On the Use of Different Dielectric Constants for Computing Individual and Pairwise Terms in Poisson−Boltzmann Studies of Protein Ionization Equilibrium

Vitor H. Teixeira, Carlos A. Cunha, Miguel Machuqueiro, A. Sofia F. Oliveira, Bruno L. Victor, Cláudio M. Soares, and António M. Baptista*

Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Av. da República, EAN, Apartado 127, 2781-901 Oeiras, Portugal

Received: May 1, 2005; In Final Form: June 3, 2005

Poisson−Boltzmann (PB) models are a fast and common tool for studying electrostatic processes in proteins, particularly their ionization equilibrium (protonation and/or reduction), often yielding quite good results when compared with more detailed models. Yet, they are conceptually very simple and necessarily approximate, their empirical character being most evident when it comes to the choice of the dielectric constant assigned to the protein region. The present study analyzes several factors affecting the ability of PB-based methods to model protein ionization equilibrium. We give particular attention to a suggestion made by Warshel and co-workers (e.g., Sham et al. *J. Phys. Chem. B* **1997**, *101*, 4458) of using different protein dielectric constants for computing the individual (site) and the pairwise (site−site) terms of the ionization free energies. Our prediction of pK_a values for several proteins indicates that no advantage is obtained by such a procedure, even for sites that are buried and/or display large pK_a shifts relative to the solution values. In particular, the present methodology gives the best predictions using a dielectric constant around 20, for shifted/buried and nonshifted/exposed sites alike. The similarities and differences between the PB model and Warshel's PDLD/S model are discussed, as well as the reasons behind their apparently discrepant results. The present PB model is shown to predict also good reduction potentials in redox proteins.

1. Introduction

The ionization state of a molecule in aqueous solution is largely a determinant of its properties and behavior. This is particularly evident for protein molecules, whose biological role is often drastically dependent on the solution pH, which directly determines the protonation states of the protein titrable sites. An analogous situation holds for redox proteins and the dependence of their properties on the solution (electrode) electrostatic potential. The theoretical approaches aimed at understanding ionization equilibrium in proteins go back to the Linderstrøm−Lang[1] and Tanford−Kirkwood[2] models, which were essentially extensions of the original Debye−Hückel model. The increase of computational power gradually led to an increased realism in the models used, and the study of ionization processes can currently be addressed at full atomic detail using a range of techniques spanning from molecular mechanics/dynamics (MM/MD) to semiclassical and quantum methods.[3−7] But simplified models remain an attractive and particularly fast route to address the full ionization (protonation and/or reduction) equilibrium of a large protein molecule, often giving better results than more detailed models and, thus, continue to be used in many computational studies of this kind. Here, we will be mainly interested in those methods that consider a single protein conformation (see below), thereby avoiding explicitly addressing the conformational reorganization that may take place in response to ionization changes. Two major methods of this type are currently in use: the so-called semimacroscopic version of the protein dipole Langevin dipole (PDLD/S) model,[8,9] and the atomic-level Poisson−Boltzmann (PB) model[10−15] (generalized Born and other implicit solvent models are also becoming increasingly popular,[16] but will not be discussed here). These methods use a simplified treatment for both the solvent (grid of Langevin dipoles or dielectric continuum) and the protein (point charges and/or induced point dipoles "scaled" by a dielectric constant), which allows for a fast calculation of the ionization free-energy terms. These energy terms can then be used to sample ionization (protonation or redox) states by Monte Carlo (MC)[13,17−19] or more approximate methods.[8,13,20−23] The resulting methodology has been widely used to study the protonation[8,12,13,24−32] and/or reduction[19,33−38] equilibrium in proteins and predict the corresponding pK_as or reduction potentials.

The great similarity between the PDLD/S and PB models extends to their major conceptual problem: the physical meaning of the protein "dielectric constant", ϵ_p, and the value that should be used for it. It is possible to restrict the use of these simplified models to address rigid molecules, whose conformational freedom is then explicitly treated via MD. This is partially done in the linear response approximation (LRA) method,[30,39,40] and more fully in the constant-pH MD methods;[41,42] other more approximate treatments also exist (e.g., see discussion in ref 42 and references therein). The more stringent this restriction is, the clearer the meaning of ϵ_p becomes. In the extreme case of constant-pH MD methods, ϵ_p is essentially due to electronic polarization, meaning that $\epsilon_p \approx 2$ should in principle be used.[43] At the other extreme, when no attempt is made to explicitly include conformational effects, the physical meaning of ϵ_p is not at all clear. In practice, it was found that, in this case, large values of ϵ_p can compensate for the rigidity of the structure.[8,9,25,26,29,44−46] However, this does not mean that there is some "true" dielectric constant (e.g., derived from the familiar

* Author to whom correspondence should be addressed. E-mail: baptista@itqb.unl.pt. Telephone: +351 214469619. Fax: +351 214411277.

10.1021/jp052259f CCC: $30.25
Published on Web 07/12/2005

statistical-mechanical polarization theories[43]) that describes the dielectric properties of the protein as a whole. In fact, those dielectric properties are clearly inhomogeneous, and the use of a single large ϵ_p must be regarded as a pragmatic approach yielding good results, rather than a physically sound one.[9,25,26,29,44−53] In general, ϵ_p is simply an empirical parameter that tries to represent contributions not included explicitly in the model.[9,44] It is true that ϵ_p is somehow related to dielectric response, and thus approximate physical reasonings can always be attempted. However, these reasonings are most often the result of importing ideas from conceptually different models and can be considerably misleading. Thus, the decision on the "best" ϵ_p values should ultimately be based on the ability to predict experimental data.

Several routes to go beyond the use of a single ϵ_p have been proposed. One route is to consider successive static and relaxation steps in the charging process and to use different ϵ_p values for each of them.[53,54] Another route is to use different ϵ_p values for different regions of the protein.[50,51,55] Both of these routes have been investigated in the context of PB models, and will not be discussed here. One further route was proposed by Warshel and co-workers,[9,30,46] namely, to compute the individual (site) and the pairwise (site−site) PB energy terms using different values of ϵ_p, hereafter referred to as ϵ_{ind} and ϵ_{pair}, respectively. More exactly, the suggestion was to use a low ϵ_{ind} and a high ϵ_{pair}. In fact, this suggestion was based on studies using the PDLD/S model and followed from the close relation of the latter with the PB model. However, to our knowledge, no PB study using $\epsilon_{ind} \neq \epsilon_{pair}$ was actually performed until now. Thus, it is a central aim of the present work to investigate how the use of different (ϵ_{ind}, ϵ_{pair}) combinations affects the simulation of protein ionization equilibrium. Our general purpose is actually broader, namely the "fine-tuning" of several model parameters, such as atomic radii and water amount, but the analysis of the dielectric combinations has a natural prominent role. This analysis will also allow us to investigate the more or less accepted view that more shifted or buried sites benefit from the use of low dielectric constants in PB models.[25,26,31,32] The present study uses our usual PB/MC methodology,[19,29,37] with the addition of new proton tautomers for carboxylic sites and new ab initio charges for heme sites, as discussed in Section 2. We start Section 3 with a preliminary study of the choice of atomic radii and the inclusion of explicit water molecules. We then proceed to investigate the effect of using different (ϵ_{ind}, ϵ_{pair}) combinations for pK_a prediction, analyzing its dependence on water and/or tautomer inclusion and on the choice of reference state. The reasons for the apparent discrepancy with similar PDLD/S studies are discussed and shown to result essentially from a misinterpretation of the so-called "effective" dielectric constant. We end up with a study of the effect of varying ϵ_p on the prediction of reduction potentials and reduction orders.

2. Materials and Methods

2.1. Protein Sets. The protein set used for pK_a calculations consists of: B2 domain of streptococcal protein G (B2), barnase (Barn), bovine pancreatic trypsin inhibitor (BPTI), hen egg-white lysozyme (HEWL), third domain of turkey ovomucoid inhibitor (OMTky3), *Escherichia coli* ribonuclease HI (RNaseH), *Streptomyces albogriseolus* subtilisin inhibitor (SubIn), and *E. coli* thioredoxin (Thrx). These proteins are common benchmarks for pK_a calculations and provide a balanced amount of shifted and nonshifted pK_a values (see Section 3). Table 1 shows for each protein in this set the PDB entry used, the number of protonable sites (total and with experimental pK_a), and the references providing the pK_a values. A complete list of these pK_a values is provided as Supporting Information.

TABLE 1: Protein Set Used to Compute pK_a Values

protein	PDB entry	no. protonable sites: total	no. protonable sites: with pK_a	references for pK_a values
B2	1IGD	21	13	56
Barn[a]	1A2P	36	12	57
BPTI	4PTI	20	14	58
HEWL	4LZT	32	21	59, 60
OMTky3[a]	1R0R	15	12	61, 62
RNaseH[b]	2RN2	55	22	63, 64
SubIn[a]	3SSI	22	1	65
Thrx[c]	2TRX	34	1	66

[a] First amino group considered neutral and nonprotonable because the true N-terminus is absent from the PDB structure. [b] Without magnesium ion. [c] Oxidized form.

TABLE 2: Protein Set Used to Compute Reduction Potential (E) Values

protein	PDB entry	no. protonable sites (without pK_a)	no. redox sites (with E)	references for E values
DvHc3	2CTH	48	4	67
DvMc3	2CDV	45	4	68
Dgc3[a]	1WAD	45	4	69
Dmnc3	2CY3	44	4	70
Ddc3	3CYR	44	4	71
Dac3	3CAO	37	4	72

[a] With calcium ion.

The protein set used for reduction potential calculations consists of cytochrome c_3 molecules from several sulfate-reducing bacteria: *Desulfovibrio vulgaris* Hildenborough (DvHc3), *D. vulgaris* Miyazaki (DvMc3), *D. gigas* (Dgc3), *Desulfomicrobium norvegicum* (Dmnc3), *D. africanus* (Dac3), and *D. desulfuricans* ATCC 27774 (Ddc3). All these c_3 cytochromes are of Type I (or basic), except for Dac3, which is of Type II (or acidic). Each of these small cytochromes contains four identical heme groups with very similar reduction potentials, the differences resulting only from variations on the local protein environment. Thus, they provide a stringent test on the ability to predict reduction potentials, having been used in previous studies.[19,33,36−38] Table 2 shows for each protein in this set the PDB entry, the number of redox and protonable sites, and the references providing the reduction potentials. Although all heme reduction potentials have been experimentally determined, no experimental pK_as are available. A complete list of the reduction potentials is provided as Supporting Information.

2.2. pK_a Calculations. The basic methodology used for the simulation of protonation equilibrium was previously described in detail.[29] Briefly, it uses PB to compute the individual and pairwise terms needed to obtain the free energies of protonation changes, which are then used to perform MC sampling of protonation states; midpoint pK_as for all sites are then obtained from the corresponding titration curves, that being assumed to correspond to the experimentally determined pK_a values. Unless otherwise stated, all calculations are performed considering alternative proton positions (tautomers) for all titrable sites plus the nontitrable sites Ser, Thr, and selected water molecules (which are treated as six-state rotamers, each obtained by adding four hydrogen atoms at the vertexes of a randomly oriented tetrahedron); details on the chosen positions, corresponding model compound pK_as, etc., were those previously described,[29] except for carboxylic acids (see Section 2.4). Unless otherwise stated, all calculations included crystallographic water molecules,

whose individual relative accessibility was below 0.50 (computed with the program ASC, version 2.14[73,74]); calculations without water molecules were also performed (Section 3.3). Nontautomeric calculations were also done using a methodology described elsewhere.[33] The atomic charges and radii used in the PB calculations were essentially derived from the 43A1 force field of the GROMOS 96 distribution.[75,76] The charges were directly taken from the force field or adapted from it.[29,33,77] Several ways of deriving the atomic radii from the force field parameters were tested, as described in Section 2.5. All PB calculations consisted of finite-difference linear Poisson−Boltzmann calculations performed with the program MEAD (version 2.2.0),[78,79] using a temperature of 300 K, a molecular surface[80] defined with a solvent probe radius of 1.4 Å, and a Stern (ion exclusion) layer of 2.0 Å. The ionic strength was set to 50 mM for Barn and to 100 mM for the remaining proteins. A two-step focusing procedure[81] was used, with consecutive grid spacings of 1.0 and 0.25 Å. The dielectric constant of the solvent was 80, and the one of the protein was assigned values between 2 and 80 (2, 4, 6, 8, 10, 15, 20, 30, 50, and 80). The MC simulations for the tautomeric and nontautomeric cases were done using the programs PETIT[19,29] and MCRP,[19] respectively, which perform multisite titrations by performing trial changes of state of both single sites and pairs of sites (see refs 19, 29 for details and also the Appendix of ref 42). MC runs were performed at intervals of 0.2 pH units, each using at least 10^5 MC cycles, one cycle consisting of sequential state changes over all individual sites and also all pairs of sites with at least one interaction term above 2.0 pK_a units.

