Consistent Calculations of pK_a's of Ionizable Residues in Proteins: Semi-microscopic and Microscopic Approaches

Yuk Yin Sham, Zhen Tao Chu, and Arieh Warshel*

Department of Chemistry, University of Southern California, Los Angeles, California 90089-1062

Received: October 31, 1996; In Final Form: January 27, 1997[⊗]

One of the most direct benchmarks for electrostatic models of macromolecules is provided by the pK_a's of ionizable groups in proteins. Obtaining accurate results for such a benchmark presents a major challenge. Microscopic models involve very large opposing contributions and suffer from convergence problems. Continuum models that consider the protein permanent dipoles as a part of the dielectric constant cannot reproduce the correct self-energy. Continuum models that treat the local environment in a semi-microscopic way do not take into account consistently the protein relaxation during the charging process. This work describes calculations of pK_a's in protein in an accurate yet consistent way, using the semi-microscopic version of the protein dipoles Langevin dipoles (PDLD) model, which treats the protein relaxation in the microscopic framework of the linear response approximation. This approach allows one to take into account the protein structural reorganization during formation of charges, thus reducing the problems with the use of the so-called "protein dielectric constant", ϵ_p. The model is used in calculations of pK_a's of the acidic groups of lysozyme, and the calculated results are compared to the corresponding results of discretized continuum (DC) studies. It is found that the present approach is more consistent than current DC models and also provides improved accuracies. Significant emphasis is given to the self-energy term, which has been pointed out in our early works but has been sometimes overlooked or presented as a small effect. The meaning of the dielectric constant ϵ_p used in DC models is clarified and illustrated, establishing the finding (e.g. King et. al., *J. Phys. Chem.* **1991**, *95,* 4366) that this parameter represents the contributions that are not treated explicitly in the given model, rather than the "true" dielectric constant. It is pointed out that recent suggestions to use large ϵ_p to obtain improved DC results might not be much different than our earlier suggestion to use a large effective dielectric for charge−charge interactions. This ϵ_p reduces the overestimate of charge−charge interactions relative to models that use small ϵ_p while not considering the protein relaxation explicitly. Unfortunately, the use of large ϵ_p does not reproduce consistently the self-energies of isolated ionized groups in protein interiors. The recent interest in taking protein flexibility into account in pK_a calculations is addressed. It is pointed out that running MD over protein configurations will not by itself lead to a more consistent value of ϵ_p. It is clarified that a smaller value of ϵ_p, which is not really more (or less) consistent with the physics of the proteins, will be obtained if one uses our LRA (linear response approximation) formulation, generating configurations of both neutral and ionized states of the protein. It is also stated that such studies have been a standard part of our approach for some time. The present model involves a consecutive running of all-atom MD simulations of solvated proteins and an automated used of the electrostatic PDLD model. This allows one to move consistently to any level of explicit solvent model, keeping an arbitrary number of solvent molecules in an explicit all-atom representation, while treating the rest as dipoles. This capacity is used in examining the microscopic basis of the PDLD models by comparing its free energy contributions to those obtained by the all-atom linear response approximation treatment. The agreement appears to be quite encouraging, thus further verifying the microscopic character of the PDLD model. Finally it is reclarified that real continuum models cannot provide proper descriptions of charges in protein and that current DC models are becoming more and more microscopic in nature.

1. Introduction

Electrostatic energies play a major role in controlling the functions of proteins[1−6] and provide what is probably the most important element in structure−function correlation of biological molecules.[3,4,7,8] Thus, the ability to determine accurately electrostatic energies is a key requirement in any attempt to predict functional properties of proteins.

One of the most direct and challenging benchmarks for electrostatic models is provided by the pK_a's of ionizable groups in proteins. In fact, it has been argued that the ability to predict enzyme rate constants is limited by the accuracy of the corresponding electrostatic calculations and therefore by the accuracy of pK_a calculations.[9]

The challenge of evaluating pK_a's and the corresponding titration curves has been addressed on a macroscopic level quite early in the pioneering works of Linderstrom-Lang,[10] Tandford and Kirkwood,[11] and others.[12] However, these early works overlooked the fact that the pK_a of a given ionized group depends on the corresponding self-energy, which is determined by the local environment. These studies concentrated only on the interaction between ionizable groups and considered the intrinsic pK_a as an adjustable parameter, thus avoiding the most challenging problem altogether (see discussion in refs 3, 13). Such approaches have been justified at the time of the influential Tanford and Kirkwood (TK) work[11] when it was not clear what proteins looked like, and it could have been assumed that all ionized groups are located on the surfaces of the proteins. However, with the emergence of protein crystal structures of

* Author to whom correspondence should be addressed.
[⊗] Abstract published in *Advance ACS Abstracts,* May 1, 1997.

S1089-5647(96)03412-8 CCC: $14.00

proteins it became clear that ionizable groups can be located quite far from the surfaces of proteins. This finding appeared to be inconsistent with the implicit assumptions of the TK model and the corresponding calculated pK_a's (for example a consistent use of the TK model would produce incorrect pK_a's as is demonstrated in Tables 12−16 of ref 14). Nevertheless, the TK model continued to be popular for quite some time[5,15] because of its simplicity and rigorous derivation (of what turned out to be an incomplete model) and perhaps because the crucial role of self-energies of charges in proteins[13] was not widely appreciated.

The fundamental problems associated with the self-energy and the corresponding intrinsic pK_a was realized[9,13,16] in the mid 1970s when it was recognized that the local environment, which was not considered in macroscopic models, plays a crucial role in determining the energetics of ionized residues. This realization led to the first consistent treatment of the pK_a of ionizable groups in proteins by a series of simplified microscopic models[3,9,16,17] and subsequent semi-microscopic models.[14,18] Most of these early studies involved the use of the protein dipoles Langevin dipoles (PDLD) model. The main idea behind the PDLD approach has been the realization that the safest way to elude the traps in the continuum treatments of the electrostatic energies in macromolecules is to use microscopic models where all interactions are considered explicitly even if this requires the introduction of simplified potential functions. The resulting model has been discussed and examined extensively elsewhere (e.g. ref 14). The justification of this model and its consistency with the actual polarization of water molecules and other polar models have been demonstrated.[3,10,19,20] The PDLD model was sometimes misunderstood[21] including recent suggestions that this explicit dipolar model is a macroscopic model.[25] While we disagree with these suggestions,[26] we think it is useful to recognize that regardless of what name is chosen to describe explicit dipolar models, one fact remains: the PDLD model treated electrostatic energies in protein consistently at least a decade before any alternative macroscopic models.

The understanding of the relationship between pK_a and solvation energies has increased significantly in recent years. Evaluation of pK_a's in solution using experimental gas phase energies and calculated solvation energies has been reported quite early.[29] Evaluating pK_a's using quantum mechanical calculations of gas phase energies and macroscopic estimates of solvation energies has also become quite common recently.[30,31] However, evaluating pK_a's in solution is trivial as compared to the challenges in evaluating pK_a's in proteins.[32] In solution one can obtain almost perfect agreement by calibrating empirical van der Waals radii (see ref 29) or Born's radii, while in proteins the pK_a is quite different in different regions, and a given radius cannot reproduce the correct value everywhere.

Discretized continuum (DC) methods[34−36] were developed partially in order to be able to construct "realistic" shapes of the systems of interest. However, despite the ability to represent the actual shape of proteins and the dielectric effect of the surrounding solvent, DC methods did not give reasonable pK_a values for ionizable groups in proteins until the gradual realization that the local environment must be treated in a microscopic way (see discussion in subsequent sections). DC approaches with semi-microscopic treatments of the local environment and the corresponding self-energies started to emerge in the early 1990[4,37−39] and are now commonly used.

Attempts to evaluate pK_a's in proteins by fully microscopic free energy perturbation (FEP) approaches were also reported,[14,40−42] and a very instructive attempt to use the linear response approximation (LRA) was reported recently by Levy and co-workers.[43]

Despite the above mentioned progress there are still major problems and challenges with regard to the meaning of the dielectric constant used in macroscopic models[8,44] and the convergence of microscopic models. As much as the evaluation of pK_a's in proteins is concerned, there are still large deviations between calculated and observed values[17,38] and some confusion with regard to the difference between obtaining precise results by macroscopic models (where a large dielectric constant leads automatically to such results) and obtaining reliable results by microscopic models (see discussion in refs 8, 45).

The present work revisits the challenge of evaluating pK_a's of ionizable residues in proteins. This is done in a more extensive and systematic way than in our earlier works, using more powerful computers and more extensive averaging procedures. The main emphasis is placed on our semi-microscopic model since it can be compared directly to alternative DC models. This allows us to demonstrate the crucial role of protein reorganization during the charging process and its relationship to the dielectric constant used in the DC models.

The structure of the paper is as follows: Section 2 describes our theoretical approaches. Crucial concepts such as self-energies and their role in consistent evaluation of electrostatic energies in macromolecules are pointed out. The importance of consistent introduction of microscopic elements in the so-called macroscopic treatment is reemphasized, pointing out that the protein configuration should be relaxed with the charged and uncharged configuration in order to be consistent with the correct physics of electrostatic effects. Our treatment of interactions between ionized groups and the corresponding treatment of titration curves is outlined. The implementation of automated and consistent configuration averaging in the PDLD/S and PDLD methods is briefly described. The main features of our all-atom LRA approach are outlined emphasizing the treatment of long-range electrostatic effects. Section 3 describes our computation studies, comparing first the results of the PDLD/S methods to related DC approaches, pointing out the advantages of our consistent treatment. Next we establish the close relationships between the LRA and PDLD models and demonstrate that the semi-microscopic version of the LRA model (LRA/S model) is as accurate as the PDLD/S model. Finally, we discuss in section 4 the implications of the present study in terms of both fundamental concepts of electrostatic effects in proteins and practical aspects of pK_a calculations.

2. Theoretical Approaches and Simulation Strategies

2.1. General Formulation. *2.1.1. The Energetics of Ionized Groups in Proteins.* The energy balance associated with ionizing a group in a protein can be described by the thermodynamic cycle of Figure 1. This cycle which has been introduced in ref 9 gives the pK_a of an ionizable residue by

$$\Delta G^{p}(AH_p \rightarrow A_p^- + H_w^+) = \Delta G^{w}(AH_p \rightarrow A_p^- + H_w^+) + \Delta G_{sol}^{w\rightarrow p}(A^-) - \Delta G_{sol}^{w\rightarrow p}(AH) \quad (1)$$

where p and w designate protein and water, respectively, and $\Delta G_{sol}^{w\rightarrow p}$ represents the free energy difference of moving the indicated group from water to its protein active site. This free energy difference is considered formally as a change in "solvation" free energies.

