states for the protonated and deprotonated forms of a residue are computed  
3 exactly using minimization of the position of the Drude particles (i.e. performing the polarization SCF  
4 calculation) in the field of the implicit solvent. Thus, polarization effects for the most populated microstates  
5 are taken into account exactly, while free energies of less populated microstates perturbed by the  
6 polarization response to the change of the protonation state are computed less accurately during the MC  
7 simulation. To calculate  $\Delta G_{\text{BB}}(\bar{x})$ ,  $\Delta G_{\text{Born},\mu}(x_{\mu}, \bar{x})$ ,  $\Delta G_{\text{back},\mu}(x_{\mu}, \bar{x})$  and  $W_{\mu\nu}(x_{\mu}, x_{\nu}, \bar{x})$  in Equation 7 the  
8 position of all Drude particles should be defined. In the method, the highly populated protonation states at  
9  $\text{pH} = \text{p}K_{a,\mu}$  are used to calculate these energies for the protonated and deprotonated forms of residue  $\mu$ .

### 10 **pK<sub>a</sub> calculations with the Drude force field and Poisson-Boltzmann model**

11 In this section the calculation protocol of the new method is given. A flow chart of the computational  
12 protocol is presented in Scheme 1. Protonation states for all residues are predefined in the initial calculation  
13 of energy terms appearing in Equation 7, with titratable residues assigned neutral protonation states. These  
14 predefined states will be refined iteratively in subsequent steps. The method starts with molecular  
15 mechanics (MM) and Poisson-Boltzmann calculations of free energies needed to perform MC simulations:

16 Step 1. Calculate protein free energies for both ionization states of all titratable residue with the  
17 remaining titratable residues assigned neutral protonation states. For each protonation state of titratable  
18 residue  $\mu$ , neutral protonation states are used for all other titratable residues giving the vector defining the  
19 protonation microstate  $\bar{x}_i$ . These protonation microstates  $\bar{x}_i$  are used to optimize the Drude particles. The  
20 free energies of the protein in each of these protonation microstates is calculated as  $G_i = G(\bar{x}_i)$ , based on  
21 the system MM energy and the PB implicit solvation energy, with these energies including the polarization  
22 energy following the Drude SCF calculation.

23 Step 2. Interaction free energies between titratable residues, which include MM electrostatic  
24 interactions and the solvent contribution, are calculated. This involves individually calculating the  
25 electrostatic potential for each titratable residue  $\mu$ , by zeroing the charges on all atoms in the protein  
26 (including lone pairs and Drude particles) except those on the residue  $\mu$ . The positions of Drude particles  
27 optimized in step 1 and corresponding to selected protein protonation microstates for residues  $\mu$  and  $\nu$  are  
28 used, so no optimization of Drude particles is needed at this step. To avoid the problem of artificial  
29 contributions arising when interaction energies are computed between neighboring residues due to 1,2 and  
30 1,3 dipole-dipole interactions included in the Drude model, the contribution to the interaction energy from  
31 solvent is computed using the PB model and combined with the MM energy to obtain the total interaction  
32 free energy between residues. The PB equation is solved to obtain the electrostatic potential  $\varphi_{R\mu}(\epsilon_{ext} =$

1  $\varepsilon_w, \varepsilon_{int} = \varepsilon_p$ ), due to the charges of residue  $\mu$  being in the protonation state  $x_\mu$ . Calculations are repeated  
2 using the protein dielectric constant for the protein exterior to obtain the electrostatic potential  $\varphi_{R\mu}(\varepsilon_{ext} =$   
3  $\varepsilon_p, \varepsilon_{int} = \varepsilon_p)$ . The electrostatic potential is used to calculate the electrostatic interaction  $W_{\mu\nu}^{x_\mu, x_\nu^*}$  between  
4 the titratable residues  $\mu$  and  $\nu$  being in protonation state  $x_\mu$  and  $x_\nu$ , respectively, according to  $W_{\mu\nu}^{x_\mu, x_\nu^*} =$   
5  $1/2 \sum_{ij} q_i q_j / \varepsilon_p / r_{ij} + \sum_j q_j (\varphi_{R\mu_j}(\varepsilon_{ext} = \varepsilon_w, \varepsilon_{int} = \varepsilon_p) - \varphi_{R\mu_j}(\varepsilon_{ext} = \varepsilon_p, \varepsilon_{int} = \varepsilon_p))$ , where  $q_i$  and  $q_j$   
6 are charges of residues  $\mu$  and  $\nu$ , respectively. Note that in principle  $W_{\mu\nu}^{x_\mu, x_\nu^*} \neq W_{\nu\mu}^{x_\nu, x_\mu^*}$ , and these  
7 interaction energies are different from those appearing in Equation 7 since the polarization used for residues  
8  $\mu$  and  $\nu$  corresponds to different protein protonation microstates. We use an asterisk to distinguish these  
9 energies from the interaction energies in Equation 7.

10 Step 3. For each free energy,  $G_i$  computed in step 1 it is possible to write Equation 7 as follows:

$$11 \quad G_i = G_{BB}(\bar{x}_i) + \sum_\mu \left( \Delta G_{\text{Born},\mu}(x_{i,\mu}, \bar{x}_i) + \Delta G_{\text{back},\mu}(x_{i,\mu}, \bar{x}_i) \right) + \frac{1}{2} \sum_{\mu \neq \nu} W_{\mu\nu}(x_{i,\mu}, x_{i,\nu}, \bar{x}_i), \text{ [Eq. 8]}$$

12 The latter expression does not form a closed system of linear equations relative to the terms  
13  $\Delta G_{\text{Born/back},\mu}(x_{i,\mu}, \bar{x}_i) = \Delta G_{\text{Born},\mu}(x_{i,\mu}, \bar{x}_i) + \Delta G_{\text{back},\mu}(x_{i,\mu}, \bar{x}_i)$ , since the latter terms are different for  
14 different protonation microstates  $\bar{x}_i$ . To recover  $G_i$  later in MC simulations, instead of using Equation 8 we  
15 introduce a system of linear equations:

$$16 \quad \sum_\mu G_{\text{Born/back},\mu}^{x_\mu^1} + G_{BB} = G_1 - \frac{1}{2} \sum_{\mu \neq \nu} W_{\mu\nu}^{x_\mu^1, x_\nu^{1*}}$$

$$18 \quad \sum_\mu G_{\text{Born/back},\mu}^{x_\mu^2} + G_{BB} = G_2 - \frac{1}{2} \sum_{\mu \neq \nu} W_{\mu\nu}^{x_\mu^2, x_\nu^{2*}}$$

17 ... [Eq 9]

19 where  $G_{BB}$  is again due to interactions between background atoms with themselves, but invariant relative  
20 to the protonation state of titratable residues;  $W_{\mu\nu}^{x_\mu^1, x_\nu^{1*}}$  is the interaction energy between residues  $\mu$  and  $\nu$   
21 computed in step 2;  $G_{\text{Born/back},\mu}^{x_\mu^i}$  and  $G_{BB}$  can be regarded as unknowns that satisfy the system of equations.  
22 The right hand expressions in the system are calculated in steps 1 and 2. The system of linear equations can  
23 be resolved to find all  $G_{\text{Born/back},\mu}^{x_\mu^i}$  and  $G_{BB}$ .

24 We note that  $G_{\text{Born/back},\mu}^{x_\mu^i}$  are not calculated directly in step 1 as was performed in the original constant-  
25 pH MC method. This is due to the need to calculate free energies from step 1 in the MC simulations as

1 required to identify the most likely protonation microstates for each titratable residue as a function of pH  
 2 when residues titrate (at  $\text{pH} = \text{p}K_{a,\mu}$ ) rather than  $G_{\text{Born/back},\mu}^{x_\mu^i}$  energies. In other words,  $G_1, G_2 \dots G_n$  are  
 3 used in MC simulations to sample probabilities of protonated and deprotonated states and, thus are required  
 4 to calculate the titration curves. It should be emphasized that in MC simulations with the Drude force field  
 5 it is prohibitively expensive to calculate the free energies of all protein microstates in contrast to the  
 6 calculations with additive force fields; instead, we recover free energies of the most important states using  
 7 the above method.

8 It may happen that the most likely protein microstates are identical for protonation states of different  
 9 residues at the pH where they are half-protonated. In this case, equations for the protonation states of these  
 10 residues are identical in the system of equations 9, and the system is not complete as required to define  
 11  $G_{\text{Born/back},\mu}^{x_\mu^i}$  and  $G_{\text{BB}}$ . To complete the system we introduce additional equations in the free energy  $G_l$   
 12 computed with zero charges on all titratable residues except residue  $\mu$ . The additional equation added to the  
 13 system of equations 9 is:  $G_{\text{Born/back},\mu}^{x_\mu^1} + G_{\text{BB}} = G_l$ .

14 Step 4. Perform MC simulations. During the MC simulations at the pH corresponding to the  $\text{p}K_{a,\theta}$  of  
 15 residue  $\theta$ , the free energy of microstates is computed according to:

$$16 \quad G(\bar{x}) = G_{\text{BB}} + \sum_{\mu} G_{\text{Born/back},\mu}^{x_\mu} + \frac{1}{2} \sum_{\mu \neq \nu} W_{\mu\nu}^{x_\mu, x_\nu^*} \text{ [Eq. 10]}$$

17 In Equation 10,  $W_{\mu\nu}^{x_\mu, x_\nu^*}$  are the same energies used in the system of equations 9 and  $G_{\text{Born/back},\mu}^{x_\mu}$  and  $G_{\text{BB}}$   
 18 are the solutions. For the most populated microstate  $\bar{x}_i$ , selected in Step 1, this equation should give exactly  
 19  $G_i$ . Thus, this approximation allows the free energies to be recovered in the MC simulations computed with  
 20 the correct polarization (*e.g.* SCF Drudes). It should be noted that  $G_{\text{BB}}$  is a constant for all microstates and  
 21 thus, cancels out when relative free energies of microstates are computed in the MC simulations. The  
 22 dependence of  $G_{\text{BB}}(\bar{x})$  on the protonation state does not appear in Equation 10 explicitly. However, for the  
 23 most populated states it is included in  $G_{\text{Born/back},\mu}^{x_\mu}$ , as they are solutions of the system of equations 9.