2.3. Redox Potential Calculations. The methodology used to compute the reduction potentials of redox sites consists of simulating the joint-binding equilibrium of electrons and protons, and was previously described.[19,37] It is essentially an extension of the methodology of Section 2.2, obtained by adding the redox groups of the protein to the global set of titrable sites and linking their reduction to an additional external parameter, the solution electrostatic potential E. In this way, the protein solution becomes characterized by the temperature, pH, and E. The general methodology and software used for the calculations, as well as the chosen parameters, are essentially those given in Section 2.2. As for the charges used for the heme groups, we have derived an improved charge set to substitute the Mulliken set used before.[33] Single-point B3LYP calculations were run with Gaussian 98[82] for the heme model compound, using the basis set 6-31G(2df) for iron and 6-31G(d) for non-iron atoms, followed by charge fitting to the electrostatic potential using RESP;[83] the structure was the one from heme I of DvHc3. The atoms used in the calculation and the resulting charges are given as Supporting Information. The reduction potential of our heme model compound is not directly related to known experimental values, although it is expected to be similar to the −220 mV value measured for an octapeptide bis-histidinyl derivative of the heme of cytochrome *c*.[84] Following the rationale used before,[19,37] we set the model reduction potential to zero and determined afterward the value that leads to best predictions under the current methodology; the suitability of this procedure will be analyzed in Section 3.6. The MC simulations were done at pH 7.0, except for Dmnc3, where the experimental 8.1 value was used. The E values were varied in intervals of 10 mV.

2.4. Tautomers for Carboxylic Sites. The inclusion of tautomers in carboxylic sites was previously done considering only the two identical syn positions for the proton, one per oxygen atom.[29] However, although the anti positions are in principle less stable, they should also contribute to the overall energetics of the protonation reaction. Furthermore, some particular local environments may actually turn an anti position into the most stable configuration. Thus, we have decided to include also anti positions for the protons in carboxylic sites. From simple considerations of chemical equilibrium, it directly follows that the general relation between the macroscopic pK_a of a carboxyl site and the microscopic pK_as of each of its syn and anti forms (a total of four microscopic pK_as) is

$$\begin{aligned} \mathrm{p}K_a(\mathrm{syn}) &= \mathrm{p}K_a(\mathrm{macro}) - \log[2/(1-a)] \\ \mathrm{p}K_a(\mathrm{anti}) &= \mathrm{p}K_a(\mathrm{macro}) - \log(2/a) \end{aligned} \tag{1}$$

where a is the fraction of the anti proton relative to the proton total (syn + anti) population. Several experimental evidences indicate that the free-energy difference between the syn and anti forms of acetic acid should be between 1 and 2 kcal/mol (see refs 85, 86 and references therein). High-level quantum chemical studies of acetic acid[86−89] confirm this result, many reporting values around 1.7 kcal/mol. We also performed calculations for acetic acid, at the B3LYP/6-311+G(d,p) level, using the solvation model combination IPCM//PCM for the solvated terms in the thermodynamic cycle, obtaining 1.77 kcal/mol for the free-energy difference; the calculations were done with Gaussian 98.[82] Given these evidences, we decided to adopt the value 1.7 kcal/mol for this free energy, which at 300 K corresponds to

$$a = [1 + \exp(1.7\ \mathrm{kcal\ mol^{-1}}/RT)]^{-1} = 0.055 \tag{2}$$

This value of a was used in eq 1 in order to derive the solution microscopic pK_as for the syn and anti forms of Asp, Glu, C-terminus, and propionates, using the macroscopic model compound values previously reported.[29]

2.5. Derivation of Atomic Radii. In previous works,[29,33,37,77] we used atomic radii equal to half of R_{min}, the distance corresponding to the minimum Lennard−Jones energy, computed using a pair of identical atoms; this is a common procedure for deriving atomic radii from a force field.[26,45,90−92] An alternative yielding smaller radii has also been used,[91,92] namely, using half of σ, the distance corresponding to zero Lennard−Jones energy ($\sigma = 2^{-1/6}R_{min}$). Another physically meaningful possibility could be to use half of the thermal radius R_{th}, usually defined in cavitation studies as the distance at which the Lennard−Jones energy raises RT above the minimum;[93,94] in general, the resulting $R_{th}/2$ radii are smaller than $\sigma/2$. To get even smaller radii, we decided to define a "doubly thermal" radius, corresponding to a raise of $2RT$ above the minimum; the resulting $R_{dth}/2$ values are more similar to the van der Waals radii taken from Pauling[95] or the PARSE parametrization.[96] Furthermore, it must be noted that the use of identical atoms when considering the Lennard−Jones interaction is not entirely consistent with the definition of a solute−solvent interface. A more sound alternative would be to use the Lennard−Jones energy between the atom of interest and the oxygen of the water molecule, OW,[97] and then to subtract from that the radius of OW using the one-half rule. We have thus decided to test the following atomic radii definitions for each atom type X:

$$\begin{aligned} r^*_{min}(\mathrm{X}) &= R_{min}(\mathrm{X\text{-}X})/2 \\ r_{min}(\mathrm{X}) &= R_{min}(\mathrm{X\text{-}OW}) - R_{min}(\mathrm{OW\text{-}OW})/2 \\ r_0(\mathrm{X}) &= \sigma(\mathrm{X\text{-}OW}) - \sigma(\mathrm{OW\text{-}OW})/2 \\ r_{dth}(\mathrm{X}) &= R_{dth}(\mathrm{X\text{-}OW}) - R_{dth}(\mathrm{OW\text{-}OW})/2 \end{aligned} \tag{3}$$

The first definition corresponds to the like-atom approach we

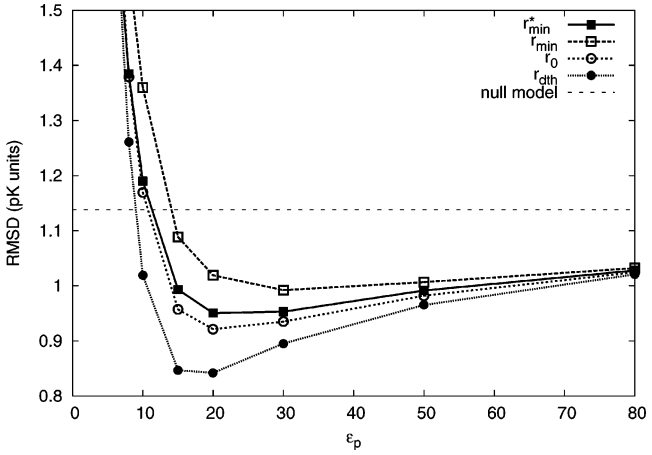

Figure 1. RMSD of predicted pK_as (proteins in Table 1), using the different choices of atomic radii given in eq 3. See the text for details.

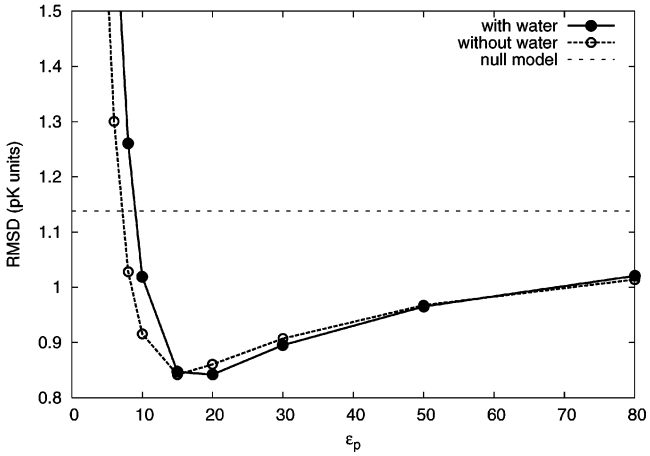

Figure 2. RMSD of predicted pK_as (proteins in Table 1), with and without inclusion of water molecules.

have used before, included for comparison despite its inconsistency with the existence of the protein−water interface. Strictly speaking, any of these definitions implies that the radii of atoms belonging to titrable sites will depend on the protonation state because the latter would usually affect the types of those atoms. However, this is not acceptable for the usual PB/MC methodology because a state-dependent dielectric boundary would invalidate the simple pairwise decomposition of the PB energies (the problem resembles that posed by using the nonlinear form of the Poisson−Boltzmann equation[98]). To avoid that, atomic radii are computed using "typical" states for the protonable sites: charged Asp, Glu, Lys, Arg, His, C- and N-termini, and propionates, the remaining being neutral. This issue does not arise for the heme groups because the same atom types, and thus the same atomic radii, are used for the oxidized or the reduced states.

3. Results and Discussion

3.1. Choice of Atomic Radii and Water Content. Before addressing our main issue, the combination of dielectric constants discussed in the Introduction, we performed some preliminary studies aimed at testing the use of alternative radii definitions and the inclusion of water molecules.

pK_a calculations were done with the protein set in Table 1 for several protein dielectric constants, ϵ_p, using the different radii definitions given in eq 3; water molecules were used in all these calculations. The resulting root-mean-squared deviation (RMSD) relative to the experimental pK_as is shown in Figure 1. Also shown is the RMSD for the so-called "null model", which corresponds to the use of the model compound pK_as for all sites.[25] All radii obtained using atom−water pairs are ordered as $r_{min} \geq r_0 \geq r_{dth}$; the like-atom r^*_{min} radii span values comparable to all the atom−water versions (although the r_{dth} radii are always smaller), depending on the atom. This indicates that the calculations generally benefit from the use of a smaller protein volume. The lowest RMSDs were obtained with $\epsilon_p = 20$ for all cases except the one using the r_{min} radii. Given these results, we decided to use the r_{dth} radii in all subsequent calculations.

As discussed in Section 2.2, crystallographic water molecules were included in the calculations whenever their individual relative accessibility is below 0.50. This cutoff value was previously found to give the lowest RMSD for the predicted pK_as.[29] Because the present study uses a different set of radii (the one chosen in the previous paragraph), we made a simple check to test whether that cutoff still lead to good results. Figure 2 shows the RMSD profiles for the pK_as obtained using the 0.50 cutoff and using no water molecules at all. The relation between both curves is similar to the one previously observed,[29] with best results obtained at $\epsilon_p = 20$ with waters and at $\epsilon_p = 15$ without waters. While the previous study indicated a better overall prediction with water inclusion, now both alternatives perform equally well. Given these results, we decided to keep the inclusion of water in all subsequent calculations, except in some of those analyzed in Section 3.3. In fact, we will end up concluding that the main results from our study do not seem to be much affected by the inclusion of water molecules.

3.2. Combination of Dielectric Constants. Several pK_a calculations for all sites of the proteins in Table 1 were performed using different combinations of the dielectric constants used to compute the individual and pairwise PB terms (ϵ_{ind} and ϵ_{pair}). The values used for the dielectric constants were the same as above (2, 4, 6, 8, 10, 15, 20, 30, 50, and 80), but not all possible combinations (100) were computed for all proteins. We started with all $\epsilon_{ind} = \epsilon_{pair}$ cases and then gradually added other heuristically chosen points as the study proceeded; a total of 42 combinations was thus generated. When necessary, interpolation/extrapolation was used for other combinations (see below).

The counterpart of a pK_a RMSD *curve* (as in Figures 1 and 2) is now an RMSD *surface* as a function of (ϵ_{ind}, ϵ_{pair}) combinations, which can be shown as a three-dimensional plot or, alternatively, as a set of two-dimensional contours. Such RMSD contours are shown in Figure 3 for all sites and also for different types of sites. The global prediction quality considering all sites (upper-left plot) is seen to be better for $\epsilon_{pair} \geq 10$ and $15 \leq \epsilon_{ind} \leq 30$. A wide plateau exists around the (20, 20) combination, where the RMSD takes the minimum value 0.84. This RMSD value is comparable with what is obtained with other single-structure PB models[25−28] and reasonably below the null model value (1.14). The RMSD raises sharply as either ϵ_{ind} or ϵ_{pair} approach very small values, as observed in calculations using a single ϵ_p (see Figures 1 and 2). These results indicate that, overall, there is no advantage in using (ϵ_{ind}, ϵ_{pair}) combinations; a single $\epsilon_p = 20$ still gives the best result (for the particular PB model used here).