Equation 1 can be rewritten as

$$pK_{a,i}^{p} = pK_{a,i}^{w} + \frac{1}{2.3RT}\Delta\Delta G_{sol}^{w\rightarrow p}(AH_i \rightarrow A_i^-) \quad (2)$$

where the $\Delta\Delta G$ term consist of the last two terms of eq 1. This

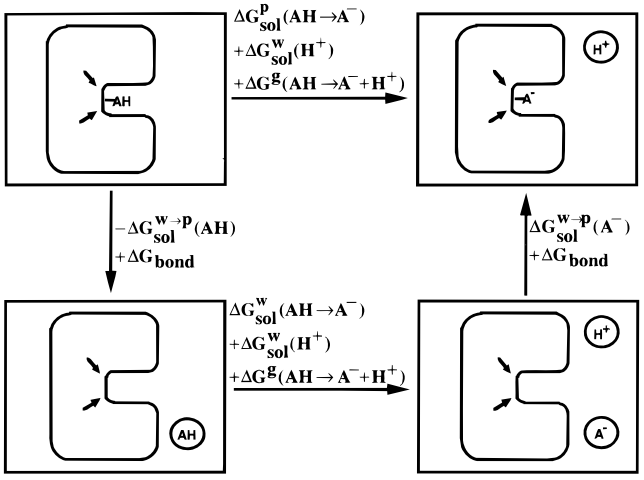

Figure 1. Thermodynamic cycle used in calculating the pK_a's of ionizable residues. ΔG^{p}_{sol} and ΔG^{w}_{sol} designate the corresponding contributions in protein and solution, respectively. ΔG_{bond} is the energy of the bond between the acid and the protein, which is assumed to have the same strength for AH and A^-.

fundamental equation (that might seem obvious today) and the corresponding thermodynamic cycle have been formulated in ref 9 and later by others.[37,38]

Using eq 2 converts the problem of evaluating a pK_a in a protein to evaluation of the change in "solvation" energy associated with moving the charge from water to the protein site. This is significantly simpler, and at present more reliable, than the evaluation of the absolute pK_a, which involves the determination of the gas phase proton affinity and the solvation of A^- and H_3O^+.[13,29,33,46]

In order to evaluate the free energy of an ionized group in a protein, it is useful and convenient to consider first the self-energy of ionizing this group when all other ionizable groups are uncharged and then to consider the effect of charging the other groups to their given ionization state. Thus, we can express the ΔG_{sol} of eq 1 as

$$(\Delta G^{w\rightarrow p}_{sol})_i = (\Delta G^{p}_{self} - \Delta G^{w}_{self})_i + \sum_{i\neq j}\Delta G^{p}_{ij} \quad (3)$$

$$= (\Delta G^{p}_{q\mu} + \Delta G^{p}_{q\alpha} + \Delta G^{p}_{qw} - \Delta G^{w}_{self})_i + \sum_{i\neq j}\Delta G^{p}_{ij}$$

where ΔG_{self} is the self-energy associated with charging the ith group in its specific environment. In the case of a charge in a protein we decompose ΔG_{self} into the interaction between the charge and its surrounding permanent dipoles ($\Delta G_{q\mu}$) and induced dipoles ($\Delta G_{q\alpha}$) as well as with the water molecules in and around the protein (ΔG_{qw}). Thus eq 3 can be viewed as the sum of the loss of "solvation" energy associated with removing the charges from water ($-\Delta G^{w}_{self}$) plus the "solvation" of the charge by its surrounding protein environment (the protein dipoles and water molecules) and finally the interaction between the charge and the ionized groups.

It is important to note that the crucial self-energy terms in eq 3 were later adopted by other workers[37] and renamed, introducing a cycle that involves a hypothetical nonpolar protein (see ref 18 for such a cycle) where the charge is first moved from water to a hypothetical nonpolar environment, without the protein permanent dipoles, followed by activation of the dipoles. In this new notation we have

$$\Delta G^{w\rightarrow p}_{sol} = \Delta G_{Born} + \Delta G_{back} = \Delta G^{p}_{self} - \Delta G^{w}_{self} \quad (4)$$

where the free energy of the first step is denoted by ΔG_{Born}, while the interaction between the ionized group and its polar environment has been named "ΔG_{back}". The ΔG_{back} term is given to a good approximation by $\Delta G_{q\mu}/\epsilon_p$, where ϵ_p is the assumed dielectric "constant" of the protein (the meaning of this parameter will be discussed in subsequent sections). It is also possible to relate our original energy decomposition to other recent expressions by writing

$$\Delta G^{p}_{self} - \Delta G^{w}_{self} = \Delta G^{p}_{q\mu} + (\Delta G^{p}_{q\alpha} + \Delta G^{p}_{qw} - \Delta G^{w}_{self}) \quad (5)$$

$$= \Delta G^{p}_{dipoles} + \Delta G^{p}_{desolvation}$$

where the first term designates the interaction with the protein permanent dipoles and the second term represents the electrostatic work of moving the system to a nonpolar protein. Here we are close to the notation of ref 47, except that in our microscopic treatment $\Delta G_{dipoles}$ and ΔG_{desolv} are not scaled by ϵ_p and that the "desolvation" energy involves the "dielectric effect" of the induced dipoles and the solvent when the permanent dipoles are already turned on. The PDLD treatment does not involve a thermodynamic cycle with a hypothetical nonpolar protein, but a cycle where the charge is moved directly to the real protein.

In general we can express the pK_a of each group of the protein by

$$pK^{p}_{a,i} = pK^{p}_{int,i} + \Delta pK^{charges}_{a,i} \quad (6)$$

where $pK^{p}_{int,i}$ is the so-called intrinsic pK_a that the ith group in the protein would have when all the other groups are in their neutral states, and $\Delta pK^{charges}_{a,i}$ represents the effects of the other ionized groups. Using eqs 2 and 3, we can rewrite eq 6 as

$$2.3RTpK^{p}_{a,i} = 2.3RTpK^{w}_{a,i} + (\Delta\Delta G^{w\rightarrow p}_{self})_i + \sum_{i\neq j}\Delta G^{p}_{ij} \quad (7)$$

where ΔG_{ij} represents the interaction with the jth ionized group. The evaluation of this term will be considered below.

2.1.2. Interactions between Ionizable Residues. After evaluating the self-energy of each of the ionizable residues (in the reference system where all other residues are in their neutral state) we can evaluate the perturbation due to the interactions between different ionized residues. In other words, after determining the electrostatic work of bringing a charge to the neutral protein we may now ask how much does this reversible work (or free energy) change when other groups are ionized. It is important to comment that this issue has been frequently considered to be the main and sometimes the only problem in electrostatic calculations (perhaps because of the difficulties in recognizing the importance of the self-energy term), despite the fact that the charge−charge interaction term is usually rather small. Nevertheless, it is important to be able to evaluate this contribution in a practical and consistent way. The approach used here for this purpose is similar to that used by others[37,38,48] and is described in detail below.

Our starting point is the free energy of different charge configuration that can be expressed as[49,50] (see a closely related recent expression in ref 38)

$$\Delta G^{(m)} = \sum_i \left\{ -2.3RTq^{(m)}_i[pK^{p}_{int,i} - pH] + \frac{1}{2}\sum_{i\neq j} W_{ij}q^{(m)}_i q^{(m)}_j \right\}$$

$$= \sum_i \{-q^{(m)}_i W^0_i + \sum_{i\neq j} W_{ij}q^{(m)}_i q^{(m)}_j\} \quad (8)$$

where $q_i^{(m)}$ is the actual charge of the ith group and it can be 0 or -1 for acids and 0 or 1 for bases, and W_{ij} is the charge−charge interaction term that will be discussed below.

With the free energies of all possible charge configurations we can write

$$Z = \sum_m \exp(-\Delta G^{(m)}\beta\} \tag{9}$$

Here we have free energy rather than potential energy in the partition function but such treatment is still justified as established by Tanford and Kirkwood[11]. With the partition function of eq 9 we can calculate the average of any property and in particular we can evaluate the average charge by

$$\langle q_i\rangle = \frac{\sum_m q_i^{(m)} \exp\{-\Delta G^{(m)}\beta\}}{Z} \tag{10}$$

We can define the pK_a by using the value of the pH (pK_a = pH$_i$) where $\langle q_i\rangle$ is the median value of its neutral and ionized states. In this way we can write

$$\mathrm{p}K_{a,i} - \mathrm{pH}_i = \log\left(\frac{|\langle q_i\rangle|}{1 - |\langle q_i\rangle|}\right) = \log(1) = 0 \tag{11}$$

Thus, $|q_i| = 0.5$ leads to p$K_{a,i}$ = pH$_i$, and our problem boils down to the evaluation of $\langle q_i\rangle$. This can be done, at least in principle, by evaluating eq 10 at the specified pH, while considering all states of the system. For example, if we have a protein with two acidic groups $\mathbf{q} = (q_1,q_2)$ we will have to consider the four states (0,0), (−1,0), (0,−1), and (−1,−1). Such an explicit procedure becomes very expensive when the number of ionized residues is significant and thus cannot be used in routine calculations (a possible practical treatment is to use the Monte Carlo approach of ref 48). Another possibility is to use the effective charge approximation of Tanford and Roxby,[12] where it is assumed that the average charge of each residue depends on the average charges of all other residues. This approximation can be expressed as

$$\Delta G_i^{(\alpha)} \cong \sum_{\alpha=1}^{2}\{-q_i^{(\alpha)}W^0{}_i + \sum_{i\neq j} W_{ij}q_i^{(\alpha)}\langle q_j\rangle\} \tag{12}$$

where we have now only two states (charged and uncharged) for each residue. Using eq 12, one finds that

$$\Delta G(\mathrm{AH}_i \rightarrow \mathrm{A}_i^-) = \bar{q}_i\{-W_i^0 + \sum_{i\neq j} W_{ij}\langle q_j\rangle\} \tag{13}$$

where $\bar{q}_i$ is the charge of the ith group in its ionized form (−1 and +1 for acids and bases), respectively, and where AH is neutral and positively charged for acids and bases, respectively. Note that $\bar{q}_i$ is not identical to q_i since it cannot be zero. $\langle q_i\rangle$ of this effective two state model is given by

$$\langle q_i\rangle = \frac{q_i \exp\{-\Delta G(\mathrm{AH}_i \rightarrow \mathrm{A}_i^-)\beta\}}{\exp\{-\Delta G(\mathrm{AH}_i \rightarrow \mathrm{A}_i^-)\beta\} + 1} \tag{14}$$

Now eqs 13 and 14 are solved self-consistently where at each evaluation of $\langle q_i\rangle$ all the average charges of other residues are kept at their latest values. Once self-consistency is achieved, the pK_a at the ith group is determined as the pH where $\langle q_i\rangle$ = $^1/_2\bar{q}_i$. While eq 14 is quite useful, it is sometimes important to obtain a less approximate expression that combines the simplicity of eq 14 and the rigor of eq 9. A useful approximation can be obtained by a hybrid approach[37,38] where the charges of each residue are evaluated using

$$\langle q_i\rangle = \frac{\sum_{m_s} q_i^{(m_s)} \exp\{-\Delta G_s^{(m_s)}\beta\}}{Z_s} \tag{15}$$

where s designates all the residues within an explicit sphere of a specified cutoff ($R \leq R_s$) around the ith residue. Here we define the configuration m_s by all the possible ionization states of the residues within the cutoff range. $\Delta G_s^{(m_s)}$ is the approximated effective free energy given by

$$\Delta G_s^{(m_s)} = \sum_{i=1}^{N_s}\{q_i^{m_s}[-W_i^0 + \sum_{j\neq i}^{N_s} W_{ij}q_j^{m_s}] + q_i^{m_s}\sum_{j>N_s}^{N} W_{ij}\langle q_j\rangle\} \tag{16}$$

where N is the total number of ionizable groups and N_s are the number of groups within the specified cutoff range.