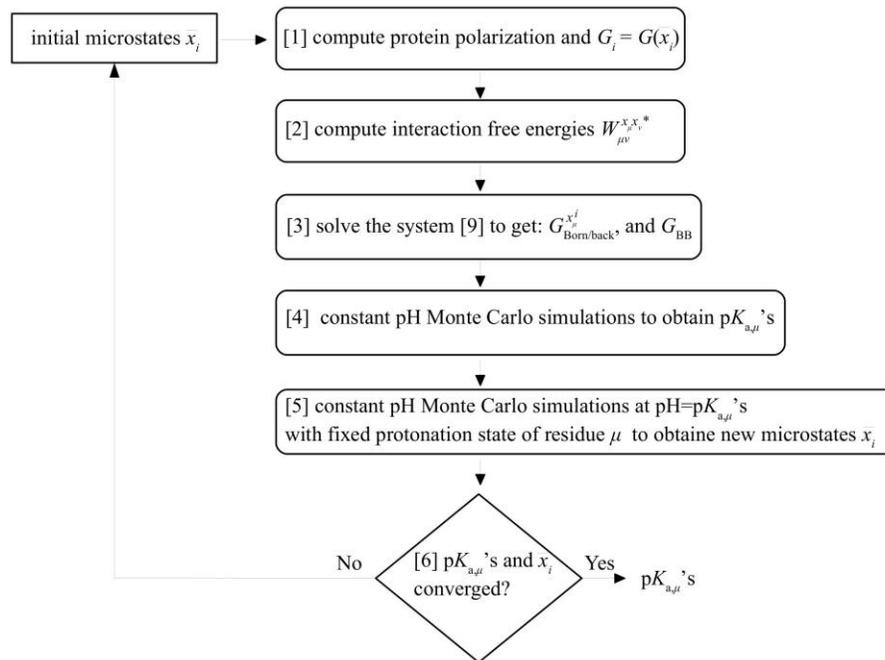
24 MC simulations are performed in the range of pH values between -10 to 30 with a step of 0.5 pH unit  
 25 to obtain a titration curve for each titratable residue. The contribution  $E(\bar{x}, \text{pH})$  computed by Equation 6 is  
 26 added to Equation 10 to obtain relative free energies of protein microstates. During the MC simulations one  
 27 randomly selected titratable residue protonation state is changed with acceptance or rejection of that change  
 28 based on the Metropolis criteria. In 50% of the MC steps a second residue is allowed to change its  
 29 protonation state. In the present study, 100,000 MC steps were performed for each titratable residue in the

1 system (eg. with 20 titratable residues  $2 \cdot 10^6$  MC steps are performed). To test the convergence of MC  
 2 simulations the number of MC steps was doubled, but the change in relative populations of protonated and  
 3 deprotonated forms was less than  $10^{-3}$  observed for residues in eight proteins. Finally, using the titration  
 4 curves the set of  $pK_{a,\mu}$  values of all titratable residues can be defined based on the pH at which they are  
 5 half-protonated.

6 Step 5. MC simulations for each titratable residue  $\mu$  and each of its protonation state  $x_\mu$  are repeated at  
 7  $pH = pK_{a,\mu}$  determined in the previous step. In contrast to the MC simulations in step 4, the targeted  
 8 titratable residue  $\mu$  is fixed in the protonation state  $x_\mu$  to find the most likely protonation states for all other  
 9 titratable residues. Note that the most likely protonation states may be different for the protonated and  
 10 deprotonated forms of the same residue  $\mu$ . The same number of MC steps was performed as in step 4.

11 Step 6. Steps 1-5 are repeated with the most likely states of each titratable residue obtained from step  
 12 5. These iterations are required since initially in step 1 the most likely protonation states are not known but  
 13 rather estimated based the neutral protonation state. Iterations over steps 1-5 are performed until the  
 14 calculated  $pK_{a,\mu}$  of all the titratable residues and the states computed in step 5 converge. Overall, the  
 15 protocol has two types of self-consistent iterations: (i) in step 1 the position of the Drudes and the PB  
 16 solvent polarization are fully optimized and (ii) globally, steps 1-6 are repeated to converge the individual  
 17 titratable residue  $pK_{a,\mu}$  values.

18



19

1 **Scheme 1.** Flow Chart of the computations performed with the Drude-PB method. Steps 1-5 are repeated until  $pK_{a,\mu}$   
2 and microstates converge. Initial microstates are updated using the computed microstates at the end of the previous  
3 iteration.

4 To summarize, using this method the polarization effects are included without any approximation in  
5 free energies for the most populated protonation microstates of a protein when residues titrate (at  $pH =$   
6  $pK_{a,\mu}$ ). Within this method, it is achieved at an additional computational cost to perform multiple iterations.  
7 It should be noted that polarization of less populated states is still incorrectly treated, since a surrogate of  
8  $G_{\text{Born/back},\mu}^{x_\mu}$  and  $W_{\mu\nu}^{x_\mu,x_\nu^*}$  corresponding to protonation states that differ from that of the less populated  
9 states is used. The latter error is expected to be small, since those microstates make small contributions to  
10 the titration curves at  $pH$  equal  $pK_{a,\mu}$ . Notice, that in principle, one could consider exact free energies for a  
11 limited number of less occupied microstates in Equation 4, however, in this work we limit to one state per  
12 protonation and rotameric state of a residue.

### 13 **Proton binding sites and protein structure relaxation**

14 In the present study, only titratable protons are allowed to change their positions to preserve the  
15 dielectric boundary. Otherwise, the PB equation would need to be solved for each  $W_{\mu\nu}$  element, which is  
16 prohibitively expensive. It should be noted that different approximations have been proposed with the  
17 sacrifice of the exact protein boundary to allow limited flexibility of sidechains<sup>8a, 14</sup>, which will be explored  
18 in future studies with the Drude force field.

19 In the case of the acidic aspartate and glutamate residues, we consider five protonation states: one  
20 ionized negative state and four neutral states with the proton on either oxygen and in the *syn* and *anti*  
21 orientations. Two rotamers were included for neutral tyrosine that differ by the orientation of the hydroxyl  
22 group, and three rotamers for the neutral lysine, distinguished by the dissociation of amino protons.  
23 Histidines had two possible neutral tautomers: protonated on  $N\epsilon$  ( $pK_a$  of the model compound 7.0) and  $N\delta$   
24 ( $pK_a$  of the model compound 6.5). In the implementation, the neutral tautomers of histidines are simply  
25 treated as "rotamers" with a different contribution to the  $pH$  dependent term due to the  $pK_a$  difference of  
26 the  $N\epsilon$  and  $N\delta$  sites. The total number of rotamers for neutral and ionized forms for titratable residues was  
27 chosen to be identical to avoid the problem of artificial biasing in MC simulations of protonation forms  
28 having a larger number of rotamers.

### 29 **Reference state**

30 Following the thermodynamic cycle shown in Figure 1, to calculate the protonation free energy in  
31 the protein the free energy of the model compound in solution, called the reference free energy, is  
32 subtracted. This free energy is estimated using the same force field model, which is needed for the  
33 cancelation of artefacts due to the employment of the empirical force field model. The force field term of

1 the reference free energy is estimated as the free energy of the model compound in solution averaged over  
2 all possible compound conformations. In this work, we neglect the contribution from the bonded terms not  
3 associated with the Drude particles, since a single conformation for the protein calculations is used. Thus,  
4 the reference free energy of a model compound with a titratable residue  $x$  in solvent is:

$$5 \quad G_x^{\text{ref}} = E_{\text{elec}} + E_{\text{bond}}^{\text{Drude}} + G_{\text{solv}}^{\text{PB}}, \text{ [Eq 11]}$$

6 where  $E_{\text{elec}}$  is the intramolecular electrostatic energy computed with the same dielectric constant  $\epsilon_p$ , which  
7 is used to calculate the solvation free energy  
8  $G_{\text{solv}}^{\text{PB}} = G_{\text{ext}=80}^{\text{PB}} - G_{\text{ext}=\epsilon_p}^{\text{PB}}$ . The same dielectric constant is also used for the protein calculations.  $E_{\text{bond}}^{\text{Drude}}$   
9 is the bond energy from the atomic core-Drude particle bonds (i.e. self-polarization energy term or  
10 polarization work).<sup>15</sup> N-acetyl- $x$ -N-methylamide with the corresponding titratable residue  $x$  was used as the  
11 model compound in solution. In this compound, charges involved in all 1-4 electrostatic interactions,  
12 including Drudes are identical to those charges in the protein system, leading to the cancelation of artefacts  
13 arising from the employment of the force field. To obtain  $\text{p}K_a$  's in the protein, the computed  $\text{p}K_a$  shifts due  
14 to the protein environment were added to  $\text{p}K_a^{\text{model}}$ 's given in Table S2. The experimental  $\text{p}K_a$  shifts were  
15 computed as the difference between the  $\text{p}K_a$  in the protein environment and the  $\text{p}K_a$  of the corresponding  
16 model compound.