When we consider the plots for the different site types shown in Figure 3, the Asp case is clearly the one for which the pK_a predictions are more difficult. The computed minimum is now 1.10 at (30, 30); the estimated lower value at (30, 80) is probably an artifact from the interpolation procedure. This difficulty in predicting the pK_as for Asp sites is probably related to the fact

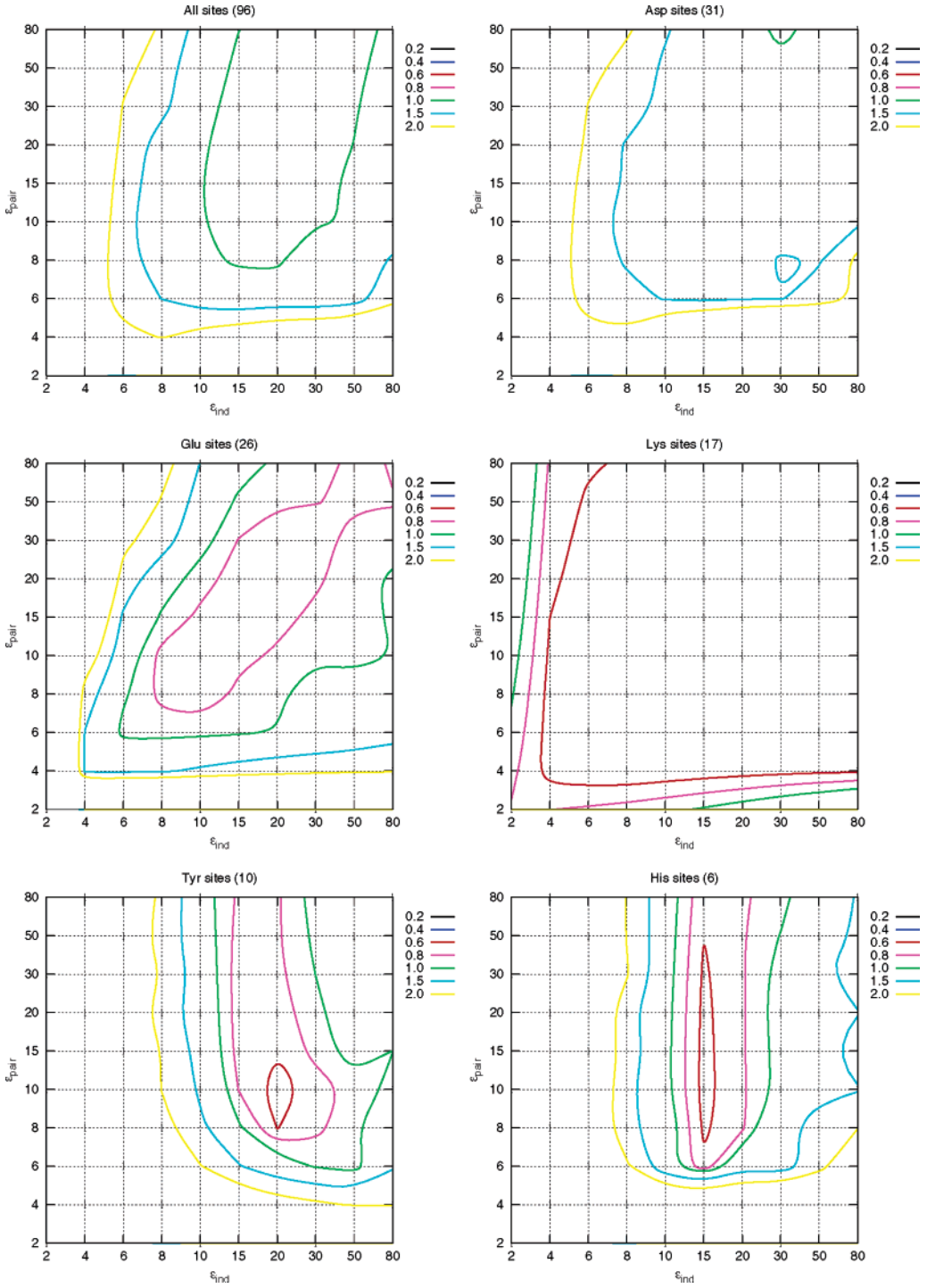

Figure 3. RMSD contours of predicted pK_as (proteins in Table 1) for all sites and specific site types. The total numbers of sites are shown in parentheses. Dots indicate (ϵ_{ind}, ϵ_{pair}) combinations for which actual calculations were performed. The remaining grid points are obtained by a Neville's second-order interpolation/extrapolation algorithm,[99] using the inverse-error-weighted average of the values obtained in each direction; for $\epsilon_{ind} = 2$ and $\epsilon_{pair} = 2$, only the corresponding direction was used because comparison with full-grid calculations (Section 3.3) showed this to give best results.

that, together with Tyr, this is the site type with a larger fraction of very shifted sites, i.e., sites whose pK_a value deviates considerably from the typical solution value (about half the sites have shifts of at least one pH unit). The pK_as for Glu sites are much more easy to predict. They show a least-RMSD basin consistent with the global results, although more oriented along the diagonal line. The Lys sites are even more easy to predict, as long as a value above 4 or 6 is used for both ϵ_{ind} and ϵ_{pair}. This is not really surprising, given the fact that the Lys sites are usually well-solvated and do not show significant pK_a shifts. The Tyr and His sites show similar RMSD contours, with least-RMSD basins more elongated toward high ϵ_{pair}. It is interesting to note that, although many Tyr sites have very large pK_a shifts, the overall prediction is, in this case, better than obtained for the Asp sites.

Some of the proteins analyzed are more representative than others, as can be seen from the RMSD contours in Figure 4. HEWL is the one whose contour pattern more closely resembles the global one in Figure 3 (upper-left plot). This is due to the fact that HEWL has representative amounts of different site

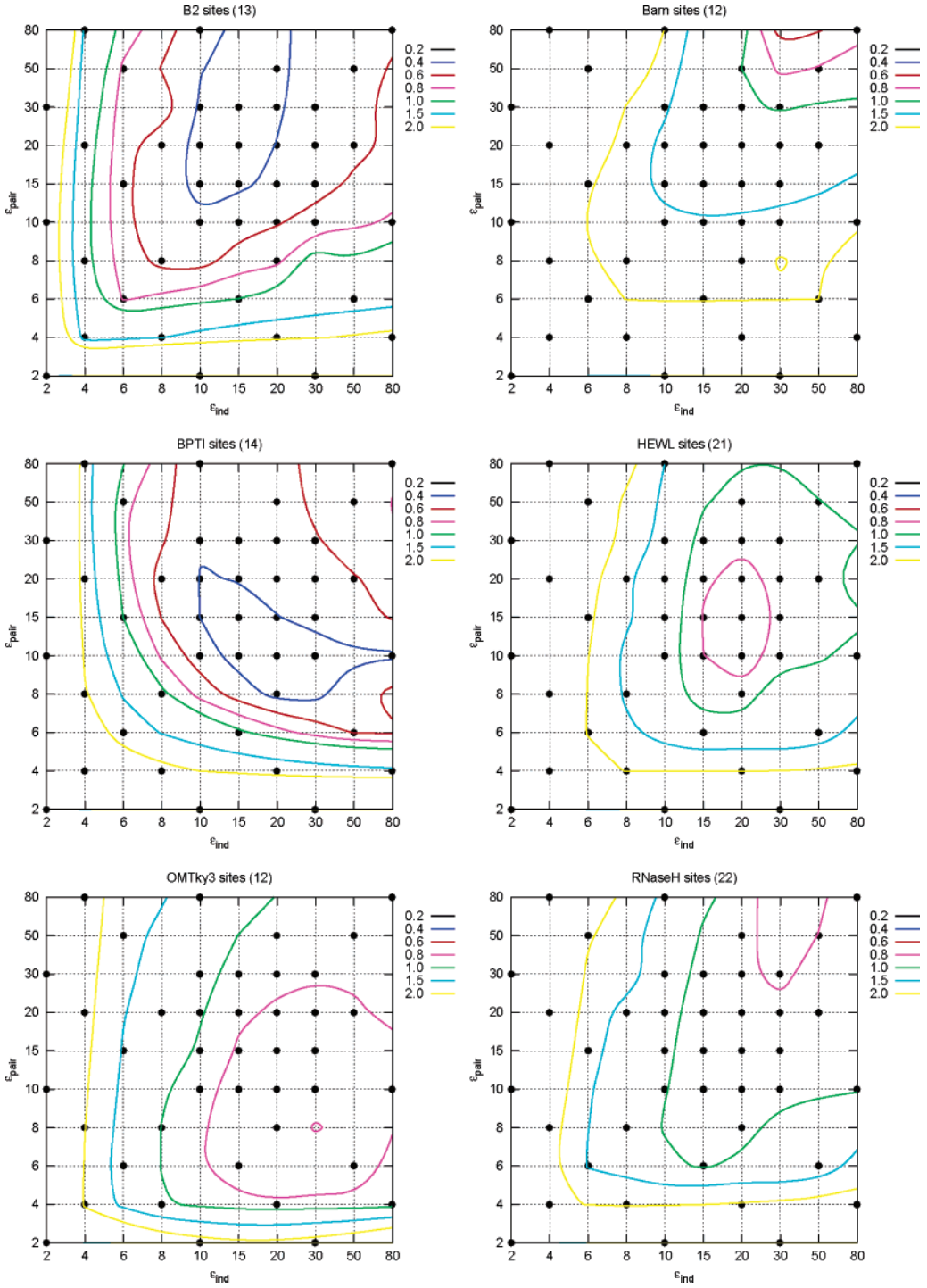

Figure 4. RMSD contours of predicted pK_as for different proteins in Table 1. See caption of Figure 3 for further details.

types and also of shifted sites, corroborating its traditional use as a benchmark for pK_a prediction methods. Two of the proteins, B2 and BPTI, are really more "well-behaved" than the average, yielding particularly low least-RMSD basins, whose shape is more or less consistent with the global one. These two proteins have few shifted sites and none of them to a very great extent (no shifts above 1.5 pH units). Predictions comparable to HEWL are obtained for OMTky3, whose least-RMSD basin is wider than that of HEWL, but shifted toward higher ϵ_{ind} and lower ϵ_{pair} values. In the case of RNaseH, the higher contours are similar to the global ones, but the bottom of the basin is shifted to high ϵ_{ind} and ϵ_{pair} values. This protein does not have many shifted sites, when compared to HEWL or OMTky3, but still its pK_as are more difficult to predict. Finally, Barn turns out to be the most difficult protein in terms of pK_a prediction, with all contours strongly shifted toward very high ϵ_{ind} and ϵ_{pair} values. This atypical profile (but not necessarily the high dielectric values; see below) may be due to the fact that this is the protein in our set with the largest amount of shifted sites, having half of them shifted at least 1 pH unit and one-third at least 1.5 pH units. These results show that different proteins can give somewhat different contours, pointing to the importance of using a reasonably large set of proteins. HEWL turns out to be the most representative protein of the set studied here. As for SubIn and Thrx, only a single site is compared with experimental data (see Section 2.1), being discussed below.

When we try to understand the variation in terms of RMSD contours observed among site types and proteins, one of the

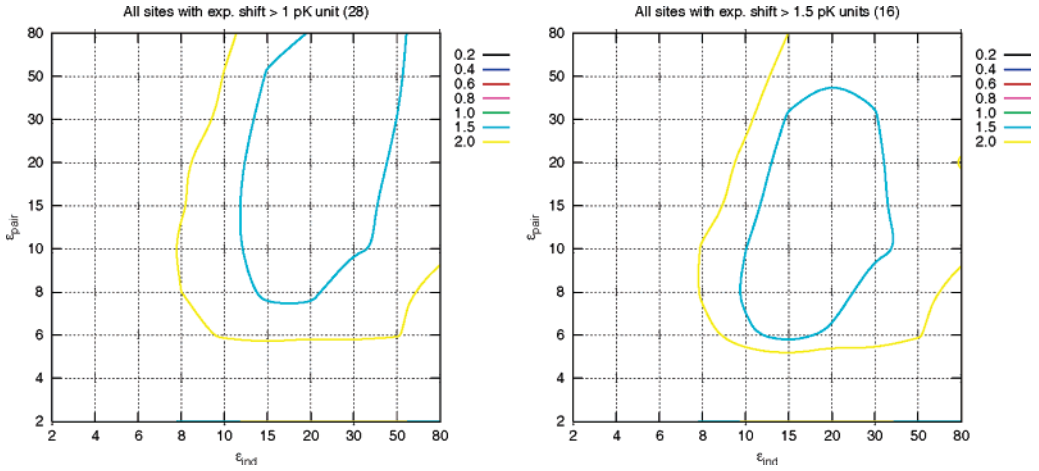

Figure 5. RMSD contours of predicted pK_as (proteins in Table 1) for all sites with large pK_a shifts. See caption of Figure 3 for further details.

most natural explanations would be perhaps the magnitude of the pK_a shifts in each analyzed category. As noted in the Introduction, it is often assumed that more shifted or buried sites are better treated using a low dielectric constant in PB models,[25,26,31,32] so maybe the particular mixture of shifted/buried and nonshifted/exposed sites in each category explains the different trends observed. However, RMSD contours for sites with large pK_a shifts (Figure 5) do not show significant differences relative to the global contours, although the predictions for sites with a shift of at least 1.5 pK_a units benefit from a slight decrease of ϵ_{pair}. As expected, more shifted pK_as tend to be more difficult to predict, yielding higher RMSD values. However, no clear systematic trend seems to exist for the overall contour shape of shifted sites. This can be further investigated by looking at the RMSD contours obtained for individual sites, a large diversity becoming immediately evident. This can be seen in Figure 6, which shows contours for some of the most shifted sites. Clearly, there is no direct relation between the contour pattern and either the magnitude or the signal of the pK_a shift. Some sites (SubIn-His43, HEWL-Tyr53, and RNaseH-Asp10) yield best predictions along a stripe of $\epsilon_{ind} \approx 20$ and a reasonably high ϵ_{pair} value, which could by itself suggest some kind of rule. However, the other cases shown do not at all confirm this rule. Thrx-Asp26 shows a similar stripe of low RMSD, but shifted toward $\epsilon_{ind} = 6$. Furthermore, no stripe is observed for either Barn-Asp93 or HEWL-Asp66. The pK_a of Barn-Asp93 turns out to be reasonably easy to predict, illustrating that strongly shifted sites are not necessarily difficult cases. In contrast, no reasonable pK_a can be predicted for HEWL-Asp66, regardless of the dielectric combination used. The curious aspect here is that both Barn-Asp93 and HEWL-Asp66 are, at first sight, under similar conditions: both are close to the protein surface (although the latter is more buried) and involved in strong interactions with positively charged groups (Arg and Lys sites). Yet, their RMSD contours are totally different. This may indicate that HEWL-Asp66 requires explicit account of structural reorganization; this issue is not the purpose of the present study, though (see Introduction).