The residues are now numbers from i (for the reference residue) to N_s. The first term represents the contribution of the residues within R_s, while the second corresponds to the average effect of the residues outside the range. With eqs 15 and 16 we can evaluate the pK_a of each ionized group provided we know pK_{int} and W_{ij}. The evaluation of the intrinsic pK_a has been described in the previous section, and thus we only have to address the evaluation of W_{ij}. This interaction term can be evaluated by explicit PDLD/S or LRA calculations, considering any given pair of groups and using

$$\Delta G_{ij} = \bar{q}_i\bar{q}_j W_{ij} = \Delta G(q_i{=}0\rightarrow q_i{=}\bar{q}_i)_{q_j=\bar{q}_j} - \Delta G(q_i{=}0\rightarrow q_i{=}\bar{q}_i)_{q_j=0} \tag{17}$$

This equation is evaluated by calculating the difference between the free energy of charging A$_i$ when A$_j$ is charged and the free energy of charging A$_i$ when A$_j$ is neutral. This is done while placing A$_i$ and A$_j$ in region I and II of the PDLD model, respectively. The same calculation can be performed by reversing the role of A$_i$ and A$_j$, and the agreement between the two calculated results can serve as a consistency check. The resulting W_{ij} can be rewritten as

$$W_{ij} = 332/r_{ij}\epsilon_{ij} \tag{18}$$

where r_{ij} is the average distance (in Å) between the ith and jth charge centers and where the energy given is in kcal/mol. This expression defines the effective dielectric constant ϵ_{ij} by

$$\epsilon_{ij} = 332/r_{ij}W_{ij} \tag{19}$$

The explicit evaluation of eq 17 is quite expensive and not justified in most cases. That is, in almost all cases when the distance between a charge pair is larger than 5 Å, the effective dielectric constant for charge−charge interaction can be approximated by a large number between 40 and 80 or by the function[13]

$$\epsilon_{ij} = \epsilon_{\mathrm{eff}} \cong 1 + 60[1 - \exp(-0.1r_{ij})] \tag{20}$$

A similar function has been used recently in the study of ref 51. In fact, as will be argued in section 3.5, many times the ϵ_{ij} of eq 20 gives more reliable estimates than that obtained by explicit calculations (it should be clear at this point that the asymptotic value of ϵ_{ij} at $r \rightarrow \infty$ is irrelevant since the corresponding interaction is zero). Thus, our procedure involves the use of the ϵ_{eff} of eq 20 except in cases of very strongly interacting groups, where we use eqs 17 and 19.

2.2. Calculations of Electrostatic Free Energies. After defining the free energy terms of pK_a^p we have to examine the most effective ways of evaluating these terms. Basically, we are dealing with calculations of electrostatic energies in proteins, and the relative reliability of different approaches is far from obvious and is in some respect the major subject of this work.

As stated in the previous section, our primary objective is to produce a reliable strategy for evaluating the self-energy term. In this work we will examine and discuss the performance of the PDLD, PDLD/S, LRA, and LRA/S approaches. Since the details of these three approaches have been discussed in recent works.[14,52] We will discuss below only the main points about these approaches and their current implementation.

2.2.1. The PDLD Method. The PDLD method[8,16,17] was introduced in the 1970s and has provided an early consistent way of evaluating electrostatic energies in proteins. The introduction of this microscopic model was essential in order to avoid the uncertainties and conceptual problems associated with the use of the macroscopic models of that time. The PDLD model was discussed extensively somewhere (see also the introduction section of this paper and further discussion below). Here we review only the main aspects of this model.

The PDLD model considers explicitly the proteins/solvent system with all its electrostatic components. Thus, the effective potential of a reference charged group is given by

$$\Delta V_{\mathrm{pdld}} = \Delta V^{\mathrm{p}}_{q\mu} + \Delta V^{\mathrm{p}}_{q\alpha} + \Delta V^{\mathrm{p}}_{qq} + \Delta V^{\mathrm{p}}_{qw} + \Delta G^{\mathrm{p}}_{\mathrm{bulk}} \quad (21)$$

where $\Delta V^{\mathrm{p}}_{q\mu}$ is the interaction between the charge and the protein permanent dipoles, $\Delta V^{\mathrm{p}}_{q\alpha}$ is the interaction between the charge and the protein induced dipoles, $\Delta V^{\mathrm{p}}_{qq}$ represents the interaction with other ionized groups, and $\Delta V^{\mathrm{p}}_{qw}$ is the interaction between the charge and the Langevin dipoles (which represent the average polarization of the water molecules in and around the protein). $\Delta G^{\mathrm{p}}_{\mathrm{bulk}}$ is the solvation energy due to the bulk solvent, which surrounds the region of explicit solvent molecules. Early PDLD treatments approximated the free energy terms associated with each ΔV contribution by the corresponding terms evaluated at the average structure (although energy minimization that relaxed the protein in different charge configurations was already implemented in the original work[16]). In this approach it has been assumed that solvation free energies can be represented by considering the effective potential for interaction between the solute charges and the average polarization of the solvent (or protein) dipoles. This was done with the implicit assumption of the LRA, and the resulting effective energy was considered as ΔG_{pdld} and parametrized accordingly. More recent approaches[14] used the explicit LRA to describe the protein reorganization by considering the relaxed structures in both the ionized and neutral states of the relevant charge (see below).

The $\Delta V^{\mathrm{p}}_{q\mu}$ term is evaluated by considering the Coulombic interaction between the given charge and the residual charges of the protein atoms. These residual charges are assigned according to the atom type and residue type as described in ref 14. The effect of the protein induced dipoles are evaluated as described elsewhere[14,16] by attaching an induced dipole to each protein atom and evaluating self-consistently the interaction of these dipoles with the permanent charge distribution of the system as well as with each other. The solvation of the protein and its charges by the water in and around the protein is evaluated by the following procedures. The protein is surrounded by a three-dimensional cubic grid, and each point that is within a specific van der Waals distance from a protein atom is deleted. The grid is truncated to a sphere, and each of the remaining points is occupied by a point dipole that represents

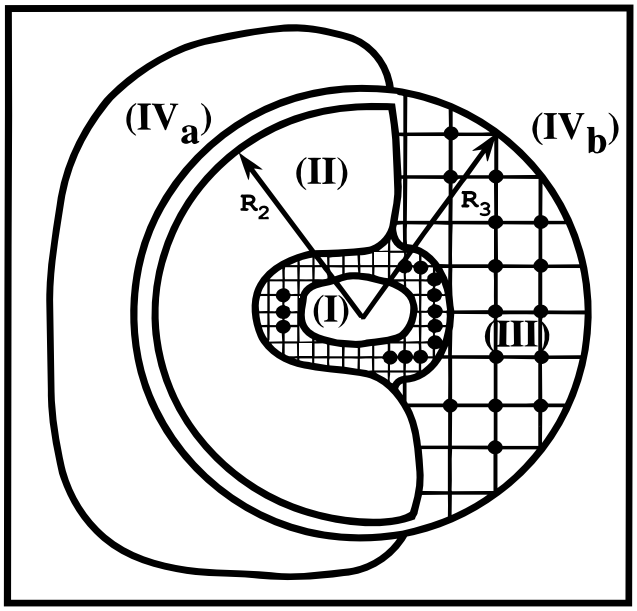

Figure 2. Regions of the protein/solvent system in the PDLD method. Region I contains the charged groups of interest. Region II contains the protein atoms found within a radius R_2 from the center. Region III is the Langevin grid truncated to a sphere of a radius R_3. The inner part of the grid has 1 Å spacing, and the outer part has a 3 Å spacing. Region IV_a contains the rest of the protein atoms outside region II. The electrostatic effects of regions I, II, and III are treated explicitly, while those of region IV (IV_a and IV_b) are considered as bulk solvent regions and are treated by a macroscopic continuum formulation. Note that the protein in region IV_a is replaced by bulk solvent.

the average polarization of a water molecule at that site. Each point dipole is allowed to be polarized toward the local field due to the protein atoms as well as other solvent dipoles except its nearest neighbors.[14] The consistency of this model with the polarization of water molecules in particular and other dipolar models in general has been demonstrated and discussed elsewhere.[3,19,20,53] A systematic study that relates dipolar lattice models to macroscopic models is presented in ref 54. The original PDLD treatment involved an average over a significant number of randomly generated grids. Later it was found that the number of averaging steps can be reduced if the grid points near the solute surface are converted to a finer grid (1 Å spacing instead of 3 Å spacing) with a corresponding reduction in the magnitude of the dipole. This treatment does not necessarily increase the accuracy of the model (in fact, it makes it less accurate in treating water molecules in protein cavities), but it produces more stable results for those who are concerned with the precision beyond the decimal point.

It has also been found recently that the original approximation that represents the solvent dipoles by Langevin type dipoles can be relaxed in many cases without a major loss in accuracy[55] (also see below) and with faster convergence. Thus the PDLD version used in the current work replaces the Langevin dipole polarization law by

$$\mu_i^{(n+1)} = \alpha_{\mathrm{L}} \xi_i^{(n)} \quad (22)$$

where ξ_i is the field on the ith dipole from its surrounding (with the exception of its nearest neighbors). α_{L} is the effective polarizability of the solvent dipole (6.3 Å^3 for 3 Å grid spacing), and n is the iteration number. It is important to recognize that eq 22 is just an approximation that is found to reproduce well the more rigorous results of the original Langevin dipoles model. Apparently this fact is not yet clear.[56] Thus we would like to emphasize that the original LD model is used as one of the options on POLARIS 6.3 and in some of our most recent

programs (ref 57), and it does give basically the same PDLD/S results as those obtained by eq 22. The bulk contribution is evaluated by the continuum approach described in ref 14. Such an approach has been, of course, implemented in early PDLD treatments and related approaches and has also been proposed in other recent studies (e.g. ref 58). The contribution of charge−charge interaction can be calculated by an explicit use of the PDLD model (by explicit inclusion of the V_{qq} term), as was demonstrated repeatedly.[17,55] However, in the present work we prefer to evaluate such contributions in a macroscopic way (see below). The present version of the PDLD model, as implemented in the program POLARIS, divides the protein/solvent into three regions as discussed in ref 14 and depicted in Figure 2. This model guarantees the correct treatment of long-range electrostatic effects by the use of the spherical boundaries[19] and by the implementation of the local reaction field (LRF) treatment,[42] where the interaction between each dipole and its surrounding is divided into short-range interaction, which is evaluated each iteration, and long-range interaction, which is updated only once in 10 self-consistent iterations.