17 To obtain average free energies in solvent we performed molecular dynamics (MD) simulations of  
18 the N-acetyl- $x$ -N-methylamides immersed in a cubic solvent box. The minimum distance between the  
19 compound atoms and the edge of the system was 12 Å. Periodic boundary conditions were assumed. All  
20 long range electrostatic interactions were computed efficiently by the particle mesh Ewald method<sup>16</sup> using  
21 a real space cutoff of 12 Å. The Lennard-Jones term was evaluated out to 12 Å with a force switch  
22 smoothing function from 10 to 12 Å. MD simulations were performed at a constant temperature of 298 K  
23 and pressure of 1 ATM after 20 ps of thermalization. During MD simulations the center of mass of the  
24 model compound atoms was weakly harmonically restrained to the origin of the system with a force  
25 constant of 1.0 kcal·mol<sup>-1</sup>·Å<sup>-2</sup>. For the model compounds the CHARMM36 (C36)<sup>17</sup> and Drude<sup>18</sup> protein  
26 force fields were used along with the CHARMM TIP3P<sup>19</sup> and SWM4-NDP<sup>20</sup> model for water for the  
27 additive and polarizable calculations, respectively. Simulations were done with the NAMD program.<sup>21</sup> 50  
28 nanoseconds of MD were performed at constant temperature and pressure for the compound containing  
29 each titratable residues. To calculate PB free energies, structures from the MD simulations were saved every  
30 100 ps. The final PB free energies were averaged over these structures. The convergence was confirmed by  
31 dividing the data into five blocks corresponding to 10 ns MD simulations and computing the standard  
32 deviation, which was lower than 0.1 kcal·mol<sup>-1</sup> in all cases.

1 For the protonated form of the carboxylic acids, Asp and Glu, the *syn* and *anti* positions of the OH  
 2 proton were simulated separately. The reference energy of the protonated form of Asp and Glu was  
 3 Boltzmann-averaged over the free energies of the two forms.

#### 4 **Internal dielectric constant**

5 As demonstrated and discussed in the work of Warshel et al, the dielectric constant ascribed to the  
 6 protein medium is meant to represent physical contributions that are not considered explicitly.<sup>22</sup> In the early  
 7 model of Tanford and Roxby a protein was treated as a medium with a dielectric constant  $\epsilon_{int} = 4$  and  
 8 solvent with a dielectric constant of 80, the experimental value. The protein dielectric constant of 4 is larger  
 9 than the electronic polarizability estimate of 2, presumably to take into account the contribution due to the  
 10 fluctuations of protein polar groups about their equilibrium positions.<sup>9, 23</sup> In the model of Tanford and  
 11 Roxby, the uniform continuum medium representing the interior of the protein, itself treated as a fixed  
 12 object, was meant to implicitly incorporate the effects of the atomic fluctuations. This model is clearly an  
 13 approximation. Obviously, the choice of the dielectric constant ascribed to the protein interior depends on  
 14 the physical effects that are treated explicitly in the model.<sup>10a, 24</sup> In this work, we do not treat fluctuations of  
 15 protein atoms explicitly, which justifies the use of a higher dielectric constant for the protein interior ( $\epsilon_{int} >$   
 16 1). However, since reorganizations in the protein electronic structure are treated explicitly in the polarizable  
 17 model, the protein dielectric constant is expected to be smaller than in the model with the additive force  
 18 field. This conjecture will be verified with practical examples below. Following our previous work, the  
 19 ionic strength was set to 0 M.<sup>15</sup>

#### 20 **Poisson-Boltzmann free energy calculations with the Drude Force field**

21 The Poisson-Boltzmann free energy with the Drude force field is calculated in accord with our  
 22 previous work.<sup>15</sup> In brief, we need to calculate the electrostatic free energy,  $G_{\epsilon_{ext}=\epsilon_w, \epsilon_{int}=\epsilon_p}$  of a solute with  
 23 an internal dielectric constant of  $\epsilon_p$  immersed in a dielectric medium with a high dielectric constant of  $\epsilon_w$ .  
 24 The free energies computed using the potential obtained by numerically solving the Poisson-Boltzmann  
 25 equation and Equation 3 contain the artificial contributions of the grid as well as from electrostatic  
 26 interactions between 1-2 and 1-3 bonded atoms. These contributions in the PB model should be removed  
 27 by subtraction. To correct the electrostatic component of the free energy we modify  $G_{\epsilon_{ext}=\epsilon_w, \epsilon_{int}=\epsilon_p}$  by the  
 28 free energy computed with a uniform dielectric constant of  $\epsilon_p$ :

$$29 \quad G_{\epsilon_{ext}=\epsilon_w, \epsilon_{int}=\epsilon_p} = G_{\epsilon_{ext}=\epsilon_w, \epsilon_{int}=\epsilon_p} - G_{\epsilon_{ext}=\epsilon_p, \epsilon_{int}=\epsilon_p} + G_{\epsilon_{ext}=\epsilon_p, \epsilon_{int}=\epsilon_p}, \text{ [Eq 12]}$$

30 where  $G_{\epsilon_{ext}=\epsilon_p, \epsilon_{int}=\epsilon_p}$  is the contribution from the solute-solute interactions in a uniform dielectric medium  
 31 with a dielectric constant of  $\epsilon_p$  and is computed using  $G_{\epsilon_{ext}=\epsilon_p, \epsilon_{int}=\epsilon_p} = \frac{1}{2} \sum_{i \neq j} \frac{q_i q_j}{\epsilon_p r_{ij}}$ . The first two terms

1 are computed using the Poisson-Boltzmann equation using the same set of parameters including those that  
2 define the grid, except the external dielectric constant. In this case, the artificial contributions cancel out,  
3 since the internal dielectric constant in both calculations is the same. In these calculations the state with the  
4 uniform dielectric constant,  $\epsilon_p$ , is used as a reference state. To obtain the total free energy of a solute, the  
5 electrostatic component given by Equation 12 should be supplemented by self-polarization work, which is  
6 computed within the Drude force field as the bond energy contributed by the atomic core-Drude particle  
7 bonds.

8 An additional complication with a polarizable force field is that the interaction energy  
9  $W_{\mu\nu}(x_{i,\mu}, x_{i,\nu})$  in Equation 7 includes the electronic energy of the entire system that includes the self-  
10 polarization energy and the 1-2, 1-3 contributions from Drude particles. These terms disallow the  
11 calculation of  $W_{\mu\nu}(x_{i,\mu}, x_{i,\nu})$  for two neighboring residues using only the Poisson-Boltzmann model. This  
12 is not the case for additive force fields where charges on the backbone atoms are normally fixed to the same  
13 values in the protonated and deprotonated forms, and thus these contributions cancel out for neighboring  
14 residues when the protonation free energy is computed. Thus, for the Drude force field the combination of  
15 the MM energy and PB solvation free energy are used to calculate the interaction energy,  $W_{\mu\nu}(x_{i,\mu}, x_{i,\nu})$ ,  
16 as described above.

17 We use the solvation radii that were optimized in our previous work to reproduce experimental  
18 solvation free energies of a set of small molecules.<sup>15</sup> The solvation radii were defined for all atom types  
19 except the deprotonated hydroxyl oxygen in tyrosine. The missing solvation radius of the O<sup>-</sup> oxygen was  
20 optimized to reproduce the experimental absolute solvation free energy of the deprotonated tyrosine as  
21 described in the Supplementary Information.

22 PB free energy calculations were performed with the PBEQ module<sup>25</sup> implemented in the CHARMM  
23 program.<sup>26</sup> To include polarization effects explicitly the positions of Drude particles were optimized with  
24 the nuclear positions constrained in each protein microstate in step 1 using 50 steps of the Steepest Descent  
25 minimizer. Previously we showed that 20 steps of optimization was adequate for the minimization  
26 convergence for a set of protein complexes.<sup>15</sup> As previously, dummy atoms were added to fill internal  
27 cavities not accessible by water molecules with a low dielectric medium.<sup>15</sup> The protein PB energies were  
28 computed using the focusing method with a coarse grid of 0.8 Å resolution and fine grid with 0.4 Å  
29 resolution. The ion concentration was set to zero; we continue to call this method PB for the sake of  
30 simplicity, but use the finite-difference Poisson equation with no electrolyte present in the continuum  
31 solvent. The program to perform Monte-Carlo simulation for pK<sub>a</sub> calculations was written in C++. The  
32 system of linear equation 9 was solved using the Eigen library for linear algebra.<sup>27</sup>

## 1 **Protein data set for pK<sub>a</sub> calculations**

2 The data set includes 94 titratable residues from eight proteins (Table S1, Supporting information).  
3 Protein structures were retrieved from the Protein Data Bank (PDB) and used for the position of heavy  
4 atoms in all calculations. Hydrogens were built using CHARMM,<sup>26</sup> and optimized with a uniform dielectric  
5 constant of 4 and titratable residues set to the standard protonation states at pH 6.5 (carboxylic acids  
6 deprotonated; lysines and tyrosines protonated; histidines doubly protonated). In this work we consider  
7 Asp, Glu, His, Lys, and Tyr as titratable, while Arg residues were present only in the protonated form. The  
8 protein data set did not contain any titratable cysteines. The N- and C-termini were not considered as  
9 titratable and were fixed in the standard protonation state, i.e. the terminal amino group is protonated and  
10 terminal carboxylate group is deprotonated. Thus, the data set included 31 aspartic acids, 30 glutamic acids,  
11 10 tyrosines, 17 lysines and 6 histidines. Most of the experimental pK<sub>a</sub> values used in this study were  
12 compiled by Georgescu et al.<sup>14</sup> The experimental pK<sub>a</sub>'s for the SNase variant Δ+PHS were taken from  
13 Castaneda et al.<sup>3</sup>