To have a more complete view of the prediction trends over all sites, a direct scatter plot can be made of all experimental versus calculated pK_a shifts (the shifts are more appropriate for visual analysis than the actual pK_a values[30]). These plots are shown in Figure 7 for some (ϵ_{ind}, ϵ_{pair}) combinations, indicating also the accessibility of each site. We start by noting that points at the left and right extremes of the plots are generally very dark, meaning that sites displaying large experimental shifts tend to be buried. However, the converse is not generally true because many sites with small shifts are also buried. Thus, even though it is not our purpose to investigate this issue here, this shows that low solvent accessibility does not necessarily imply a shifted pK_a (because the charged form can be efficiently stabilized by surrounding preoriented dipoles[100]). Now, the plots show that sites more shifted and/or buried do not generally benefit from the use of low dielectric constants. The use of a single dielectric constant gives very bad predictions for $\epsilon_{ind} = \epsilon_{pair} = 4$, which are slightly improved for $\epsilon_{ind} = \epsilon_{pair} = 8$, and become reasonable good for $\epsilon_{ind} = \epsilon_{pair} = 20$; the predictions get worse for $\epsilon_{ind} = \epsilon_{pair} = 80$, approaching the null model, as expected. The results become very bad for the dielectric combinations (4, 20) and (20, 4). In particular, more shifted and/or buried sites do not benefit from the combination procedure, with the (4, 20) choice being particularly deleterious for them. An overview of the scatter plots for all determined (ϵ_{ind}, ϵ_{pair}) combinations can be seen in Figure 8 (no accessibility shown). The figure shows that none of the combinations manages to yield better global predictions than those of the (20, 20) region. In particular, very shifted sites do not seem to behave differently, with most of them giving the best predictions within this same region. Of course, these are the global trends, and exceptions do occur. By looking at the (20, 20) plot in Figure 7, one finds two particularly bad points: the rightmost one, corresponding to Thrx-Asp26, and the lowermost one, corresponding to Barn-Asp75. The large experimental shift of Thrx-Asp26 (3.5) probably results mainly from its pronounced internalization, suggesting that this effect may be underestimated by a too-high ϵ_{ind} in the (20, 20) combination. Indeed, using $\epsilon_{ind} = 6$ or 8 and $\epsilon_{pair} \geq 6$ gives a good prediction. The moderate experimental shift of Barn-Asp75 (-1.2) results from a close balance between two opposite effects because the site is buried (positive shift) but strongly interacting with two Arg sites (negative shift). The charge−charge interactions are probably excessive using the (20, 20) combination, and indeed, better results are obtained with higher ϵ_{pair}. The discussion of these two sites may seem to indicate that, after all, the usual approximate reasonings are correct. However, we cannot forget that these sites are exceptions to the (20, 20) plot, even among the shifted sites; the plain fact is that shifted sites do not generally benefit from the use of lower or higher values for either ϵ_{ind} or ϵ_{pair}. Our discussion of Thrx-Asp26 and Barn-Asp75 was intentionally made in order to illustrate how easy it is to "rescue" the approximate reasonings and ideas by using nonrepresentative sites. However, no matter how tempting it may be to devise some kind of trend in accordance with those ideas, there is no evidence at all supporting them.

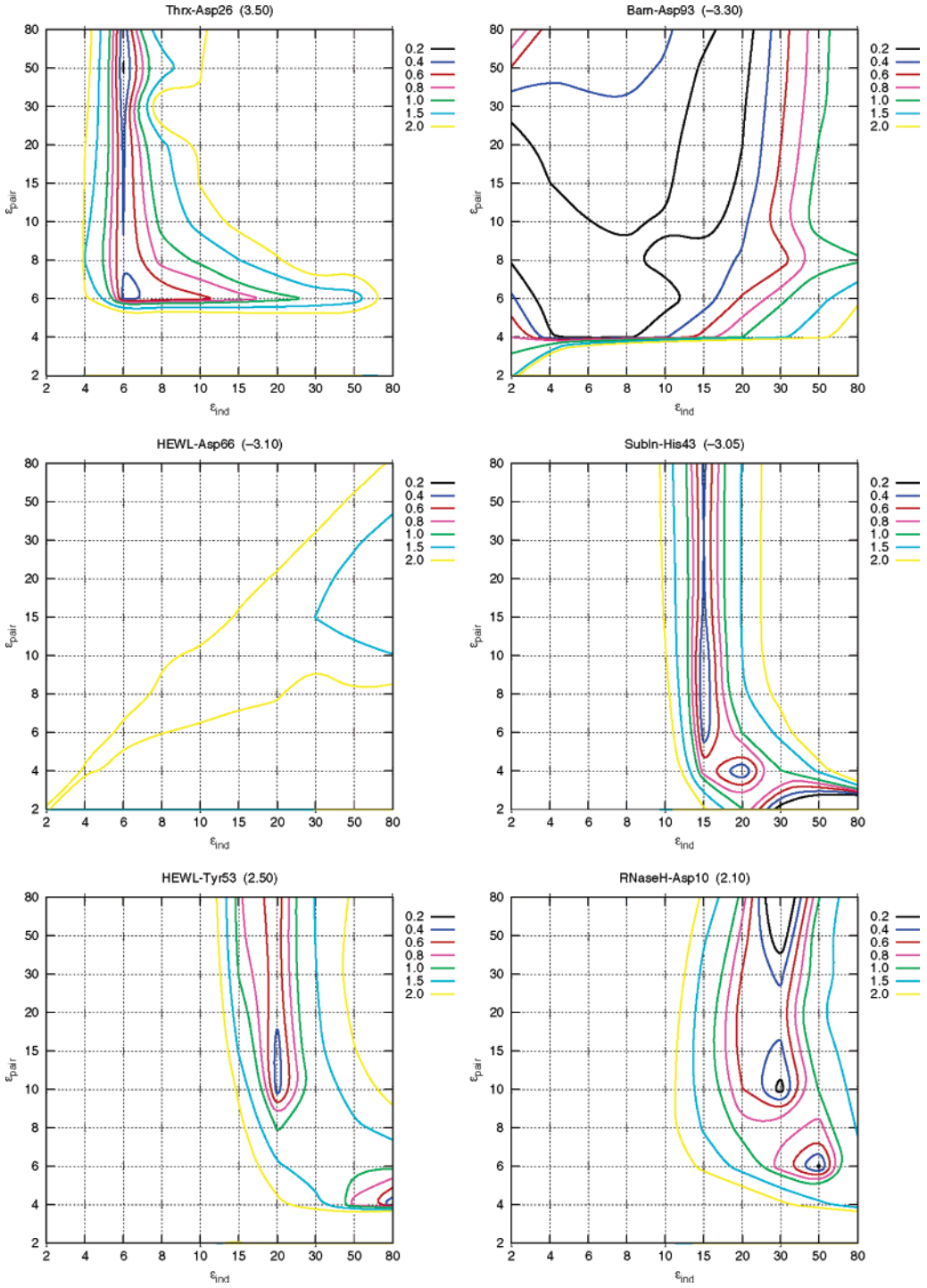

Figure 6. RMSD contours of predicted pK_as for some sites with very large experimental pK_a shifts, shown in parentheses. See caption of Figure 3 for further details.

In conclusion, the combination (20, 20) turns out to be generally the best one for "normal", shifted, and buried sites alike. Actually, very similar results are obtained with $15 \leq \epsilon_{ind} \leq 30$ and $\epsilon_{pair} \geq 10$. Therefore, these results go against the suggestion of using different ϵ_{ind} and ϵ_{pair} values,[9,30,46] and also against the common view that more shifted or buried sites benefit from the use of a low ϵ_p.[25,26,31,32] The simple procedure of using a single $\epsilon_p \approx 20$ to compute all PB energy terms is generally superior. These conclusions hold for the present methodology; their generalization to other PB models is discussed below.

3.3. Effect of Water, Tautomers, and Reference State. The above results on the use of (ϵ_{ind}, ϵ_{pair}) combinations refer to our particular methodology and cannot be immediately extended to PB models in general. Because a large diversity exists on the choice of parameters and algorithms (atomic radii and charges, charge/dielectric mapping, numerical method for solving Poisson−Boltzmann equation, etc.), one may ask whether different choices would lead to different results. Naturally, it is computationally unfeasible to examine all possible choices and their combinations. The use of different radii was investigated in Section 3.1, and although different minima were obtained for different radii, no significant changes were found for the shape of the curves; this will not be further analyzed here. There are, however, two other aspects of our methodology that are not standard in PB models and thus deserve further consider-

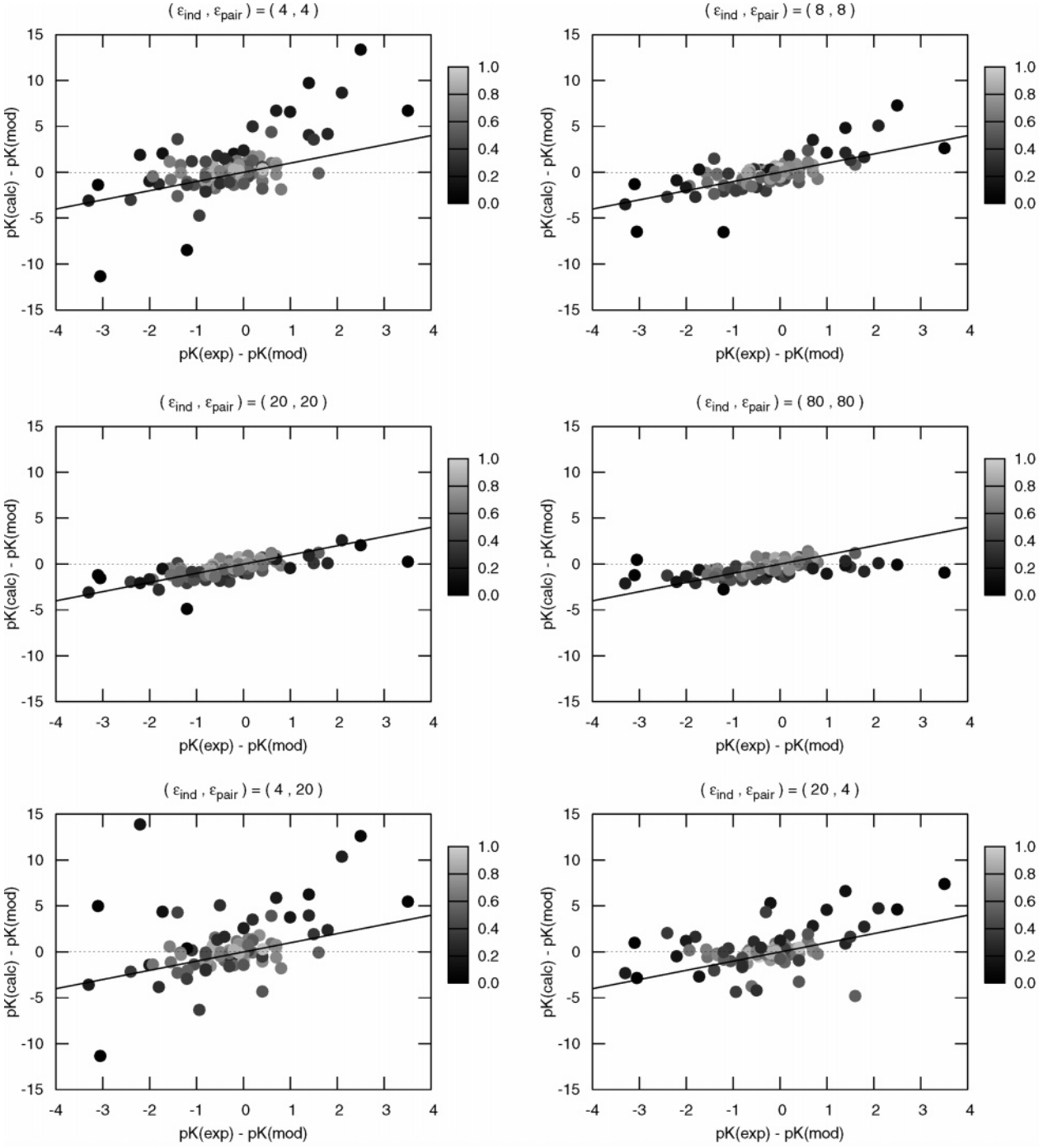

Figure 7. Experimental vs calculated pK_a shifts for all sites (proteins in Table 1), using some (ϵ_{ind}, ϵ_{pair}) combinations. Solid lines correspond to the exact prediction and dashed lines to the null model. Gray shading indicates the relative accessibility of the site. The accessibility of each site was computed using only the atoms whose charge depends on the protonation state.