One of the unique features of the PDLD approach is the consistent treatment of the protein structural relaxation upon formation of charges (this can be easily accomplished since all electrostatic contributions are treated explicitly). In the present treatment we achieve this consistency by combining the ENZYMIX simulation program and the POLARIS program in such a way that the PDLD results are averaged automatically over the relevant MD generated protein configurations. This is done in the framework of the LRA approach using the expression[14,59]

$$\Delta G_{\mathrm{pdld}} = \frac{1}{2}[\langle \Delta V_{\mathrm{pdld}} \rangle_{\mathbf{r}^{\mathrm{p}}_{(q=0)}} + \langle \Delta V_{\mathrm{pdld}} \rangle_{\mathbf{r}^{\mathrm{p}}_{(q=\bar{q})}}] \qquad (23)$$

where $\langle\ \rangle_{\mathbf{r}^p}$ designates an average over protein configurations generated with the indicated charge (q) and where ΔV_{pdld} is defined in eq 21. In other words, we evaluate the PDLD energy of an ionized group by averaging it over configurations generated with the charge set to its full final value and to its initial neutral value.

An interesting and crucial element of our averaging procedure is the fact that protein configurations are generated by MD simulation of a consistently solvated protein, where explicit water molecules are present in the protein cavities. This is quite different than recent proposals and attempts of averaging DC results over MD runs (e.g. ref 60) in that such proposals currently do not seem to involve explicit water molecules in the first solvation shell of the protein and its cavities and channels. Simulating charged groups by such a model may suffer from a local collapse of the protein. In our approach the explicit water molecules keep a consistent protein structure during the MD simulations and are converted to Langevin dipoles only after the given configuration is generated. Also, the PDLD model has been used in the buffer regions of our all-atom ENZYMIX program.[14] An analogous attempt to add a DC buffer region to MD simulation programs has recently been made.[60] The actual MD simulations involve running continuous trajectories with 1 fs time steps at 300 K and sending the protein configuration after each 2 ps segment to the PDLD module of POLARIS (see Figure 3).

2.2.2. The PDLD/S Model. The PDLD model provides large microscopic contributions, and the final solvation energy involves very significant compensation effects. Obtaining this compensation is a major challenge that is essential for true understanding of electrostatic energies in proteins.[45] Yet it might be beneficial to obtain more stable results by scaling the microscopic contributions, provided the scaling can be done in a consistent manner, as done in the semi-microscopic PDLD or the scaled PDLD/S approach introduced by Warshel et al.[18] The method, which is described in detail by Lee et al.,[14] assigns to the protein a "dielectric constant", ϵ_p, that represents the contributions that are not included explicitly in the model[44] (as will be argued in section 3.4, this ϵ_p has little to do with the true protein dielectric constant but serves mainly as a scaling factor). The PDLD/S effective potential is obtained from the PDLD energy contributions and is given by[14]

$$\Delta V^{\mathrm{w}\rightarrow\mathrm{p}}_{\mathrm{pdld/s}} = -[\Delta G^{\mathrm{w}}_{\mathrm{qw}} + (\Delta G^{\mathrm{p}}_{\mathrm{qw}}(q=\bar{q}) - \Delta G^{\mathrm{p}}_{\mathrm{qw}}(q=0))]\left(\frac{1}{\epsilon_{\mathrm{p}}} - \frac{1}{\epsilon_{\mathrm{w}}}\right) + (\Delta V^{\mathrm{p}}_{\mathrm{qq}}(q=\bar{q}) + \Delta V^{\mathrm{p}}_{q\mu}(q=\bar{q}))\frac{1}{\epsilon_{\mathrm{p}}} \qquad (24)$$

where $\Delta G^{\mathrm{w}}_{\mathrm{qw}}$ is the self-energy of the given charge in water (the $\Delta G^{\mathrm{w}}_{\mathrm{self}}$ of eq 3), the $\Delta G^{\mathrm{p}}_{\mathrm{qw}}$ term represents the change in the solvation energy of the protein with and without the charged group, and $\Delta V^{\mathrm{p}}_{\mathrm{qq}}$ and $\Delta V^{\mathrm{p}}_{q\mu}$ are the same terms used in the PDLD expression of eq 20. As in the case of the PDLD treatment, we consider the ΔV as free energy when we use a single protein configuration, but we consider it as an effective potential when we average over protein configurations in the more rigorous LRA treatment.

Our PDLD/S approach is implemented in the LRA framework in the same way as the PDLD method described in the previous section. That is, we evaluate the PDLD/S free energy using

$$\Delta G_{\mathrm{pdld/s}} = \frac{1}{2}[\langle \Delta V_{\mathrm{pdld/s}} \rangle_{\mathbf{r}^{\mathrm{p}}_{(q=0)}} + \langle \Delta V_{\mathrm{pdld/s}} \rangle_{\mathbf{r}^{\mathrm{p}}_{(q=\bar{q})}}] \qquad (25)$$

where the average is obtained in the same way as the corresponding average in eq 23.

The PDLD/S has features similar to current DC models (which treat the protein dipoles explicitly) since it also assigns a "dielectric constant" to the protein. However, the consistent LRA treatment of the PDLD/S method is not yet implemented in DC models. Since ϵ_p represents only the factors that are not treated explicitly, the ϵ_p of the PDLD/S method is expected to be smaller than that of the DC models. This point will be considered in section 3.4 and in the Discussion section.

2.2.3. The LRA and LRA/S Approximations. Although the simplified solvent models described above seem to give reasonable results, it is important to relate the relevant energy contributions to the corresponding results obtained by the more rigorous all-atom model. In fact the most rigorous results should in principle be obtained by free energy perturbation (FEP) approaches using all-atom solvent models (e.g. refs 61, 62). Such approaches were used in the first FEP calculations of pK_a's and electrostatic energies in proteins,[40] and in subsequent studies,[14,42] but they are not the subject of the present work. What we like to accomplish here is to use all-atom approaches only as a way to establish the consistency of our semi-microscopic treatment. The simplest and most direct way of relating all-atom to simplified solvent models is the linear response approximation (LRA). That is, simulation studies have indicated that the linear response approximation (which is the basis of macroscopic electrostatic models) is valid even on a microscopic level both in solution[53,63−65] and in proteins.[24,43,61,66,67]

When a system can be described as a collection of harmonic oscillators and therefore follows the LRA approximation, one can use the relationship[67−69] (see also ref 65 for a related derivation)

$$\Delta G_{\mathrm{A}\rightarrow\mathrm{B}} = \frac{1}{2}(\langle V_{\mathrm{B}} - V_{\mathrm{A}} \rangle_{\mathrm{A}} + \langle V_{\mathrm{B}} - V_{\mathrm{A}} \rangle_{\mathrm{B}}) \qquad (26)$$

where V_A and V_B are the potential energies of the system in state A and B, respectively. This equation in the case a single ion (where $V_A = 0$) is converted to the familiar result of $\Delta G = (1/2)\langle V_A \rangle_B$,[3,65] which, of course, is identical to the continuum results. The more general form of eq 26 is not so widely known, and its validity for treatment of ionized groups in proteins and other properties is of significant current interest.[14,52,65,66,68] Furthermore, our motivation in exploring the LRA model is associated with its close relationship to other electrostatics models. In particular we would like to establish in this work the close relationship between the LRA and PDLD models. This point will be demonstrated in the Results section.

The LRA approximation is related to the corresponding FEP treatment (it is just the initial and final integration points in the FEP approach). Thus, one can assume that if the LRA reproduces the FEP results it would give the exact pK_a's. Unfortunately, both the LRA and FEP methods involve major convergence problems and require correct treatment of long-range effects and boundary conditions. One of the most effective ways of obtaining reliable results with a limited number of solvent molecules is the use of spherical boundary conditions with special surface constraint.[3,19] Such constraint should force the finite system to behave as the corresponding region in an infinite system. The present version implemented in the program ENZYMIX[14] is the surface-constrained all-atom solvent (SCAAS) model.[19] This approach emphasizes electrostatic constraints, forcing the polarization of the finite system in response to the field of internal charges, to approximate the polarization of the infinite system.[29] Alternative approaches[70] emphasize correct heat transfer between the system and its surroundings but do not guarantee that the electrostatic response of the finite system will follow that of the complete system. It is also important to recognize that the frequently used periodic boundary conditions do not have the proper symmetry for the treatment of ions.[8]

The present SCAAS version focuses on obtaining a reliable treatment of long-range forces. This is accomplished by dividing the protein/solvent system into regions as described in detail elsewhere[14] and by using the local reaction field (LRF) method.[42] The LRF method allows one to evaluate the results that would have been obtained without any cutoff, while using a relatively small cutoff. Thus, in contrast to many of the available simulation packages, the SCAAS provides a proper electrostatic treatment without the pathologic effect of truncation of long-range forces. It is instructive to point out in this respect that the SCAAS treatment does not only represent the protein and a limited number of water molecules as seems to be implied by ref 43, but considers the protein and an infinite number of water molecules, where some of these solvent molecules are represented explicitly while the outer regions are represented by Langevin dipoles surrounded by a bulk solvent (which is represented by a reaction field model). Such a representation is in fact the new direction in some recently developed approaches.[58] At any rate, this work examines the effect of the long-range treatment on the reliability of pK_a calculations.

The force field used in the present simulation is the standard ENZYMIX force field, which has been described in detail elsewhere.[14] The van der Waals parameters for carboxyl oxygens were modified, however, to account for the use of induced dipole forces (ref 14 considered the energy of induced dipoles but ignored, in most cases, the corresponding induced forces so that the induced energy could be evaluated once in 10 MD time steps), and the present values are 1070.0 Å^6 kcal$^{1/2}$ mol$^{-1/2}$ and 25.0 Å^3 kcal$^{1/2}$ mol$^{-1/2}$, respectively, for the A and B parameters of a negatively charged oxygen atom.

2.2.4. Calculating Charge−Charge Interaction and Ionic Strength Effects. As explained in section 2.1.2, we treat the interaction between ionized groups on a macroscopic level, and only in specific cases, when a given interaction is expected to be large, do we evaluate the microscopic estimate of this interaction. In our macroscopic model we use Coulomb's law and the effective dielectric constant, ϵ_{eff}, of eq 20. It seems to us that the use of Coulomb's law with a large dielectric constant is more justified than the customary DC treatment that involves a small protein dielectric constant. In particular, we believe that the charge−charge interaction in most DC calculations is largely overestimated when these charges are in the interior of proteins (when the charges are near the surface, the results are almost independent of the value of ϵ_p and the effective ϵ is large due to the compensating effect of the solvent). This problem is probably the reason that recent studies[71] were forced to use large values for ϵ_p. Further discussion and examination of this issue will be given in subsequent sections.

When two ionizable groups are in very close proximity, it might be useful to evaluate the relevant ΔG_{ij} explicitly by the PDLD/S procedure, rather than to use our ϵ_{eff}. In doing so we go beyond what is done in current DC treatment and allow the protein to reorganize during the charging process. That is, when we evaluate the ΔG_{ij} of eq 17, we use the LRA approach. This treatment reflects automatically the structural relaxation of the proteins and allows one to use smaller and more consistent values of ϵ_p than what is needed otherwise.