## 14 **RESULTS**

### 15 **Polarization effect on interaction free energies between titratable residues**

16 We first examine the effect of polarization due to protonation of protein titratable sites on  
17 interaction free energies,  $W_{\mu\nu}^{x_{\mu}x_{\nu}}$  to test the approximation that these terms do not change significantly in  
18 the polarizable force field. Within classical additive force fields  $W_{\mu\nu}^{x_{\mu}x_{\nu}}$  are independent of protonation  
19 states of all residues except the protonation state  $x_{\mu}$  and  $x_{\nu}$  of the corresponding pair of residues  $\mu$  and  $\nu$ .  
20 With polarizable force fields, in principle  $W_{\mu\nu}^{x_{\mu}x_{\nu}}$  depends on the protonation state of all protein titratable  
21 sites:  $W_{\mu\nu}^{x_{\mu}x_{\nu}} = W_{\mu\nu}^{x_{\mu}x_{\nu}}(\bar{x})$ . To estimate the magnitude of this dependence we computed  $W_{\mu\nu}^{x_{\mu}x_{\nu}}$  for  
22 different pairs  $\mu$  and  $\nu$  in the eight proteins from the data set and random protein protonation states as  
23 follows. Random protonation states for each of the proteins were generated with the number of the  
24 generated random protonation states proportional to the number of titratable residues. The positions of the  
25 Drude particles were then fully optimized for each of these protonation states using the PB implicit solvent  
26 model for the complete protein structures. For these calculations, the dielectric constant of two was used  
27 for the protein interior. The interaction free energies,  $W_{\mu\nu}^{x_{\mu}x_{\nu}}$ , were then calculated yielding around 20  
28 values for each  $W_{\mu\nu}^{x_{\mu}x_{\nu}}$  interaction energy when all the randomly generated models were considered. These  
29 interaction free energies for a pair of residues are different due to the protonation states of other residues  
30 through induced polarization. Table 1 gives statistics of computed interactions. The average absolute  
31 difference in the interaction free energy over all pairs of titratable residues is just  $5 \cdot 10^{-4}$  kcal·mol<sup>-1</sup> for the

1 protein 1a2p, and values of a similar magnitude were found for the other proteins in the data set. The  
 2 maximum absolute difference in  $W_{\mu\nu}^{x_\mu, x_\nu}$  due to the protein protonation state is less than or equal to 0.15  
 3 kcal·mol<sup>-1</sup> for all protein except SNase variant Δ+PHS (PDB reference code 3bdc) and ribonuclease A  
 4 (PDB reference code 3rn3). In SNase the large effect on the interaction is observed for the pair Tyr91-  
 5 Glu75. This is explained by the fact that these residues directly interact with other titratable residues: Tyr91  
 6 makes a hydrogen bond with Asp77, and Glu 75 interacts with Tyr93 and His121. Deprotonation of these  
 7 residues has a strong effect on the polarization of Tyr91 or Glu75 due to strong and unfavorable electrostatic  
 8 interactions. In fact, we expect this effect to be smaller if the protein flexibility is taken into account and  
 9 these pairs are allowed to rearrange upon titration. The maximum variation in  $W_{\mu\nu}^{x_\mu, x_\nu}$  in SNase excluding  
 10 this pair is less than 0.1 kcal·mol<sup>-1</sup>. Overall, we find that the effect of the induced polarization on interactions  
 11 between ionizable residues due to the protein protonation state to be negligible for the eight proteins in the  
 12 data set thereby allowing this term to be calculated based on a single protonation state of the system.

13 **Table 1.** Absolute difference in the interaction free energies due to randomly-generated variations in the  
 14 protein protonation state. Calculations used the protein dielectric constant of two and the Drude force  
 15 field. Energies are given in kcal·mol<sup>-1</sup>.

Protein	Abs. difference	
	Max	Average
1a2p	0.15	0.0005
1pga	0.06	0.0006
1ppf	0.02	0.0004
2lzt	0.02	0.0001
2trx	0.05	0.0003
3bdc	0.34	0.0009
3rn3	0.24	0.0003
4pti	0.01	0.0001

16

### 17 **Contribution of the polarization on background atoms induced by titration**

18 Next the polarization effect of background atoms due to changes in protonation state of titratable  
 19 residues on computed pK<sub>a</sub>'s was examined. This polarization contributes directly to interactions between  
 20 titratable residues and background atoms, i.e. to the term  $G_{\text{Born/back},\mu}^{x_\mu}$ , as well as changes the interactions  
 21 of background atoms with themselves  $G_{\text{BB}}(\bar{x})$ . To test if  $G_{\text{BB}}(\bar{x})$  can significantly influence the population  
 22 of the protonated versus deprotonated forms of a titratable residue we computed  $G_{\text{BB}}(\bar{x})$  with different  
 23 protonation states of the protein as follows. First, the most likely protein protonation state  $\bar{x}$  was computed  
 24 at the pH where a titratable residue is half protonated with a protein dielectric constant of 4.  $G_{\text{BB}}(\bar{x})$  were

1 then computed for all residues from the data set and all possible protonation states with the correct  
 2 polarization, i.e. the polarization computed in the first step presented in the Methods section. The results  
 3 are given in Table 2. As can be seen  $G_{\text{BB}}(\bar{x})$  depends on the protonation state of titratable residues only  
 4 moderately. For all studied proteins, the average values of  $G_{\text{BB}}(\bar{x})$  are close to those obtained through  
 5 solution of the system of equations 9. For example, for lysozyme (PDB 1a2p), the standard deviation of  
 6  $G_{\text{BB}}(\bar{x})$  due to residue protonation states is just 0.3 kcal·mol<sup>-1</sup>. Further analysis demonstrated that the largest  
 7 variations in  $G_{\text{BB}}(\bar{x})$  are associated with either interactions with arginines treated as background non-  
 8 titratable atoms in the present work or very unfavorable interactions with the background atoms, explained  
 9 by the fact that no explicit relaxation is taken into account. Thus, the results in Table 2 indicate that the  
 10 polarization of the background charges induced by titration can be neglected in the calculation of  $G_{\text{BB}}(\bar{x})$   
 11 for pK<sub>a</sub> calculations thereby avoiding recalculation of this term for all protonation states.

12 **Table 2.** Average contribution of background charges,  $G_{\text{BB}}(\bar{x})$ , to the calculated total free energy  
 13 (kcal·mol<sup>-1</sup>).

protein	<sup>A</sup> exact $G_{\text{BB}}$	<sup>B</sup> $G_{\text{BB}}^{\text{sol}}$
1a2p	-467.3 (0.3)	-467.3
1pga	-29.6 (0.2)	-29.5
1ppf	-156.8 (0.1)	-156.7
2lzt	-1092.6 (0.3)	-1092.5
2trx	-117.8 (0.2)	-117.8
3bdc	-443.3 (0.3)	-443.3
3rn3	-524.1 (0.9)	-524.2
4pti	-465.1 (0.1)	-465.2

14 <sup>A</sup>The average value of the exact  $G_{\text{BB}}(\bar{x})$  computed for the most populated protonation states for each  
 15 titratable residue in the proteins; standard deviations are given in parenthesis; <sup>B</sup> $G_{\text{BB}}(\bar{x})$  obtained as a  
 16 solution to the system of equations 9.

## 17 pK<sub>a</sub> calculation with the Drude-PB model

### 18 Comparison to the exact solution

19 Initially, the method for pK<sub>a</sub> calculations with the Drude model was tested on a simple system with  
 20 fewer titration sites, for which the direct application of Equation 4 is still feasible. Lysozyme (PDB  
 21 reference code 2LZT) was chosen as a test protein. To allow the application of Equation 4 only aspartates  
 22 and glutamates were considered in the calculations as titratable and all other titratable residues were fixed  
 23 in the standard protonation state at physiological pH, i.e. lysines and tyrosines protonated. Only one *syn*  
 24 orientation for the proton in the protonated form was considered. With 7 aspartic and 2 glutamic acids, it  
 25 gives 512=2<sup>9</sup> possible protonation states. The structures corresponding to all possible protonation states  
 26 were generated, and Drude particles were fully optimized in the field of the PB implicit solvation model in

1 each of the structures. The internal dielectric constant of two was used. The total free energies were used  
2 to compute an average number of bound protons using Equation 4.  $pK_a$ 's were estimated as the pH where  
3 residues were half-protonated on average.  $pK_a$ 's were also calculated using the new method.

4 For the lysozyme system the new method converged within two self-consistent iterations as  
5 computed  $pK_a$ 's were invariant with more iterations. The results indicate that the computed  $pK_a$ 's with the  
6 new method and two iterations are practically identical to those estimated with the exact form of Equation  
7 4. The RMS deviation between  $pK_a$ 's computed with the two methods is just 0.02 pH units.  $pK_a$ 's computed  
8 with one iteration of the new method differ more from the ones computed with the exact statistical approach,  
9 by 0.07 pH units.

10  $pK_a$  calculations were performed with the protein dielectric constant of 4 and the Drude-PB model  
11 for all 8 proteins. The self-consistent iterations were repeated four times. The results for the  $pK_a$  calculations  
12 versus the experimental values as well as subsequent iterations as a function of the number of iterations are  
13 given in Table 3. The RMS deviation between  $pK_a$ 's computed after the second iteration relative to those  
14 after the first iteration is 0.15 pH units, and reduces to 0.10 and 0.08 pH units after the third and the fourth  
15 iterations, respectively. However, that RMS deviation between computed and experimental  $pK_a$ 's only  
16 changes insignificantly from 1.94 to 1.93 pH units after the second iteration and stays practically the same  
17 after the third and fourth iterations. The linear correlation between computed and experimental  $pK_a$ 's,  $R$ ,  
18 does not improve. However, the computed  $pK_a$ 's slightly change as a function of the number of iterations.  
19 Importantly, the difference between the first and subsequent iterations is that the polarization is inconsistent  
20 in the first round of  $pK_a$  calculations, but it is improved in the subsequent iterations. Though we find only  
21 a moderate change due to the consistent treatment of the polarization, it may be attributed, at least in part,  
22 to the lack of the protein flexibility in this work. In the following sections, all results of  $pK_a$  calculations  
23 with the Drude-PB model will be presented using two iterations, since the computed  $pK_a$ 's change less than  
24 0.1 pH units with more iterations and the exact  $pK_a$ 's were reached within two iterations for the reduced  
25 lysozyme system.