ation: the inclusion of tautomers for titrable protons and of explicit water molecules (as rotamers). The consideration of proton tautomerism can be seen as a partial (although very limited) inclusion of reorganization effects, and therefore, one may expect that a lower dielectric constant would be needed. Somewhat unexpectedly, this was not observed using a single ϵ_p value,[29] but a different scenario may arise when ϵ_{ind} and ϵ_{pair} are considered separately. As for the inclusion of explicit water molecules, it corresponds to the "appropriation" of part of the high-dielectric solvent region by the low-dielectric cavity and may thus be expected to affect how the results depend on the dielectric constant of this cavity. Again, no such effect was observed using a single ϵ_p value[29] (see also Figure 2), but the situation may be different when (ϵ_{ind}, ϵ_{pair}) combinations are considered.

A third aspect that is not standard in PB models is the use of a charged reference state. As discussed in more detail in Section 3.5, the usual reference state is a fully neutral one, both in the PB and PDLD/S models. Although the choice of reference state is computationally irrelevant if a single ϵ_p is used (see Section 3.5), that is not the case when different (ϵ_{ind}, ϵ_{pair}) combinations are considered because the terms which are "mixed" upon changing reference state are now dependent on the particular combination. Furthermore, the physical argument given by

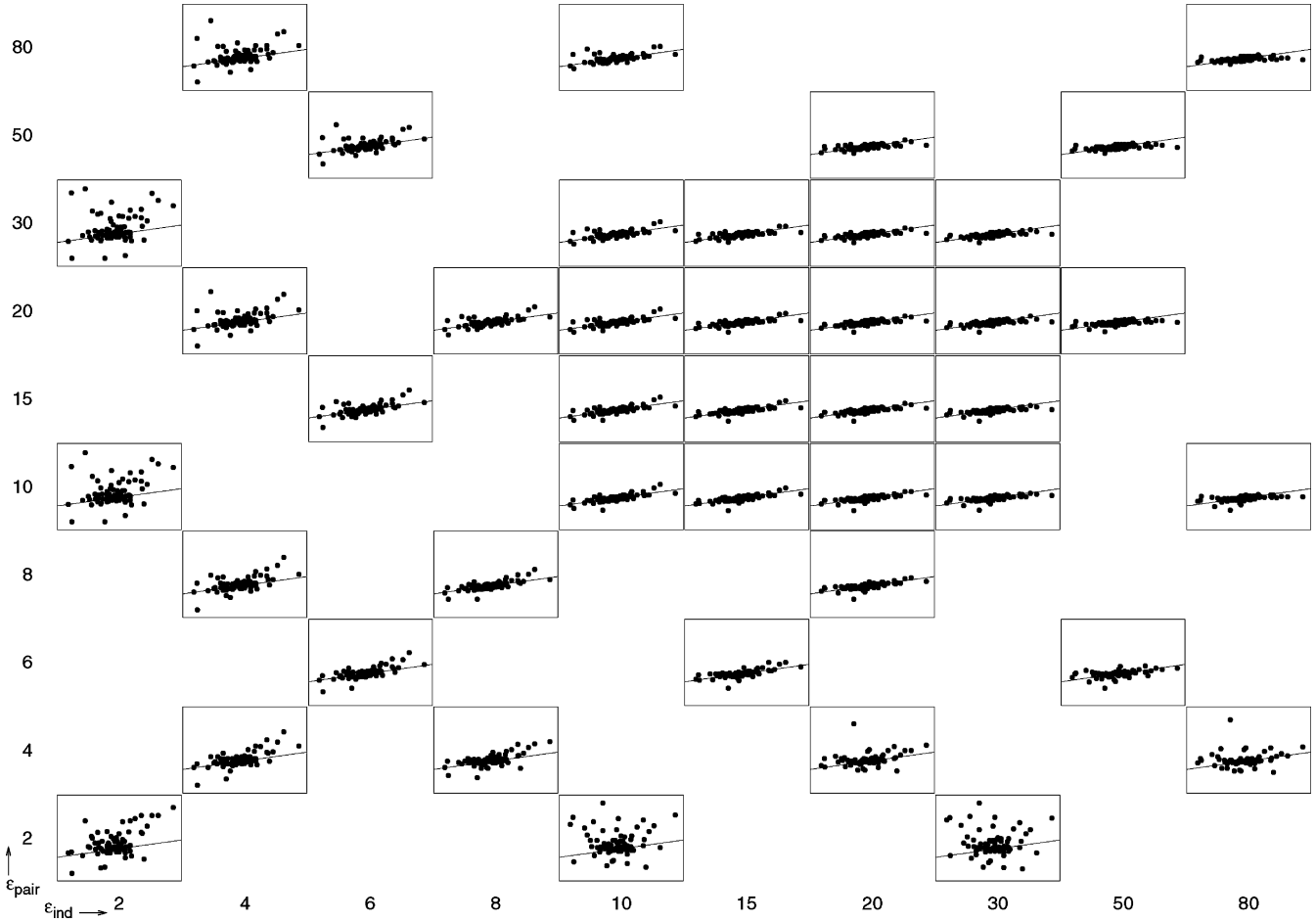

Figure 8. Experimental vs calculated pK_a shifts (proteins in Table 1) for all determined (ϵ_{ind}, ϵ_{pair}) combinations. The *x*- and *y*-axes of each small plot are the same as in Figure 7, their ranges being respectively [−4, 4] and [−15, 25].

Warshel and co-workers for the use of different ϵ_{ind} and ϵ_{pair} values originally refers to the use of a fully neutral reference state (see Section 3.4). Therefore, the (ϵ_{ind}, ϵ_{pair}) dependences obtained in Section 3.2 are not necessarily the ones that would be obtained using a neutral reference state and thus do not directly rule out Warshel's suggestion.

A study of these three aspects (proton tautomers, explicit water, reference state) for the complete protein set in Table 1 would amount to a huge computational effort. Thus, we decided instead to use HEWL for this purpose, which, as seen above, seems to be the most representative protein of the whole set; all 100 possible (ϵ_{ind}, ϵ_{pair}) combinations were analyzed. Figure 9 shows RMSD contours for HEWL using four different setups, obtained by gradually introducing cumulative changes to the full PB/MC methodology used in Section 3.2: (a) full methodology (with water molecules and tautomers, using a fully charged reference state); (b) no inclusion of water molecules; (c) no inclusion of water molecules or tautomers; (d) no inclusion of water moleculesnor tautomers, using a fully neutral reference state. The immediate observation is that no drastic changes occur, the overall contour pattern being quite similar in all cases. The removal of water molecules leads to an overall small displacement toward lower ϵ_{ind} and ϵ_{pair} values. In particular, this has the effect of displacing the RMSD minimum from (20, 15) to (15, 10). A further displacement of ϵ_{pair} takes place when proton tautomers are also removed, bringing the minimum to (15, 4). Hence, the removal of both water molecules and proton tautomers (two nonstandard aspects in PB models) leads to a best combination of dielectrics for which $\epsilon_{ind} > \epsilon_{pair}$, i.e., the opposite of what was suggested by Warshel and co-workers. On the other hand, the RMSD value depends much more strongly on ϵ_{ind} than on ϵ_{pair}, so that one may say that Warshel's statement that any sufficiently large ϵ_{pair} should work,[30,101] is roughly confirmed. As noted in the previous paragraph, it could be suspected that these results were due to our choice of the charged reference state. However, the RMSD contours for the case of the neutral reference state show that this is actually not the case. The overall contour pattern is very similar to the other three plots, although the minimum is higher in this case; this minimum is found at (20, 8), i.e., again for $\epsilon_{ind} > \epsilon_{pair}$. Once more, the RMSD value depends much more strongly on ϵ_{ind} than on ϵ_{pair}. Furthermore, the scatter plots of the experimental versus calculated pK_a shifts (given as Supporting Information) do not reveal any particular trend for very shifted sites. Thus, as for the charged reference state, we find that more shifted sites do not particularly benefit from the use of lower ϵ_{ind} (or ϵ_{pair}) values. Overall, "normal", shifted, and buried sites in HEWL are all better predicted with a high ϵ_{ind} and a not-too-low ϵ_{pair}.

The removal of water molecules and proton tautomers, as well as the use of a neutral reference state, makes our methodology very similar to other, more standard PB-based methods for pK_a prediction. Thus, the results in the present section suggest that the main conclusions obtained in Section 3.2 should be largely applicable to PB models in general (using a single, rigid structure; see Introduction). In particular, we find again that the data presented here does not support the ideas of using $\epsilon_{ind} < \epsilon_{pair}$[9,30,46] or of using a low ϵ_p for shifted/buried sites.[25,26,31,32] The main difference with respect to Section 3.2

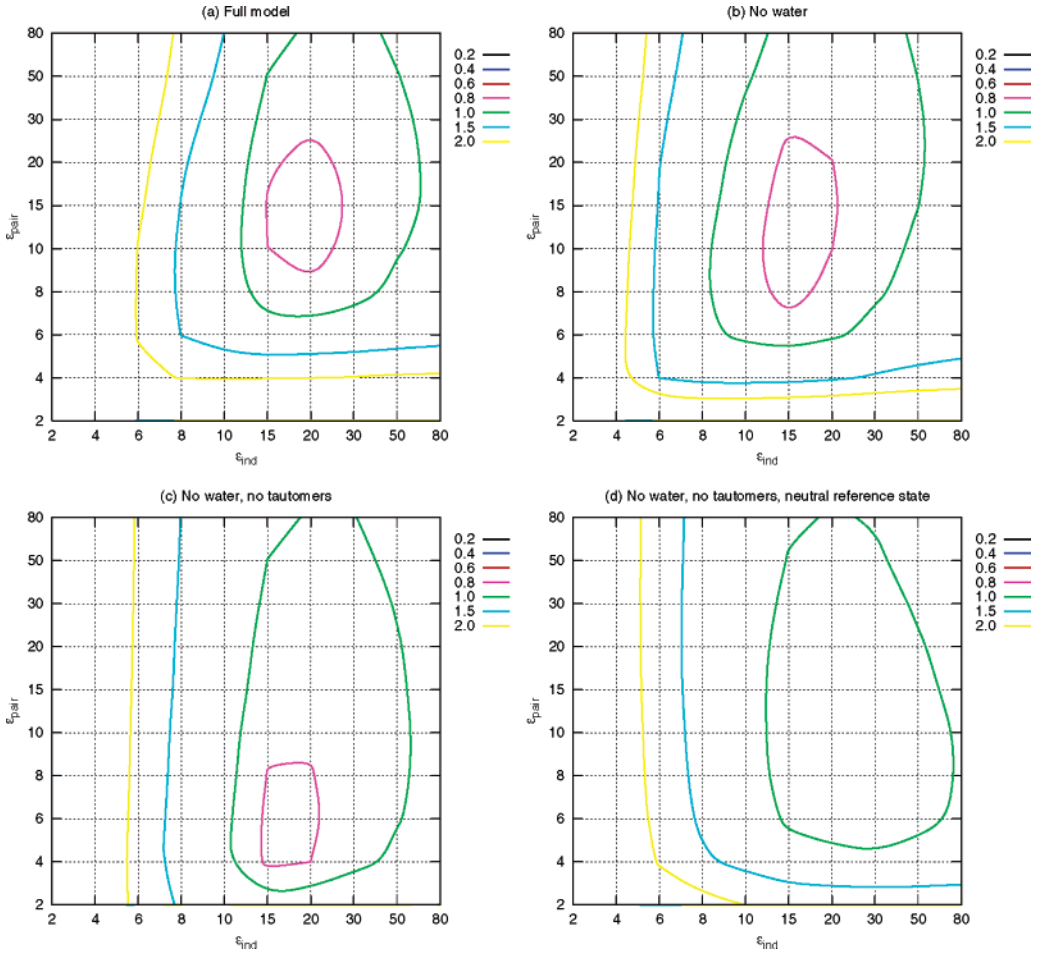

Figure 9. RMSD contours of predicted pK_as for HEWL, obtained with the full PB/MC model (including tautomers and water molecules), and also with elimination of water molecules, and/or tautomers, and/or change of reference state.

is that the more standard setup produces the best results with $\epsilon_{ind} > \epsilon_{pair}$ rather than with $\epsilon_{ind} = \epsilon_{pair}$.