In treating the effect of ionic strength we use a fully macroscopic model, following a previously described procedure.[14] This procedure, which is largely based on an approach of Pack and co-workers,[72] places fractional charges on a grid in the solvent region and evaluates the corresponding probability using a Boltzmann distribution. The interaction between the fractional charge is evaluated with the ϵ_{eff} of eq 20. Some recent aspects of this procedure are described in ref 55, and a validation study is described in ref 14.

3. Results

This section examines the performance of our models for pK_a calculation and focuses on the results of the PDLD/S model, which is closest in spirit to recent DC methods.

3.1. The Semi-microscopic PDLD/S Approach and Current DC Models. The earliest consistent evaluation of pK_a's in protein involved the PDLD study of Asp52 and Glu35 in lysozyme.[9] This was followed by FEP and LRA studies.[40,41,43] The evaluation of the pK_a's of all ionizable groups in lysozyme has recently become a benchmark for pK_a calculations.[33,43,65,73] While we believe that better benchmarks must reflect more emphasis on cases with large pK_a shifts, we felt that it is useful to address this specific benchmark due to its current popularity. Thus we focus here again on the pK_a of the ionizable acids of lysozyme. The starting points for our calculations are the crystal structures of the triclinic (2LZT) and the tetragonal (1HEL) forms of the protein.[74,75] Using these two starting points is a useful way of examining whether the given procedure is able to sample the relevant phase space of the proteins (a perfect approach should give similar results regardless of the starting point). The PDLD/S results are presented in Tables 1 and 2 and compared to the DC results[37,73] in Table 3. As seen from the Table 3 we obtain an improved agreement relative to DC studies where the rms deviation of the PDLD/S model is 0.73 pK_a units as compared to the deviation of 2.07 and 1.58 pK_a units in refs 37 and 73 respectively. This is, however, not the main point of the present work since statistical agreement by itself might be quite misleading (see Discussion section), and even physically inconsistent models can give very good results

TABLE 1: Contributions to the PDLD/S Free Energies and the Corresponding Calculated pK_a's for 1HEL[a]

residue	$\Delta G^p_{q\mu}$	ΔG^p_{qw}	ΔG^p_{bulk}	ΔG^p	ΔG^w_{qw}	$\Delta\Delta G$	pK^p_{int}	$\sum\Delta G_{ij}$[b]	pK^{calc}_a	pK^{obs}_a
7	−4.2	−10.6	−2.0	−16.8	−17.7	0.8	4.9	−1.4	3.5	2.6
18	−8.9	−7.7	−2.1	−18.7	−17.9	−0.8	3.3	−0.9	2.5	2.8−3.0
35	−2.5	−8.8	−2.1	−13.4	−17.1	3.7	7.0	−0.5	6.4	6.1
48	−6.9	−8.2	−1.9	−17.1	−18.0	0.9	4.6	−1.0	3.6	4.3
52	−5.4	−8.1	−2.1	−15.6	−17.5	1.9	5.3	−0.1	5.2	3.5−3.7
66	−13.0	−3.4	−2.1	−18.5	−17.8	−0.7	3.5	−0.3	3.1	1.5−2.5
87	−6.1	−9.9	−2.0	−18.0	−17.7	−0.3	3.7	−0.5	3.2	3.5−3.8
101	1.7	−14.5	−2.0	−14.9	−17.5	2.6	5.8	−1.3	4.5	4.0−4.3
119	−7.7	−8.5	−2.0	−18.1	−17.4	−0.7	3.4	−1.0	2.5	2.2−2.8

[a] Notation as in eqs 21 and 24 but the ΔV are replaced by ΔG since the corresponding terms are evaluated by eq 25. Each ΔG term corresponds to the process AH → A^- in the designated environment. Energies in kcal/mol where each contribution to $\Delta\Delta G$ is already scaled by $1/\epsilon_p$ with $\epsilon_p = 4$. Observed values are taken from ref 93. [b] The contribution of the interaction with all other ionizable residues evaluated in each case for pK_a= pH.

TABLE 2: Contributions to the PDLD/S Free Energies and the Corresponding Calculated pK_a's for 2LZT[a]

residue	$\Delta G^p_{q\mu}$	ΔG^p_{qw}	ΔG^p_{bulk}	ΔG^p	ΔG^w	$\Delta\Delta G$	pK^p_{int}	$\sum\Delta G_{ij}$[b]	pK^{calc}_a	pK^{obs}_a
7	−8.1	−8.1	−2.0	−18.2	−17.6	−0.6	3.9	−1.4	2.5	2.6
18	−6.8	−9.8	−2.0	−18.6	−18.1	−0.5	3.6	−1.1	2.5	2.8−3.0
35	−4.4	−8.4	−2.1	−14.9	−17.3	2.4	6.1	−0.8	5.3	6.1
48	−5.4	−8.1	−2.0	−15.5	−17.3	1.9	5.3	−0.7	4.6	4.3
52	−7.5	−6.9	−2.1	−16.5	−17.8	1.3	4.9	−0.3	4.6	3.5−3.7
66	−12.0	−3.7	−2.0	−17.7	−17.5	−0.2	3.8	−0.3	3.5	1.5−2.5
87	−7.4	−9.3	−2.0	−18.6	−17.6	−1.0	3.2	−0.7	2.5	3.5−3.8
101	4.6	−15.2	−1.9	−12.5	−17.5	5.0	7.5	−1.0	6.5	4.0−4.3
119	−7.6	−7.8	−1.9	−17.3	−16.9	−0.5	3.6	−1.1	2.5	2.2−2.8

[a] Notation as in eqs 21 and 24 but the ΔV are replaced by ΔG since the corresponding terms are evaluated by eq 25. Each ΔG term corresponds to the process AH → A^- in the designated environment. Energies in kcal/mol where each contribution to $\Delta\Delta G$ is already scaled by $1/\epsilon_p$ with $\epsilon_p = 4$. Observed values are taken from ref 93. [b] The contribution of the interaction with all other ionizable residues evaluated in each case for pK_a = pH.

TABLE 3: Calculated pK_a's for Acidic Groups in Lysozyme Obtained by the PDLD/S and Related Macroscopic Models[a]

	DC methods		PDLD/S	exptl	deviations		
residue	pK_a[b]	pK_a[c]	pK_a[d]	pK_a[e]	ΔpK_a[b]	ΔpK_a[c]	ΔpK_a[d]
7	1.7(0.9)	3.6(1.0)	3.0(1.0)	2.6	−0.9	1.0	0.4
18	2.9(0.5)	3.1(1.8)	2.5(0.0)	2.8−3.0	0.0	0.2	−0.4
35	6.3(0.1)	3.2(1.2)	5.9(1.1)	6.1	0.2	−2.9	−0.2
48	1.3(0.6)	1.8(0.1)	4.1(1.0)	4.3	−3.0	−2.5	−0.2
52	7.8(1.5)	4.6(0.8)	4.9(0.6)	3.5−3.7	4.2	1.0	1.3
66	2.0(0.5)	0.7(3.4)	3.3(0.4)	1.5−2.5	0.0	−1.3	1.3
87	1.0(0.4)	2.2(0.1)	2.9(0.7)	3.5−3.8	−1.7	−1.5	−0.8
101	6.1(3.6)	2.6(0.6)	5.5(2.0)	4.0−4.3	1.9	−1.6	1.3
119	2.3(1.9)	3.7(0.2)	2.5(0.0)	2.2−2.8	−0.3	1.1	0.0

[a] The two values reported are, respectively, the average obtained for the triclinic and tetragonal crystal structures and in parentheses the difference between the calculated results for the two crystal structures. [b] Average calculated results of ref 37 for the triclinic and tetragonal structures. [c] Average calculated results of ref 73 for MD relaxed triclinic and tetragonal structures. [d] Average calculated PDLD/S results obtained in the present work for triclinic and the tetragonal structures (1HEL and 2LZT). [e] Data from ref 93.

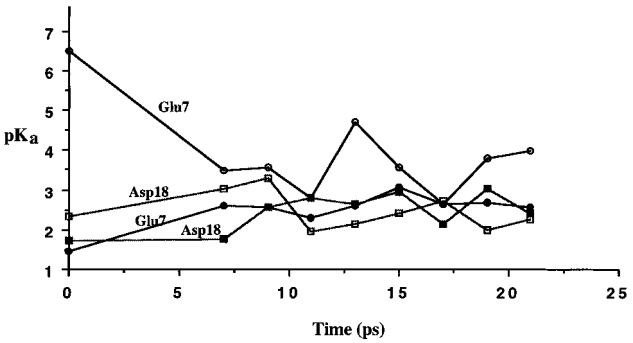

Figure 3. Convergence of the calculated pK_a's as a function of the number of MD relaxation runs (each run took 2 ps). The figure represents the PDLD/S results of eqs 7 and 25 for Asp18 and Glu7 using both the triclinic (open symbols) and the tetragonal (filled symbols) structures. The figure demonstrates how we obtain similar pK_a's after allowing the protein to relax despite starting from different crystal structures.

when the relevant data set involves mainly surface groups. This is reflected by the fact that even the "null" model that assumes a very large dielectric for the protein, or $\Delta pK_a = 0$, will give a small statistical error.[13,71] However, such a model should not be trusted when one deals with ionizable groups in the interiors of proteins and when the corresponding pK_a shifts are large. In such cases one expects significant problems from DC models despite the fact that the current versions of most of these models consider explicitly the microscopic effect of the protein permanent dipoles. The main problem is associated with the missing contribution of the orientational polarization of the protein permanent dipoles to the self-energy of ionized residues. This contribution is in general different in different sites of the protein and cannot be represented by a single dielectric constant. This problem does not exist in the PDLD/S treatment since the effect of dipolar relaxation upon formation of charges is taken automatically and consistently into account by the use of eq 25. The PDLD/S model gives smaller differences between pK_a's obtained from different starting configurations since the average over MD generated structures is more consistent. This point can also be examined by considering Figure 3, which describes the convergence of our approach as a function of the MD relaxation procedure.

3.2. The Relationship between the LRA and PDLD Models. Although the PDLD/S approach yields encouraging results, it is important and in fact crucial to examine more microscopic approaches. A step in this direction is taken by the examination of the PDLD and LRA and the corresponding LRA/S and PDLD/S results, which are summarized in Table 4.