26 **Table 3.** Convergence of the  $pK_a$  calculation method with the Drude-PB model. Calculations were done  
27 using the protein dielectric constant of 4.

Iteration	<sup>a</sup> RMSD	<sup>b</sup> RMSD	<sup>b</sup> correlation	<sup>b</sup> max  error
1	-	1.94	0.71	5.53
2	0.15	1.93	0.70	5.64
3	0.10	1.93	0.70	5.64
4	0.08	1.93	0.70	5.64

1 <sup>a</sup>RMS deviation between pK<sub>a</sub>'s computed in this step and in the previous step; <sup>b</sup>relative to the experimental  
2 pK<sub>a</sub>'s

### 3 **Comparison of the polarizable Drude and additive C36 force fields.**

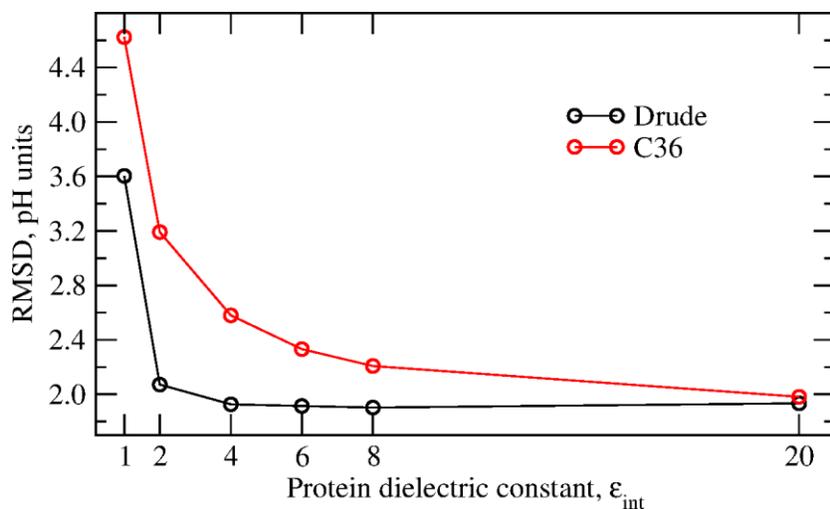
4 To test the dependence of the result on the internal dielectric constant, pK<sub>a</sub> calculations were  
5 performed with  $\epsilon_p$  in the range between 1 and 20 with the Drude and C36 force fields. For the calculations  
6 with the Drude force field, the resulting pK<sub>a</sub>'s were taken after the second self-consistent iteration. For the  
7 calculations with C36, only one iteration is required as electronic polarization is included implicitly. The  
8 results are summarized in Table 4. The computed and experimental pK<sub>a</sub> shifts are given in Table S3, and  
9 absolute pK<sub>a</sub>'s are given in Table S4 in the Supplementary Information. Figure 2 shows the dependence of  
10 the RMS deviation against the internal dielectric constant. The correlation is best with both models at the  
11 internal dielectric constant of two. However, in contrast to the results obtained with the C36 force field,  
12 with the Drude model the RMS deviation is characterized by a shallow minimum at  $\epsilon$  in the range of 4-8.  
13 With the additive force field, the RMS deviation is improving monotonically in the tested range of  $\epsilon$ .  
14 Overall, the Drude model demonstrates a better agreement with the experimental pK<sub>a</sub>'s than the C36 model  
15 at low values of the dielectric constant. The RMS deviation between the experimental pK<sub>a</sub>'s and pK<sub>a</sub>'s  
16 computed using the protein dielectric of two is 2.07 and 3.19 units with the Drude and C36 force fields,  
17 respectively. With the protein dielectric constant of four, the RMS deviation is 1.93 and 2.58 units with the  
18 Drude and C36 force field, respectively. With the Drude-PB model, the RMS deviation between the  
19 experimental pK<sub>a</sub>'s and pK<sub>a</sub>'s computed with the protein dielectric constant of 20 is 1.93, which is very close  
20 to the result of 1.93 and 2.07 units computed with the protein dielectric constant of four and two,  
21 respectively. In contrast to the results with the additive C36 model, the RMS deviation computed with the  
22 Drude-PB model is substantially less sensitive to the choice of the internal dielectric constant. However,  
23 with the Drude model, the RMS deviation sharply increases with an internal protein dielectric constant  
24  $\epsilon_p = 1$ , and the linear correlation decreases to 0.46. A Drude-PB model with  $\epsilon_p = 1$  accounts only for the  
25 induced polarization, leaving out all contributions from structural fluctuations. The poor performance  
26 suggests that such a model does not represent the protein interior as sufficiently polarizable. Interestingly,  
27 the RMS deviation for the Drude model with  $\epsilon_p = 1$  is very similar to the RMS deviation for the additive  
28 force field with  $\epsilon_p \approx 1.7$ , a value that corresponds roughly to the expected dielectric constant associated  
29 with electronic induced polarization.

30 **Table 4.** Performance of the methods for pK<sub>a</sub> calculations against experimental pK<sub>a</sub>'s. RMS deviation and  
31 linear correlation coefficient between computed and experimental pK<sub>a</sub> shifts from the model compound  
32 reference values are given.

Protein	RMSD	Correlation	<sup>a</sup> Slope
---------	------	-------------	--------------------

dielectric, $\epsilon_p$	Drude	C36	Drude	C36	Drude	C36
1	3.57	4.62	0.46	0.71	1.8	3.6
2	2.07	3.19	0.71	0.74	1.7	2.6
4	1.93	2.58	0.70	0.73	1.5	2.0
6	1.91	2.33	0.67	0.71	1.4	1.7
8	1.90	2.21	0.64	0.68	1.3	1.5
20	1.93	1.98	0.53	0.57	1.0	1.1

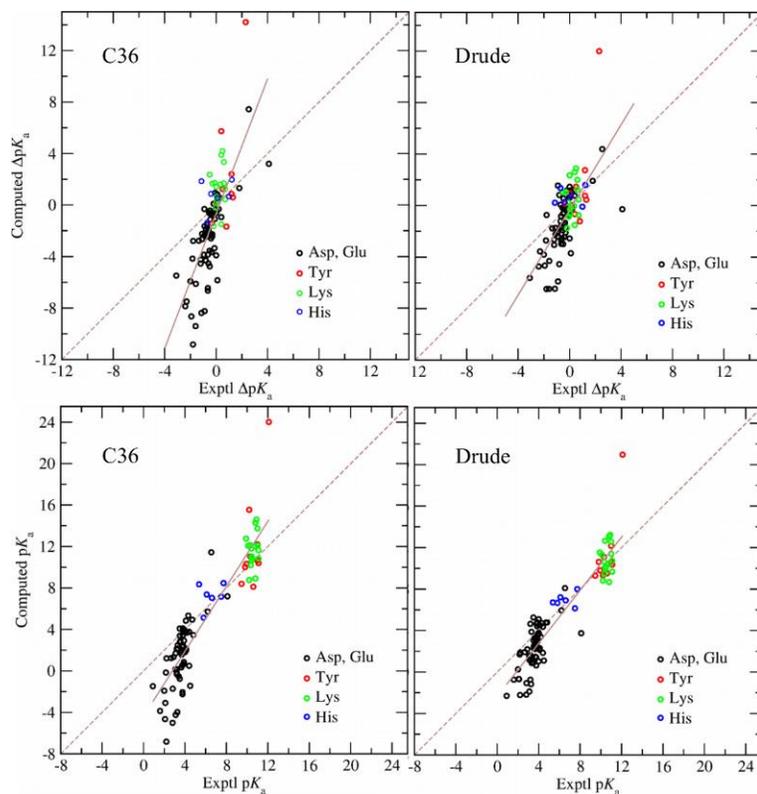
1 <sup>A</sup>The slope of the liner fit to the computed and experimental  $pK_a$  shifts.



2  
3 **Figure 2.** RMS deviation between experimental and computed  $pK_a$ 's.  $pK_a$ 's with the Drude force field  
4 were calculated using two iterations to determine the most probable protonation microstates.

5 Figure 3 gives the comparison between experimental and predicted  $pK_a$  shifts with the protein  
6 dielectric constant of two and the Drude and C36 models. As may be seen, with both Drude and C36 models  
7 computed  $pK_a$  shifts are both systematically underestimated and overestimated relative to the experimental  
8 values, so that a linear fit has a constant positive slope. This slope is also given in Table 2 as a function of  
9 the protein dielectric constant. However, the  $pK_a$  computed with the Drude model are systematically less  
10 over and underestimated in comparison with the results obtained with the C36 model. The slope with the  
11 internal dielectric constant of two is 1.7 and 2.6 with the Drude and C36 models, respectively. The slope is  
12 decreasing with the higher protein dielectric constant and with  $\epsilon_{int} = 20$  it is practically 1.0 with both  
13 models. Figure 3 also contains comparison of the absolute computed and experimental  $pK_a$  values. The  
14 correlation coefficients for the absolute  $pK_a$ 's were 0.93 and 0.91 for the Drude and C36 force fields,  
15 respectively. These values are higher than those for the  $pK_a$  shifts reported in Table 4 due to the wider range  
16 of absolute  $pK_a$ 's associated with the different classes of residues.

17



1  
 2 **Figure 3.** Experimental vs computed  $pK_a$  shifts and absolute  $pK_a$ 's. Left panels: (upper)  $pK_a$  shifts and  
 3 (lower) absolute  $pK_a$ 's computed with the C36 force field; right panels: (upper)  $pK_a$  shifts and (lower)  
 4  $pK_a$ 's computed with the Drude force field after iteration 2. In both calculations, the protein dielectric  
 5 constant of two was used. The solid line shows the linear fit to the data; the dashed line shows the perfect  
 6 match between computed and experimental  $pK_a$  shifts or  $pK_a$ 's.