3.4. Reasons for Apparent Discrepancy with PDLD/S Studies. As discussed in the Introduction, the suggestion for using $\epsilon_{ind} < \epsilon_{pair}$ in PB ionization studies followed from results obtained in PDLD/S studies.[9,30,46] Therefore, given the similarity of the PB and PDLD/S models, it is important to try to understand the apparent discrepancy of results between those studies and the present one. We start by briefly discussing the dielectric parameters used in the PDLD/S model (a detailed exposition can be found in refs 8, 30) and then proceed to investigate the origin of the discrepancy.

In the PDLD/S model, the region corresponding to the protein molecule is assigned a dielectric constant ϵ_p, which "scales" the PDLD value computed for the ionization free energy of a site in the protein environment (e.g., see Figure 5 of ref 8). This ionization free energy is computed with all other sites in the neutral state, thus corresponding to the intrinsic pK_a originally introduced by Tanford and Kirkwood.[2] To obtain the site−site interactions[102] arising when multiple sites are charged, similar PDLD/S calculations can be done using the pair of sites (but see next paragraph). A detailed analysis of the "scaling" argument used in the PDLD/S method (see Appendix of ref 30) strongly suggests that the role of ϵ_p is very nearly the same as in the PB model; thus, it seems reasonable to use the same notation, ϵ_p, to designate the protein "dielectric constant" in both models, as anticipated in the Introduction. Hence, the PDLD/S and PB approaches compute the same individual and pairwise free energies in an essentially equivalent way.

Although the consideration of ϵ_p is sufficient to perform any PDLD/S calculations, a simplifying approximation is often used. From both experimental and theoretical results (e.g., see ref 101 and references therein), it is usually a reasonable approximation to write the interaction as having a Coulombic form, whose "effective dielectric constant", ϵ_{eff}, can be obtained by approximate equations[101,103] or simply assigned a large value (e.g., 40 or higher[30,101]). Therefore, it is the usual practice in PDLD/S studies to compute the individual terms as described in the previous paragraph (using ϵ_p), and the site−site interactions with a simple Coulombic expression (using ϵ_{eff}), except for very strongly interacting pairs of sites. In particular, several studies[9,30,46] comparing PDLD-type models at different levels of conformational treatment, showed that the best performance of the simple PDLD/S model is usually obtained with a low ϵ_p to compute the intrinsic pK_as and a high ϵ_{eff} to compute the site−site interactions. Given those results and the similarity between the PDLD/S and PB models, the authors have suggested[9,30,46] that calculations using PB models should use a low ϵ_p to compute the intrinsic pK_as and a high ϵ_p to compute the pairwise interactions between charged sites (i.e., low ϵ_{ind} and high ϵ_{pair}, in the present notation). The assumption behind this suggestion is that the shielding between interacting sites is roughly given by the dielectric constant of the environment between both sites, so that $\epsilon_{eff} \approx \epsilon_{pair}$. However, this assumption neglects the fact that ϵ_{eff} reflects the joint effect of many factors: the dielectric constants (and/or induced dipole grids) assigned to both the protein and solvent regions, the shape of the interface between them, and the remaining protein charges and dipoles (some of

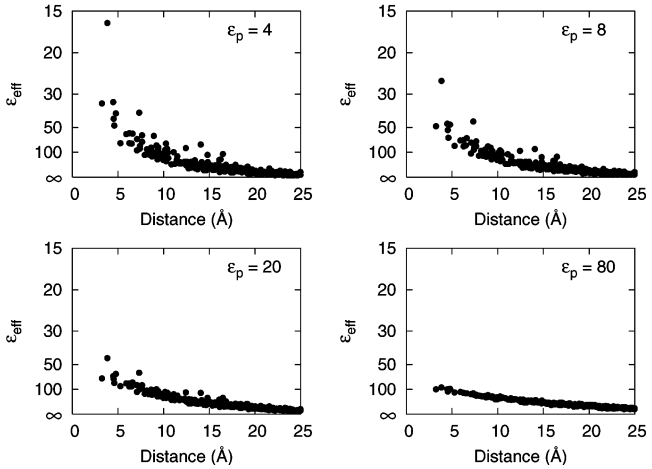

Figure 10. Computed ϵ_{eff} for site−site pairs in HEWL, as a function of the distance between the site centers, for different ϵ_p values. The center of a site is here defined as the average of the coordinates of its atoms, weighted by the charge change of each atom upon protonation; this yields a point near the most affected atoms. The ϵ_{eff} values are plotted in an inverse ($1/\epsilon_{eff}$) scale.

them preoriented in a stabilizing manner[100]). Given the close similarity between the PDLD/S and PB models, we should expect that $\epsilon_{eff} \approx \epsilon_{pair}$ should hold for both or none of them. If it holds, it would imply that a full PDLD/S calculation (i.e., not resorting to approximate ϵ_{eff} values) of a pairwise interaction should also use a high ϵ_{pair}. But this would contradict the fact that single-ϵ_p PDLD/S calulations (i.e., with $\epsilon_p = \epsilon_{ind} = \epsilon_{pair}$) give rise to high ϵ_{eff} values when a low ϵ_p is used.[46] In other words, the parallel between PDLD/S and PB models should in principle extend to the relation between ϵ_p and ϵ_{eff}, and what applies to one model should also apply to the other.

To investigate this parallel, we decided to check if the relation between ϵ_p and ϵ_{eff} in PB calculations is similar to the one found in PDLD/S models. Because the use of proton tautomers complicates the definition and computation of the site−site interactions, we used the PB calculations with no inclusion of tautomers or explicit water molecules. Furthermore, the fully neutral reference state was chosen,[104] and the same ϵ_p value was used for both individual and pairwise terms (i.e., $\epsilon_{ind} = \epsilon_{pair}$), Figure 9d. As in Section 3.3, we used HEWL, the most representative protein in our set. As usual,[101,103] we defined ϵ_{eff} by means of the Coulombic equation $W_{ij} = [\epsilon_{eff} r_{ij}]^{-1}$ (in Gaussian units), where W_{ij} is the pairwise interaction between sites i and j, and r_{ij} is the distance between them. Figure 10 shows plots of ϵ_{eff} as a function of the distance between the sites for different ϵ_p values. It is found that ϵ_{eff} increases with the distance between the sites, as experimentally observed,[101] being generally higher than ϵ_p and much higher in the case of low ϵ_p. Even for pairs of buried sites, it is wrong to assume that $\epsilon_{eff} \approx \epsilon_p$; e.g., Glu35 and Asp52 are deeply buried into HEWL, within 7 Å of each other, and yet they yield $\epsilon_{eff} = 39$, even for $\epsilon_p = 4$. This shows that the effect of preoriented dipoles[100] and/or of the solvent dielectric constant can be quite large. Overall, we find that a PB model with a single ϵ_p displays the same ϵ_{eff} behavior found from experimental and PDLD/S studies.[30,101] In particular, the use of $\epsilon_p = 20$ agrees well with the simple approach of using $\epsilon_{eff} \geq 40$.[30,101]

In conclusion, once the relation between ϵ_{eff} and ϵ_{pair} is clarified, it becomes evident that the discrepancy between our results and those of PDLD/S studies was only apparent. It resulted from the assumption of considering $\epsilon_{eff} \approx \epsilon_{pair}$, which turns out to be unjustified. In fact, the calculation of ϵ_{eff} values shows that there is excellent agreement between both models.

3.5. Physical Meaning of Using Dielectric Combinations. As shown in the previous section, the eventual advantage of using $\epsilon_{ind} < \epsilon_{pair}$ does not actually follow from the PDLD/S studies, but that alone does not necessarily imply that such a procedure should be worse than using $\epsilon_{ind} = \epsilon_{pair}$. The conclusive evidence comes only from the results presented above, which show that using $\epsilon_{ind} \neq \epsilon_{pair}$ brings no improvement to our usual methodology (Section 3.2) and that using $\epsilon_{ind} > \epsilon_{pair}$ (the opposite of the original suggestion) brings some improvement to a more standard PB methodology (Section 3.3). The physical reasons for these results are not obvious, specially given the physically vague nature of ϵ_p in PB models, noted in the Introduction. Nonetheless, we briefly discuss in this section some approximate physical arguments related with the use of $\epsilon_{ind} \neq \epsilon_{pair}$.

It may be argued[9,30,46] that the structural relaxation caused by charging a site in an otherwise neutral protein, due to charge-dipole interactions, is smaller than one involving charge−charge interactions and thus should need a smaller "scaling" parameter, meaning that we should use a low ϵ_{ind} to compute intrinsic pK_as and a higher ϵ_{pair} to compute the W_{ij} values. On the other hand, if we switch from the fully neutral reference state to the fully charged one, a similar argument would lead us to conclude that a high ϵ_{ind} and a low ϵ_{pair} should be used.[37] Thus, it seems that the "optimal" (ϵ_{ind}, ϵ_{pair}) combination depends on the reference state, as actually seen by comparing the plots (c) and (d) in Figure 9. This is actually to be expected from a formal viewpoint, following from the fact that choosing a reference state corresponds to choosing a way of grouping "low-level" PB terms (i.e., Green functions) together: while the grouping method is irrelevant if the same ϵ_p is used to compute afterward all resulting groups (e.g., note the identical diagonals of the plots (c) and (d) in Figure 9), it cannot be irrelevant if different ϵ_p values are used to compute those groups. Unfortunately, this line of approach leads to a somewhat intricate situation, because it is not at all obvious which choice is more sound in terms of physical interpretation.

There is a perhaps more fruitful line of approach,[37] which starts from the fact that the individual and pairwise PB terms are ultimately used to compute the free-energy difference between protonation states actually occurring in the protein. More exactly, they are used to compute the free-energy change of each trial move attempted during the MC simulation.[105] Therefore, when a particular trial move is considered, any physically meaningful parameters of the model should be made appropriate for the initial and final states of the move. In other words, ϵ_{ind} and ϵ_{pair} should reflect those initial and final states, and not the fully neutral, the fully charged, or any other reference state, regardless of its formal convenience. Because a typical protein would have many charged groups within a considerable pH range, any protonation changes will be largely associated with charge−charge interactions, which leads to the conclusion that a high ϵ_p should probably be used to compute *all* protonation free-energy terms in PB models. This may be the reason for the generally good performance of PB models using a single high ϵ_p value. Anyway, that could "explain" the results in Section 3.2, but not those in Section 3.3. Clearly, the physical meaning of using $\epsilon_{ind} \neq \epsilon_{pair}$ is not easy to grasp.

3.6. Redox Titrations. As a final part of our study, we investigated the effect of the dielectric constant on the calculation of reduction potentials, a study which, to our knowledge, has not been previously reported. No combinations of dielectric constants were analyzed, meaning that only the dependence on a single ϵ_p was considered. It must be noted that, in this case,

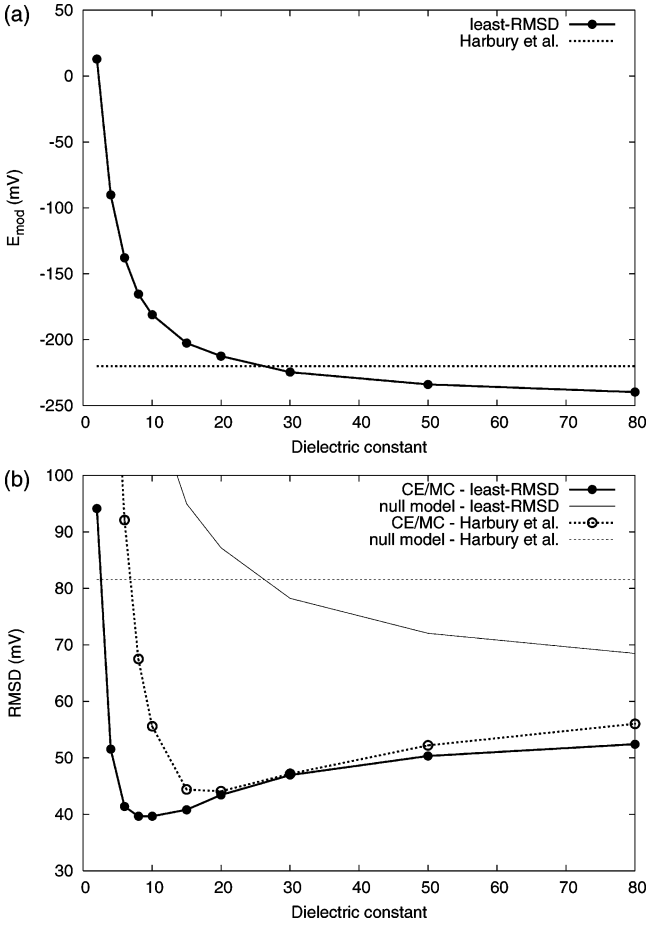

Figure 11. (a) Least-RMSD E_{mod} values computed for each ϵ_p, showing also the Harbury et al. value. (b) RMSD of predicted reduction potentials for the proteins in Table 2.

the performance of the method can only be judged from the predicted reduction potentials because no experimental pK_a values are available for the set of redox proteins in Table 2. Therefore, the analysis in this section is biased toward redox groups, and thus necessarily incomplete.