Apparently, as can be seen by inspection of Table 3, the microscopic calculations are at present less accurate than the LRA/S and the PDLD/S results as far as the lysozyme benchmark is concerned (for example, the situation is quite different in the case of highly charged iron−sulfur clusters.[76] However, this is related to the previously mentioned difficulties

TABLE 4: Calculated pK_a's for Acidic Groups of Lysozyme Obtained by the LRA, LRA/S, PDLD, and PDLD/S Models[a]

residue	LRA pK_a	LRA/S pK_a	PDLD pK_a	PDLD/S pK_a	pK_a^{obs}
7	3.0(6.7)	3.5(1.9)	2.4(2.4)	3.0(1.0)	2.6
18	3.0(2.0)	4.0(1.0)	1.6(0.6)	2.5(0.0)	2.9
35	9.2(0.0)	6.5(0.1)	4.3(0.6)	5.9(1.1)	6.2
48	6.0(0.0)	3.5(0.1)	4.1(1.6)	4.1(1.0)	4.3
52	5.6(3.1)	5.9(0.4)	3.6(0.4)	4.9(0.6)	3.6
66	2.6(7.2)	3.5(2.0)	-0.3(0.2)	3.3(0.4)	2.0
87	0.6(1.1)	3.0(1.2)	0.1(1.5)	2.8(0.7)	3.6
101	3.1(2.6)	3.5(0.0)	3.3(4.5)	5.5(2.0)	4.1
119	1.6(5.2)	3.5(2.0)	2.2(0.7)	2.5(0.0	2.5

[a] The two values reported are, respectively, the average obtained for the triclinic and tetragonal crystal structures and in parentheses the difference between the two calculated values. The LRA/S and PDLD/S results are obtained with $\epsilon_p = 4$. The observed values are taken from ref 93.

of obtaining a small error range in microscopic approaches that involve large opposing numbers. Yet the precision of microscopic models might not reflect their true accuracy particularly in cases of large pK_a shifts (see Discussion section). While we are continuously looking for ways to increase the accuracy of the LRA and PDLD methods, the main point of the present analysis is related to the very close similarity between the LRA and PDLD energy contributions, which is demonstrated in Figure 4.

The finding that the PDLD and LRA contributions are so similar is perhaps the best way to establish that the PDLD is indeed a microscopic rather than macroscopic model and also to demonstrate the meaning of a microscopic approach.

As far as the LRA results are concerned, while the agreement between the calculated and observed pK_a's is far from being satisfactory, it represents some improvement over the results obtained in the study of ref 43, probably because of the improved treatment of long-range effects and the inclusion of induced dipoles. The effect of including the LRF treatment and induced dipoles is illustrated in Table 5.

Finally, one of the most instructive points that emerge from the present analysis is the fact that the LRA/S and PDLD/S methods give similar agreement with the observed pK_a (Table 4). This illustrates our point that the accuracy of semimicroscopic models has less to do with the continuum treatment and more to do with the scaling by ϵ_p. That is, both the PDLD and LRA models are less accurate than the corresponding scaled models because the scaling reduces the problems associated with the need to obtain compensation of large energy contributions.[45]

3.3. Interaction between Ionizable Residues. The calculated pK_a's reported in Tables 1 and 2 reflect the effect of interactions between ionized residues that change their ionization state upon change in pH. Thus, the coupling between the ionizable residues should be reflected by the corresponding titration curve. Figure 5 represents single-residue titration curves when the ΔG_{ij} are artificially reduced by increasing the effective dielectric constant, ϵ_{eff}. Although the curves show a modest change upon change of ϵ_{eff}, it seems to us that in many cases it would be quite difficult to deduce the magnitude of the ΔG_{ij} from comparison of the shape of calculated and observed titration curves. This is due to the fact that such curves may reflect the effect of many residues and that sometimes an overestimate of ΔG_{ij} (by underestimating ϵ_{ij}) can be compensated for by a shift in the ionization states of the residues involved. Much more unique results are obtained from mutation experiments where one of the interacting groups is mutated and the ionization state of other groups is determined by NMR or related techniques. Mutation experiments have indicated repeatedly that the effective dielectric, $\epsilon_{eff,}$ for charge−charge interactions in proteins is large even when these groups are buried in protein interiors (for example, see discussion in ref 39). Unfortunately, DC methods with small ϵ_p may underestimate ϵ_{eff} and overestimate the corresponding ΔG_{ij}. This point is illustrated in Table 6 when we evaluate ΔG_{ij} with and without protein relaxation. Table 6 focuses on the largest interaction in the system. The most instructive result of the table is associated with the interaction between Asp52 and Glu35. The interaction is reduced drastically from the unrelaxed value of 6.1 kcal/mol to a relaxed value of 2.6 kcal/mol. Interestingly, the experimental estimate of this interaction (see ref 77) is around 1.8 kcal/mol. As is obvious from our analysis, neglecting the relaxation leads to large values of ΔG_{ij} and forces one to use large values of ϵ_p.

In order to further illustrate this point, we performed PDLD/S calculations of the interactions between Asp210 and Glu213 of the reaction center of *sphaeroides*[39] with and without relaxation (I. Muegge, personal communication). It was found that ΔG_{ij} is reduced from 7.5 to 3.2 kcal/mol when the protein is allowed to relax. This corresponds to an increase of ϵ_{eff} from ∼7 to ∼16. The above discussion does not exclude special cases when ion pairs are strongly stabilized by their local environment (for example, see discussion of $Cys^{-}\cdots His^{+}$ ion pair in papain[61] and $Asp^{-}\cdots Arg^{+}$ in aspartate aminotransferase[78]). However, all these important cases are exceptions rather than rules, as they reflect investment of folding energy that is used to create such functionally important ion pairs.

To prevent misunderstandings, we would like to clarify that we consider a direct evaluation of ΔG_{ij} for groups that are far apart a somewhat unneeded calculation considering the fact that the corresponding interactions are always close to zero (a fact that is properly captured by eq 20). The ability of eq 20 to reproduce experimental results has been repeatedly established in our studies and in those of others (see below), and the question is not whether eq 20 reproduces eq 19 or the corresponding DC results but whether eq 20 reproduces experimental facts. Of course, one would like to establish that eq 19 and explicitly evaluated ΔG_{ij} give small interaction energies, and we have reported such studies before,[17,76] but this is basically a challenging test of the stability of the explicit calculation of charge−charge interactions and not an essential procedure of proving the established fact that these interactions are small and well described by a large effective dielectric constant.

3.4. The Meaning of Protein Dielectric "Constant". The meaning of the "dielectric constant" of proteins has been discussed and analyzed repeatedly (e.g. refs 3, 8), but it still seems to be partially misunderstood and sometimes considered an unimportant semantic issue. Thus, we use the opportunity offered by the present study to reiterate our perspective on the conceptual and practical aspects of this important subject. We will start by summarizing our main points: (i) The physics of enzyme active sites is associated with a polar environment with partially fixed (constrained) permanent dipoles[2] that cannot be captured by using a uniform dielectric medium as originally conceived by TK and other early workers. (ii) The value of the "uniform" constant, $\bar{\epsilon}$, that is obtained from the fluctuations of the total dipole moment of protein regions near charges or in active sites does not correspond to a nonpolar environment. (iii) The dielectric constant, ϵ_p, used in current DC models or in our PDLD/S model has little to do with the protein dielectric constant $\bar{\epsilon}$. All of these points were raised first in our earlier works (e.g. refs 3, 8), and some of them have now been accepted and sometimes adopted and restated. Nevertheless, we will elaborate here on these points.

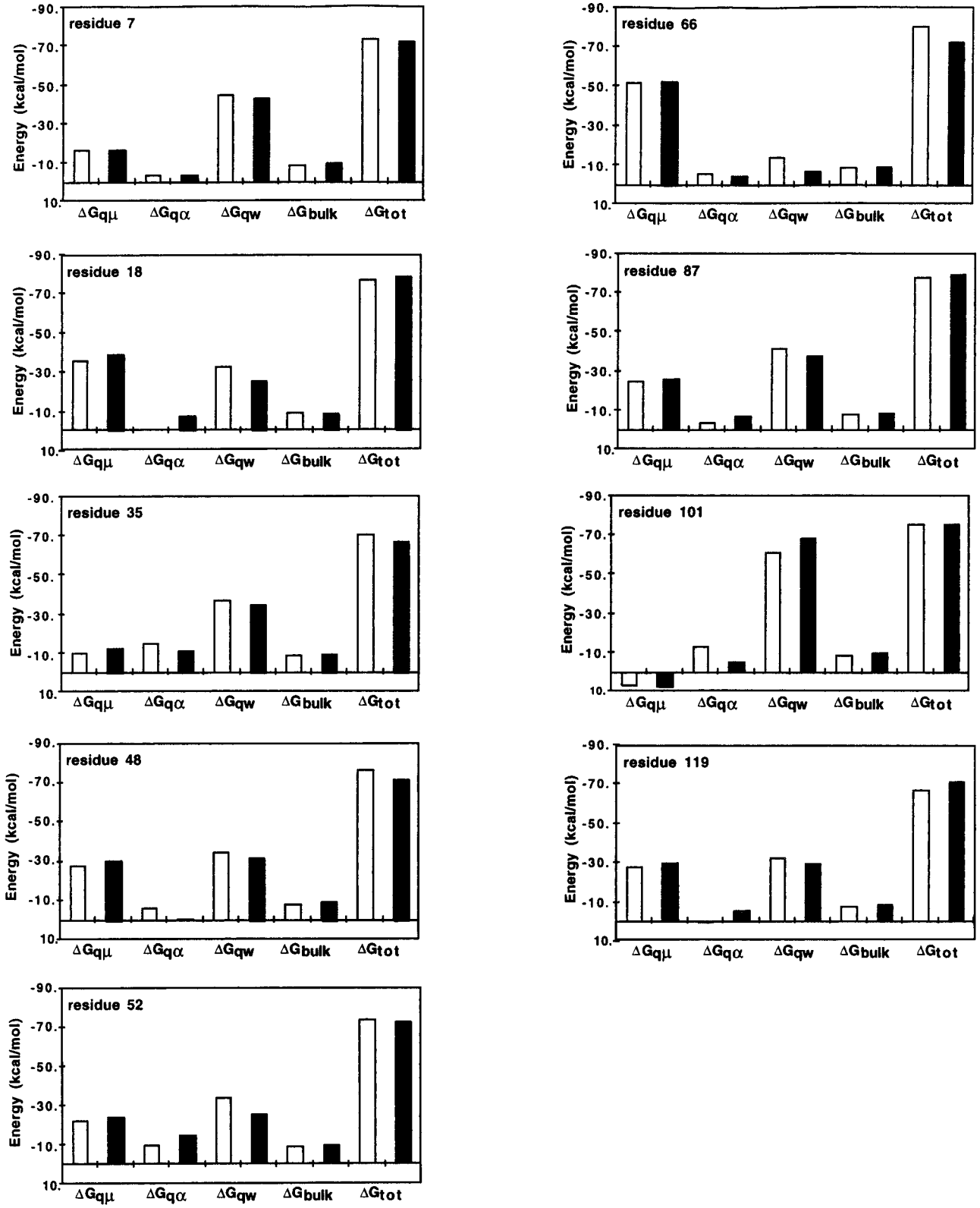

Figure 4. Comparing the PDLD (white) and LRA (black) free energy contributions for different ionizable groups of lysozyme. The calculations are done using the tetragonal crystal structure.