7 Table 5 gives the comparison between  $pK_a$  shifts computed with the Drude and C36 models. With  
 8 the low internal dielectric constant of two, the RMS deviation between  $pK_a$  shifts of the titratable residues  
 9 in the eight proteins computed with the two methods is 1.78 units and decreases with the higher dielectric  
 10 constant values. With  $\epsilon_{\text{int}} = 20$ , the  $pK_a$  shifts computed by the two methods are very close with the RMS  
 11 deviation of just 0.34 units. The linear correlation between  $pK_a$  shifts computed by the two methods is 0.92  
 12 and 0.99 with  $\epsilon_{\text{int}} = 2$  and  $\epsilon_{\text{int}} = 20$ , respectively. This further demonstrates that at the high internal  
 13 dielectric constant  $pK_a$ 's computed with the C36 model converge to those obtained with the polarizable  
 14 Drude model. This result may be understood by the fact that with the high dielectric constant, electrostatic  
 15 interactions are screened strongly, and thus polarization contributions due to those interactions are expected  
 16 to be smaller. In other words, with the high internal dielectric constant, protein polarization is close to that  
 17 observed in individual residues in solvent, so the difference in polarization observed in solvent and in the  
 18 protein plays a smaller role in  $pK_a$  calculations in accordance with the thermodynamic cycle in Figure 1.

1 **Table 5.** Comparison between  $pK_a$  shifts computed using the Drude and additive C36 models. RMS  
 2 deviation and linear correlation coefficient between  $pK_a$  shifts computed with the Drude and C36 model  
 3 are given.  $pK_a$  shifts computed with the Drude model were taken after two iterations in the method.

Protein dielectric constant, $\epsilon_p$	RMSD	Correlation
2	1.78	0.92
4	0.93	0.97
6	0.68	0.98
8	0.58	0.98
20	0.34	0.99

4  
 5 The agreement between experimental and computed  $pK_a$  shifts for different residue types is given  
 6 in Table 6.  $pK_a$ 's were computed using the C36 and Drude force fields and the dielectric constant of 2. For  
 7 all residue types, the RMS deviation with the Drude force field is better than with the additive force field.  
 8 The RMS deviation is 3.23 units for tyrosines with the Drude force field, which is higher than the RMS  
 9 deviation obtained for the other types. A similar result was obtained with the C36 force field. This may be  
 10 due to the need for larger conformational rearrangements of the protein to occur upon changes in the  
 11 protonation state of tyrosines, since they are larger than other residues and are frequently buried in the  
 12 protein. The poorer correlations for His and Lys with both force fields may indicate the need for larger  
 13 conformation changes of those sidechains upon changes in protonation. Further studies are required to  
 14 address these issues.

15 **Table 6.** Performance of the methods for  $pK_a$  calculations against experimental  $pK_a$ 's for different types  
 16 of residues. RMS deviation and correlation coefficient between computed and experimental  $pK_a$  shifts.

Residue	N sites	RMSD		Correlation	
		Drude	C36	Drude	C36
Asp	31	2.10	3.40	0.66	0.78
Glu	30	2.01	3.40	0.65	0.64
His	6	1.18	1.41	0.18	0.27
Tyr	10	3.23	4.25	0.73	0.69
Lys	17	1.38	1.88	0.19	0.23

17  
 18 **Proton orientation in the protonated form of carboxylic acids**  
 19 The majority of constant pH studies to date have limited treatment of the orientation of the proton  
 20 in neutral carboxylic acids to the *syn* form,<sup>28</sup> omitting consideration of the *anti* orientation, which is known  
 21 to be accessible in condensed phase environments.<sup>29</sup> To investigate if this approximation may be limiting  
 22 the accuracy of the  $pK_a$  estimates of acidic residues we undertook calculations of the carboxylic acid  $pK_a$

1 with and without consideration of the *anti* proton orientation in the protonated form of carboxylic acids.  
 2 Calculations with the dielectric constant of two and only with the *syn* orientation of proton were performed  
 3 and compared with the results of calculations considering both *syn* and *anti* positions of protons. The results  
 4 were obtained using the Drude-PB model and after the second iteration. Results are summarized in Table  
 5 7. The average population of the *anti* protonated form for all aspartates and glutamates in the protein data  
 6 set at a very low pH of 0, where practically all carboxylic acids are protonated, is 27.6%; for aspartic acids,  
 7 this population is 34.0% and 19.5% for glutamic acids. Accordingly, inclusion of the *anti* orientation leads  
 8 to a large improvement in the predicted  $pK_a$ 's relative to the experimental values. For aspartates the RMS  
 9 deviation is improved from 2.87 units considering only the *syn* orientations to 2.10 units when allowing  
 10 both *syn* and *anti* rotamers. A similar improvement is observed for glutamates. In barnase (PDB reference  
 11 code 1a2p), the large improvement with the *anti* orientation was found for residue Asp101. Both oxygens  
 12 of Asp101 participate in hydrogen bond interactions with the backbone and sidechain of Thr105 and the  
 13 sidechain of Thr99. These hydrogen bond interactions make energetically unfavorable the placement of  
 14 proton in the *syn* orientation in the protonated form Asp101. Thus, the calculated  $pK_a$  shift of Asp101 is -  
 15 6.7  $pK_a$  units if only the *syn* orientations are considered, and -3.8  $pK_a$  units if both *syn* and *anti* orientations  
 16 are included. The latter value agrees better with the experimental value of -2.0  $pK_a$  units for Asp101.  
 17 However, as Asp101 as well as other acid moieties may change their orientation upon protonation. The  
 18 improvement in the  $pK_a$  prediction needs to be addressed in future studies with methods that allow for  
 19 conformational changes to occur upon changes in protonation state.

20 **Table 7.** RMS deviation and correlation between computed and experimental  $pK_a$ 's. Calculations were  
 21 performed using the *syn* and *anti* rotamers or only the *syn* rotamers for the proton in the protonated form  
 22 of carboxylic acids. The Drude-PB model was used with the protein dielectric constant of two.

Residue	N sites	RMSD		Correlation	
		<i>syn/anti</i>	<i>only syn</i>	<i>syn/anti</i>	<i>only syn</i>
Asp	31	2.10	2.87	0.65	0.68
Glu	30	2.01	2.82	0.65	0.63
All	94	2.07	2.60	0.67	0.69

23

## 24 Comparison to other methods

25 Assuming the null hypothesis,<sup>30</sup> i.e. that all residues have their solution  $pK_a$  in the protein  
 26 environment, the RMS deviation with the experimental  $pK_a$ 's is 1.16 pH units, lower than the RMS deviation  
 27 obtained with the C36 or Drude force field. This implies that increasing electrostatic screening, in principle  
 28 would improve the RMS deviation, since absolute  $pK_a$  shifts become smaller.

1 We first compare to the results of the H++ server, which uses a single-conformation version of the  
2 MEAD program for  $pK_a$  calculations.<sup>31</sup> The server only provides  $pK_a$ 's for the range between 0 and 12 pH  
3 units. Thus, the comparison will be limited to  $pK_a$ 's within this range (70 values total). In principle, H++  
4 relies on the same method that we used for the calculations with the additive C36 force field, but uses the  
5 AMBER force field and van der Waals radii defined by Bondi.<sup>32</sup> With the internal dielectric constant of 4  
6 and implicit salt concentration of 0, the RMS deviation between the experimental and computed  $pK_a$ 's using  
7 the H++ server is 1.55 pH units, and the linear correlation coefficient is 0.65. The RMS deviation for the  
8 same 70  $pK_a$  values computed using  $\epsilon_{\text{int}} = 4$  and the C36 force field and the radii specifically optimized  
9 previously for PB calculations<sup>33</sup> is 2.08 pH units and the linear correlation coefficient is 0.59. However, the  
10 Bondi radii are significantly smaller than the Born radii derived by Nina et al<sup>33</sup> that were optimized targeting  
11 explicit solvent molecular dynamics simulations with an internal dielectric constant of 1. For example, the  
12 radius of the OH oxygen of tyrosine is 1.85 Å and 1.5 Å in the Nina et al<sup>33</sup> and Bondi sets, respectively.  
13 The radius of N $\delta$  and N $\epsilon$  of the protonated form of histidine is 2.3 Å and 1.55 Å in Nina et al and Bondi  
14 sets respectively. With the C36 force field and Bondi radii and  $\epsilon_{\text{int}} = 4$  and the molecular surface as the  
15 dielectric boundary (the water probe radius of 1.4 Å), the RMS deviation with the experimental  $pK_a$ 's is  
16 1.06 with a linear correlation of 0.70. However, with the Bondi radii, the absolute solvation energies of  
17 small molecules are significantly overestimated. The RMS deviation between computed and experimental  
18 absolute solvation free energies for the set of small molecules that was used in our previous study<sup>15</sup> to  
19 optimize the Drude PB radii is 4.1 kcal·mol<sup>-1</sup>, while with the optimized set of radii from Nina et al<sup>33</sup> the  
20 RMS deviation is 2.5 kcal·mol<sup>-1</sup>. In the continuum dielectric model, the induced charges in the solvent  
21 continuum dielectric medium are located within an infinitesimal layer at the boundary of the solute volume.  
22 In contrast, the solvent charge density in an atomic model is distributed over a microscopic region of space  
23 of finite dimension.<sup>33</sup> Thus, the PB model with the van der Waals radii and dielectric constant of one  
24 significantly overestimates solvation energies. The radii that were optimized specifically to reproduce  
25 results of molecular dynamics free energy simulations are significantly larger than the Bondi (van der  
26 Waals) radii. Similar to using the higher internal dielectric constant, using smaller atomic radii significantly  
27 increases solvent screening leading to smaller absolute  $pK_a$  shifts and, thus giving a lower RMS deviation.