As discussed in Section 2.3, the reduction potential for the heme model compound, E_{mod}, is expected to be similar (but not necessarily equal) to the value measured by Harbury et al. for a related compound,[84] but it can be chosen after the calculations as the one giving the least RMSD between the predicted and experimental reduction potentials.[37,106] In Figure 11a, we show both the Harbury et al. and the least-RMSD values for E_{mod} as a function of the protein dielectric constant, ϵ_p. Even considering the differences between our heme model compound and the octapeptide bis-histidinyl derivative used by Harbury et al., it is clear that the least-RMSD procedure gives unrealistic E_{mod} values at low ϵ_p, suggesting that those values are trying to compensate for some kind of systematic error. The origin of this error is actually easy to detect: at low ϵ_p, the heme−heme interactions are so high that all redox titration curves get displaced to very low (very negative) potential values, requiring a very high E_{mod} in order to yield a low RMSD. This means that if one of these high E_{mod} values is used for a system where such interactions are weaker or absent, a too-high reduction potential will generally be obtained Therefore, the use of the least-RMSD criterion to get E_{mod} values is probably not very realistic at low ϵ_p. Figure 11b shows the RMSD curves obtained for the computed reduction potentials using both the least-RMSD E_{mod} values and the Harbury et al. value; the corresponding null model curves are also shown. The Harbury et al. E_{mod} gives

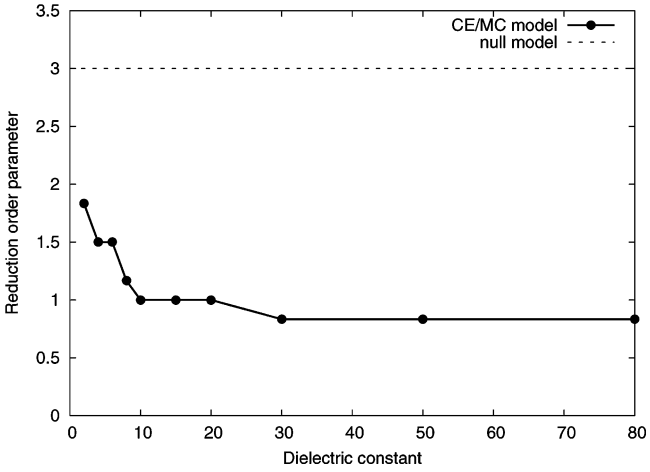

Figure 12. Reduction order parameter for the proteins in Table 2.

the lowest RMSD at $\epsilon_p = 15$ and 20, the profile closely resembling the global one obtained for the pK_a calculations (the minimum, 44 mV, corresponds to 0.74 pH units). The least-RMSD E_{mod} values give the lowest RMSD at $\epsilon_p = 10$, in a region where the computed E_{mod} starts to become arguable and, perhaps even more importantly, where pK_as start to become poorly predicted (see r_{dth} curve in Figure 1). On the other hand, at $\epsilon_p = 20$, where the pK_a prediction is better, the use of the least-RMSD E_{mod} value (−212 mV) gives an RMSD that is still quite low (43 mV) and identical to the one obtained using the Harbury et al. E_{mod} value. Therefore, the present results show that a consistent prediction of reduction potentials and pK_as can be obtained using $E_{mod} = -212$ mV, the least-RMSD value for our model compound.

A usually important issue when studying redox proteins is whether the reduction order can be correctly predicted or not. If the redox sites have reduction potentials well set apart, a good predicted order would naturally follow from a good prediction of those potentials. However, when the reduction potentials cluster in a narrow range, as is the case for some of the c_3 cytochromes used here, it may be argued that a reasonable prediction of reduction potentials could be obtained without necessarily yielding a good predicted reduction order. To remove any doubts in this respect, we decided to perform one further analysis of our results. The four heme sites of a given cytochrome c_3 can be arranged in six possible pairs, each having a definite reduction order relation. For example, the experimental reduction order of DvHc3 at pH 7.0 can be written as IV > I > II > III,[67] which corresponds to the six order relations IV > I, IV > II, IV > III, I > II, I > III, and II > III; when a PB/MC calculation is done for DvHc3, each of these six order relations may be correct or inverted. We defined a simple reduction order parameter as the number of pairs whose predicted order is wrong; a value of 0 means perfect prediction, while a value of 6 means the worst prediction (inverted reduction order). A completely random guess of the reduction order would consist of choosing any of the 24 possible orders at random, whose average order parameter is found to be 3 by direct calculation; such an average random guess would be called the "null model", even though the meaning here is slightly different from the one used throughout the rest of this study. Figure 12 shows the average order parameter computed over all the proteins of the redox set (the cytochromes c_3). For $\epsilon_p = 10$, 15, and 20, only one reduction order is wrongly predicted on average; e.g., the reduction order predicted for DvHc3 at $\epsilon_p = 20$ is IV > I > III > II, corresponding to a wrong order for hemes II and III. A slight further decrease of the order parameter

is observed for still higher ϵ_p values. These results clearly show that the prediction of reduction orders is quite good using $\epsilon_p = 15$ or 20, even for proteins with redox sites having such similar reduction potentials as cytochromes c_3.

4. Concluding Remarks

The present work investigates some factors affecting the performance of Poisson−Boltzmann (PB) methods to study protein ionization equilibrium, giving central importance to the use of different dielectric constants to compute individual (site) and pairwise (site−site) terms ($\epsilon_{ind} \neq \epsilon_{pair}$). The results in Section 3.2 show that, at least for the particular PB methodology adopted here, there is no advantage in such a procedure, yielding instead a region of best prediction around $\epsilon_{ind} = \epsilon_{pair} = 20$. Most importantly, this region is the same for buried/shifted or exposed/nonshifted sites alike. In particular, the use of $\epsilon_{ind} < \epsilon_{pair}$, suggested by Warshel and co-workers[9,30,46] to improve the prediction of significantly shifted pK_as, is not better than the use of a single $\epsilon_p = 20$. These results also go against the widespread view that more shifted or buried sites benefit from the use of low dielectric constants[25,26,31,32] because, as just noted, no such trend is observed.

The results in Section 3.3 show that these main conclusions are not significantly affected by our inclusion of water molecules or proton tautomers or by our use of a charged reference state, all nonstandard procedures in PB methodologies. Interestingly, the elimination of these nonstandard procedures tends to displace the region of best pK_a predictions toward a region with $\epsilon_{ind} > \epsilon_{pair}$, precisely the opposite of Warshel's suggestion. Again, no special trends are observed for shifted/buried sites.

Because the original proposal of using $\epsilon_{ind} < \epsilon_{pair}$ followed from studies using the PDLD/S model,[9,30,46] the present results may seem to indicate a poor parallel between the latter and the PB model, contrary to what is generally assumed. However, we show in Section 3.4 that this discrepancy is only apparent, resulting from the assumption that the so-called effective dielectric constant ϵ_{eff} is essentially the same as ϵ_{pair}, which turns out to be unjustified. In fact, the analysis of PB-derived ϵ_{eff} values reveals that even interactions involving buried sites have associated ϵ_{eff} values much higher than the ϵ_p used for the protein region, showing that the effect of preoriented dipoles[100] and/or of the solvent dielectric constant can be very significant, even in the protein interior. Hence, the use of a single ϵ_p by a PB model is perfectly consistent with the full PDLD/S model (i.e., not resorting to approximate ϵ_{eff} values). The distance-dependent profiles we obtain for ϵ_{eff} are actually in excellent agreement with both PDLD/S and experimental studies, strengthening the parallel between both models. The only real difference between the present results and those of Warshel and co-workers is that our best-prediction ϵ_p (20) is higher than theirs (4−6). Interestingly, this is consistent with a PB-LRA study we have recently performed,[40] which required a higher ϵ_p than similar PDLD/S-LRA studies.[30] These differences are probably due to the fact that both models are not strictly equivalent and several important differences exist between them (e.g., modeling the solvent either as a high dielectric medium or as a set of Langevin dipoles). Overall, our PB methodology seems to require a higher ϵ_p value than the "equivalent" PDLD/S model.

The present results tend to reinforce the view of ϵ_p as just an empirical parameter that attempts to account for all the effects not explicitly included in the single-structure PB model.[9,44] As discussed in Section 3.5, the use of approximate physical arguments to establish a link between the magnitude of ϵ_p and the extent of charging-induced reorganization are seen to be of poor value in interpreting the results. Indeed, some of the results in Section 3.2 show that apparently similar sites (in terms of shifting or exposure) often require very different best-prediction ϵ_p values or (ϵ_{ind}, ϵ_{pair}) combinations; analogous results were observed with the PDLD/S model.[46] Thus, however tempting it may be to look for general rules of thumb for selecting ϵ_p values (e.g., "shifted pK_a requires low ϵ_p"), the plain fact is that no such rules seem to hold. To actually understand the reasons for the behavior of a particular site (and to better predict it), one should explicitly account for structural reorganization. One route is to perform MD-based free-energy calculations.[8,107−109] Alternatively, this can be done within the context of the PB or PDLD/S models by means of LRA[30,39,40] and constant-pH MD,[41,42] or by more approximate methods (e.g., see discussion in ref 42 and references therein). Naturally, PB models are evolving to or being replaced by these more microscopic approaches. But, while single-structure PB calculations remain useful (due to their computational speed and predictive power), it seems advisable to rely on proper benchmarking rather than on vague physical reasoning.

Acknowledgment. We thank Prof. Maria José Calhorda for help with quantum chemical calculations. We acknowledge financial support from Fundação para a Ciência e a Tecnologia, Portugal, through fellowships SFRH/BD/6477/2001 (V.H.T.), SFRH/BPD/9466/2002 (C.A.C.), SFRH/BPD/14540/2003 (M.M.), SFRH/BD/10622/2002 (B.L.V.), and project grants POCTI/BME/32789/99 and POCTI/BME/45810/2002.

Supporting Information Available: Tables reporting experimental p*K*a values; experimental *E* values (relative to standard hydrogen electrode); atomic partial charges for heme group. Figure showing experimental versus calculated pK_a shifts for HEWL for all (ϵ_{ind}, ϵ_{pair}) combinations, without inclusion of water molecules or tautomers, using a fully neutral reference state. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Linderstrøm-Lang, K. *C. R. Trav. Lab. Carlsberg* **1924**, *15*, 1−29.
(2) Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333−5339.
(3) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395−2417.
(4) Simonson, T. *Rep. Prog. Phys.* **2003**, *66*, 737−787.
(5) Warshel, A.; Parson, W. W. *Q. Rev. Biophys.* **2001**, *34*, 563−679.
(6) Field, M. J. *J. Comput. Chem.* **2002**, *23*, 48−58.
(7) Noodleman, L.; Lovell, T.; Han, W.-G.; Li, J.; Himo, F. *Chem. Rev.* **2004**, *104*, 459−508.
(8) Lee, F. S.; Chu, Z. T.; Warshel, A. *J. Comput. Chem.* **1993**, *14*, 161−185.
(9) Sham, Y. Y.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 4458−4472.
(10) Warwicker, J.; Watson, H. C. *J. Mol. Biol.* **1982**, *157*, 671−679.
(11) Sharp, K. A.; Honig, B. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301−332.
(12) Bashford, D.; Karplus, M. *Biochemistry* **1990**, *29*, 10219−10225.
(13) Yang, A.-S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. *Proteins: Struct., Funct., Genet.* **1993**, *15*, 252−265.
(14) Simonson, T. *Curr. Opin. Struct. Biol.* **2001**, *11*, 243−252.
(15) Bashford, D. *Front. Biosci.* **2004**, *9*, 1082−1099.
(16) Feig, M.; Brooks, C. L., III. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217−224.
(17) Antosiewicz, J.; Porschke, D. *Biochemistry* **1989**, *28*, 10072−10078.
(18) Beroza, P.; Fredkin, D. R.; Okamura, M. Y.; Feher, G. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 5804−5808.
(19) Baptista, A. M.; Martel, P. J.; Soares, C. M. *Biophys. J.* **1999**, *76*, 2978−2998.
(20) Tanford, C.; Roxby, R. *Biochemistry* **1972**, *11*, 2192−2198.