It is important to be clear about the meaning of the "true" "protein dielectric constant", $\bar{\epsilon}$. First of all, there is no such "homogeneous" dielectric constant that can be assigned to all parts of a protein. The macroscopic measurements of protein dielectric constants reflect only the fluctuating part of protein polarity, and they do that only in the *average* sense; that is, they contain no information about fluctuations in any particular part of the protein molecule. The value of this $\bar{\epsilon}$ in regions near ionizable residue is significantly larger than the value $\bar{\epsilon} \cong 2$, that was used in many early studies, and even the "upgraded" value of $\bar{\epsilon} \cong 4$, which already corresponds to a fairly polar environment, does not describe properly the actual value of $\bar{\epsilon}$. Careful simulations[44] revealed that in active sites or even sites around an ionized residue $\bar{\epsilon} > 8$. In special cases when the protein is designed to destabilize charges (e.g. the heme charges in cytochrome *c*) or in regions far from ionized groups one can find relatively small $\bar{\epsilon}$.[44,52,79,80] However, in general $\bar{\epsilon}$ does not correspond to the dielectric constant of a nonpolar environment

TABLE 5: Calculated pK_a's for Acidic Groups of Lysozyme Obtained by the LRA Models with Different Treatments of Long-Range Effects

residue	pK_a[a]	pK_a[b]	pK_a[c]	pK_a[d]	pK_a[e]
7	2.7	3.8	6.0	3.0	2.6
18	3.5	−3.8	−1.3	3.0	2.8−3.0
35	8.9	−1.8	1.6	9.2	6.1
48	−2.6	2.4	4.4	6.0	4.3
52	−4.4	−2.0	−1.3	5.6	3.5−3.7
66	0.8	3.5	2.8	2.6	1.5−2.5
87	−2.8	−2.7	−1.8	0.6	3.5−3.8
101	13.7	9.1	5.3	3.1	4.0−4.3
119	2.2	−6.2	−6.1	1.6	2.2−2.8

[a] Calculated results of ref 43 for the triclinic structure with 15 Å cutoff. [b] Calculated using a cutoff of 8 Å without LRF long-range treatment and also without including the induced-dipole forces in the simulations. Simulations were done for the tetragonal crystal structure. [c] Calculated with the LRF treatment without the induced dipole forces. [d] The results present the average for the triclinic and tetragonal crystal structures. Calculated with induced-dipole forces and with the LRF treatment. [e] Data from ref 93.

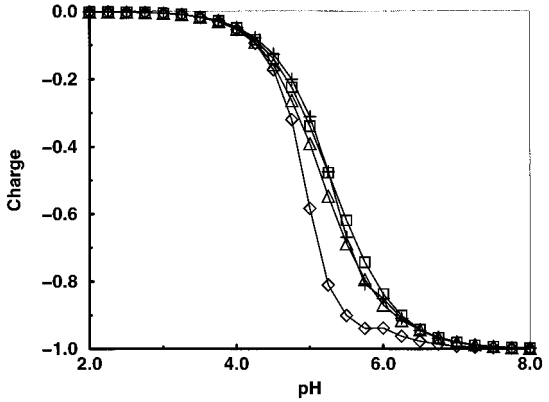

Figure 5. Titration curves for Asp52 with (square) and without the ΔG_{ij} interaction term, calculated by the hybrid approach described in section 2.1.2 using an effective dielectric function of 40 (diamond), 80 (triangle), and $\epsilon_{ij} = \epsilon_{\text{eff}} = 1 + 60[1 - \exp(-0.1r_{ij})]$ (cross).

TABLE 6: PDLD/S Estimates of the Coupling between Some Ionizable Residues in Lysozyme[a]

	interaction free energy	
residue pair	unrelaxed	relax
35−52	6.11	2.57
48−52	1.57	0.74
52−66	2.33	1.91

[a] The calculations were done starting from the tetragonal crystal structure (1hel) considering only pairs with unrelaxed interaction of more than 1 kcal/mol. Interaction energies are given in kcal/mol. Unrelaxed and relaxed designate the results obtained using the original crystal structure and the results obtained with MD relaxation of the charged and uncharged forms. The relaxed results were averaged over eight configurations obtained at 2 ps intervals.

(i.e. $\bar{\epsilon} \cong 2$) or even to $\bar{\epsilon} \cong 4$, and simulation studies[36] that obtain $\epsilon \approx 4$ have apparently omitted the effect of the reaction field around the protein that drastically increases $\bar{\epsilon}$ (see discussion and demonstration of this point in Figures 2 and 5 of ref 44). Similarly, attempts to attribute the large value of $\bar{\epsilon}$ obtained by reliable simulations[81] exclusively to ionized surface groups are not justified. That is, the consistent calculations of ref 44 produced $\bar{\epsilon} \approx 9$ in the active site of trypsin without any ionized surface groups (to the best of our knowledge this point has not been examined by other consistent calculations that include the solvent around the protein). Thus, the correct $\bar{\epsilon}$ for proteins can be significantly larger than the small value favored by those who view proteins as low dielectric environments.

At this point, it is crucial to remember that the dielectric constant $\bar{\epsilon}$ does not represent a linear measure of polarity. Such a measure is more properly provided[3] by $(1 - 1/\bar{\epsilon})$. Therefore even $\bar{\epsilon} = 4$ represents a fairly polar environment in the "macroscopic" dielectric sense. However, the issue of protein polarity cannot be properly addressed solely by conventional continuum dielectric concepts because of the preorganized nature of the dipoles within the protein. This fact should be self-evident because proteins often have to provide better solvation than the surrounding water,[2] and that would correspond to unphysical, negative dielectric constants (because of the high dielectric constant of water and the approximate proportionality of solvation to $(1 - 1/\bar{\epsilon})$). Instead, large solvation energies are provided by a combination of "nonpolar" dielectric constants and preoriented dipoles. The important message here is that neither {homogeneous, low dielectric + preoriented dipoles} nor {high dielectric over the whole protein} pictures can offer consistent explanations of protein polarity. Also, due to both the inhomogeneous and preoriented nature of the interior of a protein, the apparent "dielectric constant" depends on the property studied[3] and on the specific site and cannot be represented consistently by a single value. Nevertheless, one may still wonder, what is the origin of the reasonable results obtained with the DC and PDLD/S models that use a small value of ϵ_p? The answer to this important question has been given before (e.g. refs 8, 44), but it perhaps needs to be restated. The optimal ϵ_p is not the elusive $\bar{\epsilon}$, but simply a parameter that represents contributions that are not included explicitly. To see this point, one can start from a model where all the microscopic effects are considered explicitly. In this case obviously $\epsilon_p = 1$. Now if only the induced dipoles are not included explicitly, we will have $\epsilon_p = 2$, and if the entire protein and water are treated implicitly, then $\epsilon_p \geq 40$.[8,44] With this point in mind we may wonder, what is the optimal ϵ_p when the induced dipoles and the protein relaxation are included implicitly and when internal and external solvent molecules are represented by a DC model?[82] The answer to this question is not unique since it obviously depends on the specific reorganization in each site. In the PDLD/S model we avoid a significant part of this issue by treating explicitly the reorganization of the permanent dipoles.

3.5. The Meaning of ϵ_{eff}. This work and many of our earlier works use the ϵ_{eff} of eq 20 to estimate interactions between ionized residues. The large values of ϵ_{eff} does not reflect arbitrary assumptions but rather are the results of a long series (see for example pages 347−364 in ref 3) of computational and theoretical studies and their experimental verifications, including rather rigorous and physically consistent PDLD and FEP calculations (e.g. refs 17, 76). Despite these works it seems that the underlying microscopic physics of ϵ_{eff} is not fully appreciated (see commentaries in refs 84 and 85). Some might assume that ϵ_{eff} simply reflects the effective interactions obtained by a macroscopic model with small ϵ_p in the protein region and high dielectric constant for the solvent region or, in other words, that ϵ_{eff} only reflects the effect of the solvent around the protein. Such an approach might reflect in fact the confusion between the rigorous results of an *assumed* model and the actual physics of a real protein.[85] It is important to understand that microscopic considerations of charge separation are the only way to understand the origin of ϵ_{eff}. Such considerations do show that ϵ_{eff} reflects the compensation between vacuum charge−charge interaction and the protein reorganization and solvent penetration, as described in Figure 27 of ref 3. That is, ϵ_{eff} reflects reorientation of the protein dipoles that cannot be captured by macroscopic concepts, where the compensation is assumed rather than obtained.

TABLE 7: Calculated Intrinsic pK_a's for a Hypothetical Nonpolar Lysozyme[a]

residue	ΔG^p_{qw}	ΔG^p_{bulk}	ΔG^p	ΔG^w	p$K_{a,int}$
7	−24.0	−4.0	−28.0	−35.2	9.5
18	−23.0	−4.0	−27.0	−35.1	9.8
35	−19.8	−4.0	−23.8	−35.3	12.6
48	−19.3	−4.0	−23.3	−34.4	11.9
52	−21.0	−4.0	−25.0	−35.4	11.4
66	−12.7	−4.0	−16.7	−35.3	17.4
87	−25.1	−4.0	−29.1	−35.2	8.3
101	−27.2	−4.0	−31.2	−34.3	6.2
119	−24.4	−4.0	−28.4	−34.2	8.1

[a] Energies in kcal/mol, where each energy contribution is already scaled by $1/\epsilon_p$ with $\epsilon_p = 2$ (which corresponds to the nonpolar protein environment).

4. Discussion

Calculations of pK_a's of ionizable groups in proteins present a major challenge. On the one hand, microscopic approaches suffer from convergence problems since they involve large opposing contributions, and on the other hand, fully macroscopic models cannot take into account correctly the protein microenvironment.[13] Recent DC methods that treat the local environment in a semi-microscopic way seem to provide reasonable pK_a values, but these methods still suffer from a fundamental inconsistency since they do not take into account the protein relaxation upon the charging process and this relaxation cannot be represented by a single unique dielectric constant. The best way to realize this point is to consider a case where crystal structure was obtained at a pH where a given ionizable group is in its neutral state and where the dipoles around this group are not pointing toward it. A DC calculation that uses the crystal structure will miss the effect of reorganization of the local dipoles upon ionization of the given group. In this case, we will not have any "back field" (the $V_{q\mu}$ term of eq 21 will be zero), and accounting for the missing effect of the local dipoles will require a high value of ϵ_p. Unfortunately, this large ϵ_p will not correspond to the ϵ_p in sites where the ionizable groups are ionized in the crystallization process. If, on the other hand, one could take into account the microscopic reorganization process, one should be able to consistently use a small value of ϵ_p that will only reflect the missing induced dipoles and perhaps incomplete penetration of solvent molecules.[44] The present paper presents the results of the semi-microscopic PDLD/S methods that treat the protein reorganization effect in a consistent way. Although it is very hard to obtain perfect results in pK_a calculations, it appears that the PDLD/S approach is not only more consistent but also gives somewhat better results than current alternative DC models (this point will be further discussed below). It is useful to note, in this respect, that using a consistent approach also reduces the difference between the results obtained for different crystal structures, as is established in Table 3 and Figure 3.