28 The reported  $pK_a$ 's computed with the MCCE2 method<sup>7c</sup> were used to compare with the results of  
29 the current work. MCCE2 introduces the conformational relaxation and uses the Poisson-Boltzmann model  
30 for electrostatic calculations, which involves approximations to the protein-solvent boundary, and uses the  
31 PARSE charges and radii.<sup>34</sup> The PARSE charges and radii were optimized to reproduce experimental  
32 solvation energies, but with the dielectric constant of two. Thus, like van der Waals radii, the PARSE radii  
33 are significantly smaller than the radii optimized with the internal dielectric constant of 1. For example, the  
34 radius of the OH oxygen of tyrosine is 1.85 Å and 1.5 Å in Nina et al<sup>33</sup> and PARSE sets, respectively. The

1 radius of  $N\delta$  and  $N\epsilon$  of the protonated form of histidine is 2.3 Å and 1.5 Å in Nina et al<sup>33</sup> and PARSE sets,  
2 respectively. The RMS deviation computed for the MCCE2 results obtained with  $\epsilon_{\text{int}} = 4$  that do not  
3 include the SNase variant  $\Delta$ +PHS protein and Tyr53 in Lysozyme is 0.75 pH units with the linear  
4 correlation of 0.78. With the C36 force field and Bondi radii and  $\epsilon_{\text{int}} = 4$  using the same titratable residue  
5 sets the RMS deviation is 1.42 pH units and the linear correlation is 0.73. With the Nina et al radii the RMS  
6 deviation is 2.36 pH units and the linear correlation is 0.68. Overall, this demonstrates that the PB model  
7 strongly depends on the atomic Born radii, which is entirely expected.<sup>33, 35</sup>

## 8 **Conclusion**

9 In this study, a new method to estimate  $pK_a$  of titratable residues is presented that uses the  
10 polarizable Drude-PB model and constant-pH Monte Carlo simulations. The main challenge in using the  
11 polarizable Drude-PB model, as well as any other polarizable PB force field, is due to the dependence of  
12 the energy terms on the electronic polarization of the entire system, which in turn depends on the  
13 protonation state of all protein residues. As this represents a large computational increase in the calculation  
14 of energy matrices used in the constant-pH simulations an additional approximation is required to make the  
15 calculation feasible, which we propose and implement in the present work. In this approximation, only the  
16 polarization of the highly populated protein protonation microstates (ie. when the pH is equivalent to the  
17  $pK_a$  of the residue associated with those microstates) are treated explicitly using the corresponding protein  
18 protonation state in conjunction with optimization of the Drude particles as required to model the  
19 polarization response. The method necessitates self-consistent calculations of the most populated  
20 microstates and residue  $pK_a$ 's, since the  $pK_a$ 's are needed to define the most populated microstates and vice  
21 versa. A numerical test with a small protein, lysozyme, shows that the  $pK_a$ 's computed with the new method  
22 differ by only 0.02 pH units from the ones estimated with the exact statistical approach, demonstrating that  
23 polarization effects are correctly included in the MC simulations.

24 The present method with the Drude-PB model considerably increases the computational cost  
25 relative to the calculations with the C36 additive force field. The extra cost is arising, first due to the need  
26 to compute the solute polarization, i.e. optimize the position of Drude particles for each protonation state  
27 of all residues. To optimize the position of the Drude particles, the solvent reaction field due to the PB  
28 implicit solvent model in the current implementation is allowed to fully relax after each minimization step  
29 to calculate solvent forces. In the previous work, we demonstrated that the optimization of the Drude  
30 particles converges within 50 minimization steps. Second, additional cost is due to the need to calculate the  
31 most populated microstates and  $pK_a$ 's iteratively. The self-consistent approach converged within two  
32 iterations with  $pK_a$ 's computed after iteration 3 differing less than 0.1  $pK_a$  units from  $pK_a$ 's after iteration  
33 2. Thus, the SCF protocol of the  $pK_a$  calculation scheme increases the overall cost by two times. Overall,

1 the method for  $pK_a$  calculations using the constant-pH simulations with the Drude force field takes an  
2 average of two orders of magnitude more CPU time than the standard protocol for the  $pK_a$  calculation with  
3 an additive force field and the PB solvation model. For example, the  $pK_a$  calculation for 3bdc, the protein  
4 with the largest number of titratable residues (44 residues) consumes approximately 2 CPU Hrs. with the  
5 additive force field versus 95 CPU Hrs. for the Drude force fields on an Intel Xeon E5-2630 type processor.

6 A significant improvement for the predicted  $pK_a$ 's was observed with the Drude-PB model  
7 compared to results based on the additive force field C36 at low dielectric constants. Using the Drude-PB  
8 model with an internal protein dielectric constant of 2, the RMS deviation from the experimental  $pK_a$ 's is  
9 2.07  $pK_a$  units. In contrast, the C36 additive force field yields a RMS deviation of 3.19  $pK_a$  units with a  
10 dielectric constant of 2, and a RMSD of 2.58  $pK_a$  units with a dielectric constant of 4. The RMS is still  
11 higher than with the Drude-PB model with a dielectric of 2. Notably, the results with the Drude force field  
12 are less sensitive to the choice of internal dielectric constant, with a higher protein dielectric constant of 4  
13 and the Drude-PB model the RMS deviation is 1.93  $pK_a$  units, close to 2.07  $pK_a$  units obtained with  $\epsilon_{\text{int}} =$   
14 2. We also observe that the  $pK_a$ 's computed with the high internal dielectric constant of 20 are very similar  
15 for the two force fields with an RMS deviation of just 0.36 units and the linear correlation of 0.99. These  
16 results indicate that a model accounting explicitly for the induced polarization represents a physically more  
17 correct model that decreases the empirical requirement to ascribe an excessively high dielectric constant to  
18 the protein interior. Given the heterogeneity of the protein interior, it is likely that simply assigning a high  
19 dielectric constant to the protein interior cannot accurately substitute for an explicit treatment of  
20 polarization during protonation/deprotonation events.

21 An interesting observation was the better agreement with experimental  $pK_a$ 's when the *anti*  
22 protonated form of carboxylic acids was explicitly considered. This is due to a relatively high contribution  
23 from the *anti* protonated form of carboxylic acids of ~28% at a very low pH. However, the contribution of  
24 the *anti* orientation of the proton is expected to be impacted by the ability of the side chains as well as  
25 surrounding protein to relax upon protonation. This effect, as well as the impact of conformational  
26 flexibility on  $pK_a$  calculations using the polarizable model will be addressed in future studies.

27 The current implementation of the method for  $pK_a$  calculations with the Drude-PB models bears  
28 several limitations. Only polarization of one protonation microstate for each possible protonation state of  
29 all residues is computed exactly, while for minor microstates a surrogate of the energy components that  
30 include the residue interaction free energies and self-energies, corresponding to different pH's is used. In  
31 principle, one can consider additional protonation microstates in energy matrix calculations, and use those  
32 energies in the MC simulations. However, the main limitation of the presented method is the lack of  
33 conformational relaxation and fluctuations, which is required to preserve a fixed protein dielectric boundary

1 in the Poisson-Boltzmann calculations. Various approximations have been introduced in previous studies  
2 to circumvent this prescription.<sup>7c, 8a, 36</sup> We will explore the presented Drude-PB method in combination with  
3 existing approximations to treat protein conformational changes in future studies.

4 **Acknowledgements:** This work was supported by the French National Research Agency grant ANR-18-  
5 CE44-0002 to AA, National Institutes of Health grants GM131710 to ADM and GM072558 to BR and the  
6 Samuel Waxman Cancer Foundation. The University of Maryland Computer-Aided Drug Design Center,  
7 XSEDE, CINES (Grant 2018-A0040710436) are acknowledged for their generous allocations of computer  
8 time.

9 **Competing financial interests:** ADM is co-founder and CSO of SilcsBio LLC.

## 10 **References:**

- 11 1. Jordan, I. K.; Kondrashov, F. A.; Adzhubei, I. A.; Wolf, Y. I.; Koonin, E. V.; Kondrashov, A. S.;  
12 Sunyaev, S., A universal trend of amino acid gain and loss in protein evolution. *Nature* **2005**, *433* (7026),  
13 633-8.
- 14 2. (a) Honig, B.; Nicholls, A., Classical electrostatics in biology and chemistry. *Science* **1995**, *268*  
15 (5214), 1144-9; (b) Pace, C. N.; Grimsley, G. R.; Scholtz, J. M., Protein ionizable groups: pK values and  
16 their contribution to protein stability and solubility. *J. Biol. Chem.* **2009**, *284* (20), 13285-9; (c) Onufriev,  
17 A. V.; Alexov, E., Protonation and pK changes in protein-ligand binding. *Q. Rev. Biophys.* **2013**, *46* (2),  
18 181-209; (d) Bartlett, G. J.; Porter, C. T.; Borkakoti, N.; Thornton, J. M., Analysis of catalytic residues in  
19 enzyme active sites. *J. Mol. Biol.* **2002**, *324* (1), 105-21.
- 20 3. Castaneda, C. A.; Fitch, C. A.; Majumdar, A.; Khangulov, V.; Schlessman, J. L.; Garcia-Moreno,  
21 B. E., Molecular determinants of the pKa values of Asp and Glu residues in staphylococcal nuclease.  
22 *Proteins* **2009**, *77* (3), 570-88.
- 23 4. (a) Nielsen, J. E.; Gunner, M. R.; Garcia-Moreno, B. E., The pKa Cooperative: a collaborative  
24 effort to advance structure-based calculations of pKa values and electrostatic effects in proteins. *Proteins*  
25 **2011**, *79* (12), 3249-59; (b) Alexov, E.; Mehler, E. L.; Baker, N.; Baptista, A. M.; Huang, Y.; Milletti, F.;  
26 Nielsen, J. E.; Farrell, D.; Carstensen, T.; Olsson, M. H.; Shen, J. K.; Warwicker, J.; Williams, S.; Word,  
27 J. M., Progress in the prediction of pKa values in proteins. *Proteins* **2011**, *79* (12), 3260-75.
- 28 5. Simonson, T.; Carlsson, J.; Case, D. A., Proton binding to proteins: pK(a) calculations with  
29 explicit and implicit solvent models. *J. Am. Chem. Soc.* **2004**, *126* (13), 4167-80.
- 30 6. (a) Bashford, D., Macroscopic electrostatic models for protonation states in proteins. *Front.*  
31 *Biosci.* **2004**, *9*, 1082-99; (b) Baker, N. A., Poisson-Boltzmann methods for biomolecular electrostatics.  
32 *Methods Enzymol.* **2004**, *383*, 94-118; (c) Tanford, C.; Kirkwood, J., Theory of Protein Titration Curves.  
33 I. General Equations for Impenetrable Spheres. *J. Am. Chem. Soc.* **1957**, *79* (20), 5333-9; (d) Dolinsky, T.  
34 J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A., PDB2PQR: an automated pipeline for the setup of  
35 Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32* (Web Server issue), W665-7.
- 36 7. (a) Warwicker, J.; Watson, H. C., Calculation of the electric potential in the active site cleft due  
37 to alpha-helix dipoles. *J. Mol. Biol.* **1982**, *157* (4), 671-9; (b) Beroza, P.; Fredkin, D. R.; Okamura, M. Y.;  
38 Feher, G., Protonation of interacting residues in a protein by a Monte Carlo method: application to  
39 lysozyme and the photosynthetic reaction center of *Rhodobacter sphaeroides*. *Proc. Natl. Acad. Sci.*  
40 *U.S.A.* **1991**, *88* (13), 5804-8; (c) Song, Y.; Mao, J.; Gunner, M. R., MCCE2: Improving protein pKa  
41 calculations with extensive side chain rotamer sampling. *J. Comput. Chem.* **2009**, *30* (14), 2231-47.
- 42 8. (a) Aleksandrov, A.; Polydorides, S.; Archontis, G.; Simonson, T., Predicting the acid/base  
43 behavior of proteins: a constant-pH Monte Carlo approach with generalized born solvent. *J. Phys. Chem.*