(21) Bashford, D.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 9956−9561.
(22) Gilson, M. K. *Proteins: Struct., Funct., Genet.* **1993**, *15*, 266−282.
(23) Spassov, V.; Bashford, D. *J. Comput. Chem.* **1999**, *20*, 1091−1111.
(24) Warshel, A. *Biochemistry* **1981**, *20*, 3167−3177.
(25) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *J. Mol. Biol.* **1994**, *238*, 415−436.
(26) Demchuk, E.; Wade, R. C. *J. Phys. Chem.* **1996**, *100*, 17373−17387.
(27) Gibas, C. J.; Subramanian, S. *Biophys. J.* **1996**, *71*, 138−147.
(28) Alexov, E. G.; Gunner, M. R. *Biophys. J.* **1997**, *72*, 2075−2093.
(29) Baptista, A. M.; Soares, C. M. *J. Phys. Chem. B* **2001**, *105*, 293−309.
(30) Schutz, C. N.; Warshel, A. *Proteins: Struct., Funct., Genet.* **2001**, *44*, 400−417.
(31) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. *Biophys. J.* **2002**, *83*, 1731−1748.
(32) Warwicker, J. *Protein Sci.* **2004**, *13*, 2793−2805.
(33) Martel, P. J.; Soares, C. M.; Baptista, A. M.; Fuxreiter, M.; Náray-Szabó, G.; Louro, R. O.; Carrondo, M. A. *J. Biol. Inorg. Chem.* **1999**, *4*, 73−86.
(34) Gunner, M.; Alexov, E. *Biochim. Biophys. Acta* **2000**, *1458*, 63−87.
(35) Ullmann, G. M. *J. Phys. Chem. B* **2000**, *104*, 6293−6301.
(36) Louro, R. O.; Bento, I.; Matias, P. M.; Catarino, T.; Baptista, A. M.; Soares, C. M.; Carrondo, M. A.; Turner, D. L.; Xavier, A. V. *J. Biol. Chem.* **2001**, *276*, 44044−44051.
(37) Teixeira, V. H.; Soares, C. M.; Baptista, A. M. *J. Biol. Inorg. Chem.* **2002**, *7*, 200−216.
(38) Bento, I.; Matias, P. M.; Baptista, A. M.; da Costa, P. N.; van Dongen, W. M. A. M.; Saraiva, L. M.; Schneider, T. R.; Soares, C. M.; Carrondo, M. A. *Proteins: Struct., Funct., Bioinf.* **2004**, *54*, 135−152.
(39) Lee, F. S.; Chu, Z.-T.; Bolger, M. B.; Warshel, A. *Protein Eng.* **1992**, *5*, 215−228.
(40) Eberini, I.; Baptista, A. M.; Gianazza, E.; Fraternali, F.; Beringhelli, T. *Proteins: Struct., Funct., Bioinf.* **2004**, *54*, 744−758.
(41) Baptista, A. M.; Martel, P. J.; Petersen, S. B. *Proteins: Struct., Funct., Genet.* **1997**, *27*, 523−544.
(42) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. *J. Chem. Phys.* **2002**, *117*, 4184−4200.
(43) Böttcher, C. J. F.; van Belle, O. C.; Bordewijk, P.; Rip, A. *Theory of Electric Polarization*; Elsevier: Amsterdam, 1973; Vol. I.
(44) Warshel, A.; Åqvist, J. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 267−298.
(45) Schaefer, M.; Sommer, M.; Karplus, M. *J. Phys. Chem. B* **1997**, *101*, 1663−1683.
(46) Sham, Y. Y.; Muegge, I.; Warshel, A. *Biophys. J.* **1998**, *74*, 1744−1753.
(47) Warshel, A.; Russell, S. T. *Q. Rev. Biophys.* **1984**, *17*, 283−422.
(48) Harvey, S. C. *Proteins: Struct., Funct., Genet.* **1989**, *5*, 78−92.
(49) King, G.; Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1991**, *95*, 4366−4377.
(50) Simonson, T.; Perahia, D. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 1082−1086.
(51) Simonson, T.; Brooks, C. L., III. *J. Am. Chem. Soc.* **1996**, *118*, 8452−8458.
(52) Warshel, A.; Papazyan, A. *Curr. Opin. Struct. Biol.* **1998**, *8*, 211−217.
(53) Simonson, T.; Archontis, G.; Karplus, M. *J. Phys. Chem. B* **1999**, *103*, 6142−6156.
(54) Krishtalik, L. I.; Kuznetsov, A. M.; Mertz, E. L. *Proteins: Struct., Funct., Genet.* **1997**, *28*, 174−182.
(55) Sharp, K. A.; Jean-Charles, A.; Honig, B. *J. Phys. Chem.* **1992**, *96*, 3822−3828.
(56) Khare, D.; Alexander, P.; Antosiewicz, J.; Bryan, P.; Gilson, M.; Orban, J. *Biochemistry* **1997**, *36*, 3580−3589.
(57) Oliveberg, M.; Arcus, V. L.; Fersht, A. R. *Biochemistry* **1995**, *34*, 9424−9433.
(58) March, K. L.; Maskalick, D. G.; England, R. D.; Friend, S. H.; Gurd, F. R. N. *Biochemistry* **1982**, *21*, 5241−5251.
(59) Kuramitsu, S.; Hamaguchi, K. *J. Biochem.* **1980**, *87*, 1215−1219.
(60) Bartik, K.; Redfield, C.; Dobson, C. M. *Biophys. J.* **1994**, *66*, 1180−1184.
(61) Schaller, W.; Robertson, A. D. *Biochemistry* **1995**, *34*, 4714−4723.
(62) Forsyth, W. R.; Gilson, M. K.; Antosiewicz, J.; Jaren, O. R.; Robertson, A. D. *Biochemistry* **1998**, *37*, 8643−8652.
(63) Oda, Y.; Yoshida, M.; Kanaya, S. *J. Biol. Chem.* **1993**, *268*, 88−92.
(64) Oda, Y.; Yamazaki, T.; Nagayama, K.; Kanaya, S.; Kuroda, Y.; Nakamura, H. *Biochemistry* **1994**, *33*, 5275−5284.
(65) Fujii, S.; Akasaka, K.; Hatano, H. *J. Biochem.* **1980**, *88*, 789−796.
(66) Chivers, P. T.; Prehoda, K. E.; Volkman, B. F.; Kim, B.-M.; Markley, J. L.; Raines, R. T. *Biochemistry* **1997**, *36*, 14985−14991.
(67) Turner, D. L.; Salgueiro, C. A.; Catarino, T.; LeGall, J.; Xavier, A. V. *Eur. J. Biochem.* **1996**, *241*, 723−731.
(68) Salgueiro, C. A.; Turner, D. L.; Xavier, A. V. *Eur. J. Biochem.* **1997**, *244*, 721−734.
(69) Louro, R. O.; Catarino, T.; Turner, D. L.; Piçarra-Pereira, M. A.; Pacheco, I.; LeGall, J.; Xavier, A. V. *Biochemistry* **1998**, *37*, 15808−15815.
(70) Gayda, J.; Benosman, H.; Bertrand, P.; More, C.; Asso, M. *Eur. J. Biochem.* **1988**, *177*, 199−206.
(71) Louro, R. O.; Catarino, T.; LeGall, J.; Turner, D. L.; Xavier, A. V. *ChemBioChem* **2001**, *2*, 831−837.
(72) Pereira, P. M.; Pacheco, I.; Turner, D. L.; Louro, R. O. *J. Biol. Inorg. Chem.* **2002**, *7*, 815−822.
(73) Eisenhaber, F.; Argos, P. *J. Comput. Chem.* **1993**, *14*, 1272−1280.
(74) Eisenhaber, F.; Lijnzaad, P.; Argos, P.; Sander, C.; Scharf, M. *J. Comput. Chem.* **1995**, *16*, 273−284.
(75) van Gunsteren, W. F.; Berendsen, H. J. C. *Groningen Molecular Simulation (GROMOS) Library Manual*; Biomos: Groningen, The Netherlands, 1987.
(76) Smith, L. J.; Mark, A. E.; Dobson, C. M.; van Gunsteren, W. F. *Biochemistry* **1995**, *34*, 10918−10931.
(77) Soares, C. M.; Martel, P. J.; Carrondo, M. A. *J. Biol. Inorg. Chem.* **1997**, *2*, 714−727.
(78) Bashford, D.; Gerwert, K. *J. Mol. Biol.* **1992**, *224*, 473−486.
(79) Bashford, D. An Object-Oriented Programming Suite for Electrostatic Effects in Biological Molecules. In *Scientific Computing in Object-Oriented Parallel Environments*, ISCOPE97; Ishikawa, Y., Oldehoeft, R. R., Reynders, J. V. W., Tholburn, M., Eds.; Springer: Berlin, 1997; Vol. 1343.
(80) Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151−176.
(81) Gilson, M. K.; Sharp, K. A.; Honig, B. *J. Comput. Chem.* **1987**, *9*, 327−335.
(82) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*, revision A.7; Gaussian, Inc.: Pittsburgh, PA, 1998.
(83) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269−10280.
(84) Harbury, H. A.; Cronin, J. R.; Fanger, M. W.; Hettinger, T. P.; Murphy, A. J.; Myer, Y. P.; Vinogradov, S. N. *Proc. Natl. Acad. Sci. U.S.A.* **1965**, *54*, 1658−1664.
(85) Gao, J.; Pavelites, J. J. *J. Am. Chem. Soc.* **1992**, *114*, 1912−1914.
(86) Pranata, J. *J. Comput. Chem.* **1993**, *14*, 685−690.
(87) Chen, J. L.; Noodleman, L.; Bashford, D. *J. Phys. Chem.* **1994**, *98*, 11059−11068.
(88) Andzelm, J.; Kölmel, C. *J. Chem. Phys.* **1995**, *103*, 9312−9320.
(89) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995−2001.
(90) Jean-Charles, A.; Nicholls, A.; Sharp, K.; Honig, B.; Tempczyk, A.; Hendrickson, T. F.; Still, W. C. *J. Am. Chem. Soc.* **1991**, *113*, 1454−1455.
(91) Mohan, V.; Davis, M. E.; McCammon, J. A.; Pettitt, B. M. *J. Phys. Chem.* **1992**, *96*, 6428−6431.
(92) Bashford, D.; Case, D. A.; Dalvit, C.; Tennant, L.; Wright, P. E. *Biochemistry* **1993**, *32*, 8045−8056.
(93) Postma, J. P. M.; Berendsen, H. J. C.; Haak, J. R. *Faraday Symp. Chem. Soc.* **1982**, *17*, 55−67.
(94) Gallicchio, E.; Kubo, M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271−6285.
(95) Pauling, L. *The Nature of the Chemical Bond*, 3rd ed.; Cornell University Press: Ithaca, NY, 1960.
(96) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978−1988.
(97) In the SPC model, the one used to model water in the 43A1 force field, the hydrogen atoms are "inside" the oxygen-repulsive region and have no Lennard−Jones interaction. Therefore, only the oxygen atom is needed to compute atomic radii.
(98) Vorobjev, Y. N.; Scheraga, H. A.; Hitz, B.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 10940−10948.
(99) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: New York, 1992.
(100) Warshel, A. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 5250−5254.

(101) Warshel, A.; Russell, S. T.; Churg, A. K. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 4785−4789.

(102) Strictly speaking, the "interaction" between two sites *i* and *j* is the free energy of the process, whereby one protein with only site *i* charged and another one with only site *j* charged, give rise to one protein molecule fully neutral and another one with both *i* and *j* charged, while other protein sites remain neutral during the process.[13,45,46]

(103) Mehler, E. L.; Eichele, G. *Biochemistry* **1984**, *23*, 3887−3991.

(104) Identical results are obtained if the fully neutral reference state is chosen, because the interaction terms are the same in both cases.[29]

(105) Although the argument that follows pertains to MC simulations, it can be easily extended to mean-field treatments[8,20,21] by considering that the "moves" consist in that case of single-site (de)protonations within an environment of average-charged sites.

(106) This follows from the fact that, because we are dealing with a single type of redox site, the relative positions of the resulting redox titration curves are independent of E_{mod}, whose value can be determined afterward as the displacement leading to best agreement with experimental data (e.g., least-RMSD of reduction potentials).

(107) Warshel, A.; Sussman, F.; King, G. *Biochemistry* **1986**, *25*, 8368−8372.

(108) Kenneth, M.; Merz, J. *J. Am. Chem. Soc.* **1991**, *113*, 3572−3575.

(109) Simonson, T.; Carlsson, J.; Case, D. A. *J. Am. Chem. Soc.* **2004**, *126*, 4167−4180.