1 *B* **2010**, *114* (32), 10634-48; (b) Mongan, J.; Case, D. A.; McCammon, J. A., Constant pH molecular  
2 dynamics in generalized Born implicit solvent. *J. Comput. Chem.* **2004**, *25* (16), 2038-48; (c) Lee, M. S.;  
3 Salsbury, F. R., Jr.; Brooks, C. L., 3rd, Constant-pH molecular dynamics using continuous titration  
4 coordinates. *Proteins* **2004**, *56* (4), 738-52.

5 9. Bashford, D.; Karplus, M., The pKa's of ionizable groups in proteins: atomic detail from a  
6 continuum electrostatic model. *Biochemistry* **1990**, *29* (44), 10219-25.

7 10. (a) Wang, L.; Li, L.; Alexov, E., pKa predictions for proteins, RNAs, and DNAs with the  
8 Gaussian dielectric function using DelPhi pKa. *Proteins* **2015**, *83* (12), 2186-97; (b) Tanford, C.; Roxby,  
9 R., Interpretation of protein titration curves. Application to lysozyme. *Biochemistry* **1972**, *11* (11), 2192-  
10 8; (c) Bashford, D.; Karplus, M., Multiple-site titration curves of proteins: an analysis of exact and  
11 approximate methods for their calculation. *J. Phys. Chem.* **1991**, *95* (23), 9556-61; (d) Gilson, M. K.,  
12 Multiple-site titration and molecular modeling: two rapid methods for computing energies and forces for  
13 ionizable groups in proteins. *Proteins* **1993**, *15* (3), 266-82.

14 11. Yang, A. S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B., On the calculation of pKas in  
15 proteins. *Proteins* **1993**, *15* (3), 252-65.

16 12. Kaminski, G. A., Accurate prediction of absolute acidity constants in water with a polarizable  
17 force field: substituted phenols, methanol, and imidazole. *J. Phys. Chem. B* **2005**, *109* (12), 5884-90.

18 13. (a) Schnieders, M. J.; Baker, N. A.; Ren, P.; Ponder, J. W., Polarizable atomic multipole solutes  
19 in a Poisson-Boltzmann continuum. *J. Chem. Phys.* **2007**, *126* (12), 124114; (b) Lipparini, F.; Lagardère,  
20 L.; Raynaud, C.; Stamm, B.; Cancès, E.; Mennucci, B.; Schnieders, M.; Ren, P.; Maday, Y.; Piquemal, J.-  
21 P., Polarizable Molecular Dynamics in a Polarizable Continuum Solvent. *J. Chem. Theory Comput.* **2015**,  
22 *11* (2), 623-34.

23 14. Georgescu, R. E.; Alexov, E. G.; Gunner, M. R., Combining conformational flexibility and  
24 continuum electrostatics for calculating pK(a)s in proteins. *Biophys. J.* **2002**, *83* (4), 1731-48.

25 15. Aleksandrov, A.; Lin, F. Y.; Roux, B.; MacKerell, A. D., Jr., Combining the polarizable Drude  
26 force field with a continuum electrostatic Poisson-Boltzmann implicit solvation model. *J. Comput. Chem.*  
27 **2018**, *39* (22), 1707-19.

28 16. Darden, T., Treatment of long-range forces and potential. In *Computational Biochemistry &*  
29 *Biophysics*, Marcel Dekker, N.Y. : 2001.

30 17. (a) Mackerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J.; Field, M. J.;  
31 Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.;  
32 Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Smith, J.; Stote, R.; Straub,  
33 J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M., An all-atom empirical potential for  
34 molecular modelling and dynamics study of proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586-616; (b)  
35 Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Optimization of the  
36 Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone  $\phi$ ,  $\psi$   
37 and Side-Chain  $\chi_1$  and  $\chi_2$  Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257-73.

38 18. Lopes, P. E. M.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell, A. D., Polarizable  
39 Force Field for Peptides and Proteins Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.*  
40 **2013**, *9* (12), 5430-49.

41 19. Jorgensen, W.; Chandrasekar, J.; Madura, J.; Impey, R.; Klein, M., Comparison of simple  
42 potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926-35.

43 20. Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., A polarizable model of  
44 water for molecular dynamics simulations of biomolecules. *Chem. Phys. Lett.* **2006**, *418* (1), 245-9.

45 21. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.  
46 D.; Kale, L.; Schulten, K., Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16),  
47 1781-802.

48 22. Sham, Y. Y.; Muegge, I.; Warshel, A., The effect of protein relaxation on charge-charge  
49 interactions and dielectric constants of proteins. *Biophys. J.* **1998**, *74* (4), 1744-53.

- 1 23. (a) Harvey, S. C., Treatment of electrostatic effects in macromolecular modeling. *Proteins* **1989**,  
2 5 (1), 78-92; (b) Simonson, T.; Perahia, D.; Brünger, A. T., Microscopic theory of the dielectric properties  
3 of proteins. *Biophys. J.* **1991**, 59 (3), 670-90.
- 4 24. Schutz, C. N.; Warshel, A., What are the dielectric "constants" of proteins and how to validate  
5 electrostatic models? *Proteins* **2001**, 44 (4), 400-17.
- 6 25. Im, W.; Beglov, D.; Roux, B., Continuum solvation model: computation of electrostatic forces  
7 from numerical solutions to the Poisson-Boltzmann equation. *Comp. Phys. Comm.* **1998**, 111 (1), 59-75.
- 8 26. Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.;  
9 Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer,  
10 S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor,  
11 R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang,  
12 W.; York, D. M.; Karplus, M., CHARMM: the biomolecular simulation program. *J. Comp. Chem.* **2009**,  
13 30 (10), 1545-614.
- 14 27. Guennebaud, G. J., B. Eigen v3.
- 15 28. (a) Khandogin, J.; Brooks, C. L., 3rd, Constant pH molecular dynamics with proton tautomerism.  
16 *Biophys. J.* **2005**, 89 (1), 141-57; (b) Huang, Y.; Harris, R. C.; Shen, J., Generalized Born Based  
17 Continuous Constant pH Molecular Dynamics in Amber: Implementation, Benchmarking and Analysis. *J.*  
18 *Chem. Inf. Model.* **2018**, 58 (7), 1372-83.
- 19 29. MacKerell, A. D., Jr.; Sommer, M. S.; Karplus, M., pH dependence of binding reactions from  
20 free energy simulations and macroscopic continuum electrostatic calculations: application to  
21 2'GMP/3'GMP binding to ribonuclease T1 and implications for catalysis. *J. Mol. Biol.* **1995**, 247 (4), 774-  
22 807.
- 23 30. Antosiewicz, J.; McCammon, J. A.; Gilson, M., Prediction of pH dependent properties of  
24 proteins. *J. Mol. Biol.* **1994**, 238 (3), 415-36.
- 25 31. Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V., H++ 3.0: automating pK prediction and the  
26 preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids*  
27 *Res.* **2012**, 40 (Web Server issue), W537-41.
- 28 32. Bondi, A., van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, 68 (3), 441-51.
- 29 33. Nina, M.; Beglov, D.; Roux, B., Atomic radii for continuum electrostatics calculations based on  
30 molecular dynamics free energy simulations. *J. Phys. Chem. B* **1997**, 101 (26), 5239-48
- 31 34. Sitkoff, D.; Sharp, K.; Honig, B., Accurate calculation of hydration free energies using  
32 macroscopic solvent models *J. Phys. Chem.* **1994**, 98 (7), 1978-88
- 33 35. Roux, B.; Yu, H. A.; Karplus, M., Molecular basis for the Born model of ion solvation. *J. Phys.*  
34 *Chem.* **1990**, 94 (11), 4683-8.
- 35 36. Villa, F.; Mignon, D.; Polydorides, S.; Simonson, T., Comparing pairwise-additive and many-  
36 body generalized Born models for acid/base calculations and protein design. *J. Comput. Chem.* **2017**, 38  
37 (28), 2396-410.

38