This work reemphasizes the crucial role of self-energy of the ionizable residues in proteins.[3,13] Although this factor is starting to be widely appreciated (e.g. refs 38, 43), it is still not uncommon to see in the literature assumptions implying that it is a relatively small, second-order effect.[86] However, as has been established in our early estimates (e.g. refs 13, 17) and in any correct subsequent studies, the self-energy of an ionized acid in the interior of a hypothetical nonpolar protein would be around 35 kcal/mol (~25 pK_a units) smaller than in water. Such an enormous shift in pK_{int} is what would be obtained for internal ionizable groups that are treated by DC models that do not include explicitly the protein permanent dipoles. As a case in point it is useful to consider the calculations presented in Table 7, which present the pK_a shifts that would have been obtained if lysozyme were an entirely nonpolar protein and if no relaxation and water penetration were allowed. As seen from Table 7, even in this case (where many ionizable groups are not far from the surface) we have very large pK_a shifts. Of course, one may argue that many ionizable residues are usually located near the surface of proteins, and therefore, models that treat protein as a nonpolar environment are not so unrealistic. However, the point in developing models for pK_a calculations in proteins is the elucidation of the electrostatic energy of groups with large pK_a shifts that are located deep in protein interiors, rather than the trivial issue of finding that the pK_a's of surface groups are not shifted significantly (where any model, including entirely incorrect models, would work[13]). Finally, since the issue of self-energy might seem now rather obvious, it is important to realize that this term was missing in the pioneering TK work; the B_{kk} term in this work did not include the radius of the ionized group but the radius of the protein (see ref 13).

The relationship between pK_a and protein conformations has been the subject of recent discussion (e.g. refs 38, 83, 87, 88). It has been argued that the large differences between pK_a's calculated using different crystal structures indicate the importance of conformational effects.[37] It was also argued that accounting for the protein conformational flexibility should allow one to use a "physically reasonable" low dielectric model.[83,87] There are, however, some problems with the perspective of these proposals. pK_a simply reflects the average effect (free energy) of all the relevant conformations. Thus, the issue is how to obtain an average rather than pointing out (correctly) that the energy values used in the averaging procedure depend on the corresponding structures. Furthermore, taking different crystal structures at their face value will, of course, produce large pK_a changes. However, these results will not correspond to the actual pK_a measured in the given crystal (if such measurement is made possible), which would reflect the relaxation of the local dipoles upon ionization. This relaxation effect is reproduced by our LRA approach and can also be obtained directly if the crystal structure of the protein in its ionized and neutral forms are known. A related study of the reorganization energy of cytochrome *c* was reported recently.[52] At any rate, when the proper local relaxation is considered, one should expect a smaller difference between the pK_a's of different crystal structures, as is indeed the case in the present work. Another closely related issue is the above mentioned suggestion that averaging over protein configurations will lead to a more consistent ϵ_p.[83,87] It seems to us that conformational averaging by itself should not lead to any improvement in the calculated values except in providing more robust results. What is needed (as was argued and already demonstrated in our previous works[59,89] and in the present work) is MD averaging on the ionized and neutral states. Such a consistent implementation of the LRA approach allows the use of smaller value of ϵ_p. However, this has less to do with more physically consistent ϵ's (see section 3.4) and more to do with having more effects treated explicitly.

There is apparently still some interest in implying that the PDLD model must have been fundamentally changed (perhaps reflecting the fact that this model appeared so early in the development of the field). However, no fundamental concept has been changed in this model which is basically a model, of explicit dipoles on a grid. The model has been reparametrized in different versions and refinement states, as should be done with any microscopic model (for example, see repeated refinement of all current MD force fields) and even with macroscopic models. Perhaps it is very hard to realize what dipolar models are all about without trying them and seeing their robustness and insensitivity to details (of course with proper parametriza-

tion). Thus it may be useful for those who might have conceptual problems with the PDLD and related dipolar models to just try an LD program (e.g. ref 57). Another related issue is the perception that the PDLD/S is a new model that reflects a departure from the PDLD model. However, the PDLD/S is not a new model, as it has been introduced in ref 18, and it is perhaps the first semi-macroscopic model to correctly include the protein dipoles; and its present version presents the first semi-macroscopic model to correctly include the protein reorganization. Obviously, the PDLD/S is a different model than the PDLD in the same way that a FEP all-atom model is different from the PDLD. In fact, using different models is extremely useful for comparative studies provided they are treated consistently. The PDLD model has larger average error in the rather trivial case of surface groups than the PDLD/S or the null model does. The same is true for the LRA and FEP or any other microscopic model. However, microscopic models are developed and refined since in principle they are more realistic and must eventually be more reliable (for treating nontrivial internal groups) once the convergence problems are overcome. Furthermore, the PDLD and other dipolar models are expected to be more reliable for the truly challenging cases of strong ion pairs and highly charged clusters (e.g. the iron−sulfur protein study of ref 90). Finally, just to give this issue a proper perspective, it should be noted that the PDLD model had an error range of ∼3 kcal/mol in the 1970s when continuum models had an inherent error of 30 kcal/mol (see comparative study in Table 7 of ref 14).

The present work examined the effect of charge−charge interaction in lysozyme and concluded that these interactions are usually small. The reason for the large effective dielectric (ϵ_{eff}) for charge−charge interactions is associated with the ability of the protein and its surrounding solvent to compensate for the change in energy associated with charge separation (see for example Figure 3 of ref 13). This ability, which is partially reproduced by our LRA approach, might not be captured by DC models with small assumed ϵ_p. For example, when one deals with interactions between charges that are located far from the protein surface, ϵ_{eff} starts to approach the assumed ϵ_p. Using small ϵ_p might lead in such cases to a significant overestimation of the interaction between ionized groups. In this respect it is useful to comment about the recent conclusion[71] that ϵ_p in DC models should be quite large (i.e. $\epsilon_p \approx 20$). This conclusion probably reflects the attempt to account for the small value of ΔG_{ij} by DC models. Using such an empirically based ϵ is fully justified,[13,55,91] when one deals with charge−charge interactions. However, when one uses DC methods the ϵ_p that gives optimal ΔG_{ij} is not the one that gives the best values for the self-energies. This is the reason for the relatively poor results obtained for Glu35 with large ϵ_p.[83] Only approaches that account consistently for the protein relaxation can hope to have the same ϵ_p for pK_{int} and ΔG_{ij}. Trying to obtain the best ϵ_p by optimizing a large data-base of pK_a's can be quite misleading since most ionizable groups considered are surface groups, where the effect of ϵ_p on pK_{int} is rather small. If one really wishes to examine the consistency of different dielectric models, one should focus on internal groups with large pK_a shift rather than on surface groups.

The microscopic validity of the PDLD model has been established in this work by illustrating a very good agreement between its contribution to the self-energy and those of the all-atom LRA model (see Figure 3). As a part of the analysis we also examined the accuracy of the LRA results. At present it appears that the microscopic LRA treatment does not give sufficiently accurate results. However, we were able to establish that improved boundary conditions, inclusion of induced dipoles, and averaging over initial conditions increase the accuracy of the calculated results.

This work demonstrates that the scaling of the LRA energies according to the LRA/S formulation leads to results as accurate as those of the PDLD/S model. This finding provides a clear support to our argument[8] that obtaining better precision by macroscopic models than by microscopic models has less to do with the physics of the macroscopic models and more to do with the scaling of large compensating numbers.

The approach used in the present study allows one to explore the relationship between microscopic and semi-microscopic models in a direct and consistent way. In fact, we can easily use our method to, completely or partially, move from explicit water to simplified water models in the protein interior. This can be done by arbitrarily treating a given number of all-atom water molecules as a part of this "protein" system (region II in the notation of ref 14). Since the all-atom solvent model is always running in the background of the PDLD treatment (the LRA all-atom treatment is used in generating the protein configurations for the PDLD calculations), we have no consistency problem. This is, however, not the case in recent attempts to add explicit solvent molecules to DC calculations (e.g. ref 92). It seems that in such cases the solvent is aligned arbitrarily without a clear energy criterion and different results would be obtained with different assumptions. Of course, arbitrary addition of solvent molecules cannot describe properly the effect of solvent reorganization during the charging process.

The role of charge−charge interactions in proteins is of significant interest (e.g. ref 55). However, it seems that such interactions are significantly smaller than what is usually assumed; except in special cases when the protein is designed to stabilize ion pairs (e.g. ref 78) or when ionizable groups that are located at a nonpolar environment become charged upon change of pH.[13,83] As is illustrated in this paper, models that take into account the reorganization of the protein dipoles (in response to the development of the charges of the interacting groups) can allow one to consistently use smaller values of ϵ_p. It is also pointed out here that overestimates of ΔG_{ij} can be easily overlooked since observed titration curves can be reproduced with incorrect values of the charge−charge interaction terms. Only careful mutation experiments can be used to determine charge−charge interactions, and such experiments are repeatedly pointing toward small charge−charge interactions.

As argued above, true DC approaches such as the TK model cannot describe correctly the energetics of ionizable groups in proteins. Including the protein permanent dipoles makes such models more microscopic and starts to capture the correct physics of electrostatic energy in proteins. However, current DC models do not account at present for the microscopic effect of the protein reorganization energy. We predict, however, that eventually such models will incorporate the reorganization effect (as done in the PDLD/S model) and become more microscopic, resembling more and more the PDLD/S and LRA/S models. In this respect the fact that the solvent around the protein is modeled by a continuum has no fundamental consequences since the same or very similar results would be obtained by almost any model of the surrounding, provided it leads to a large effective dielectric for charge−charge interactions. The difference in physics is in the treatment of the microenvironment inside the protein, and this must be represented with sufficient microscopic reality. Of course one may argue that the given model is still macroscopic, but this involves significant conceptual problems. For example, arguing that treating the reorientation of the protein dipoles microscopically (as done here) is still macroscopic is problematic, since the resulting dielectric constant would have no relationship to the protein

dielectric constant $\bar{\epsilon}$; such a treatment would produce a dielectric constant of $\epsilon_w = 2$ (containing only the electronic part) for water instead of $\bar{\epsilon}_w = 80$. At any rate, regardless of the "label" of different models it is clear that the representation of the protein polar environment and its reorientation is a crucial requirement of consistent protein models.

Acknowledgment. This work was supported by Grant GM24492 of the National Institutes of Health.

References and Notes

(1) Perutz, M. F. *Science* **1978**, *201*, 1187.
(2) Warshel, A. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 5250.
(3) Warshel, A.; Russell, S. T. *Q. Rev. Biol.* **1984**, *17*, 283.
(4) Sharp, K. A.; Honig, B. *Ann. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301.
(5) Matthew, J. B. *Annu. Rev. of Biophys. Biophys. Chem.* **1985**, *14*, 387.
(6) Nakamura, H. *Q. Rev. Biophys.* **1996**, *29*, 1.
(7) Warshel, A. *Acc. Chem. Res.* **1981**, *14*, 284.
(8) Warshel, A.; Åqvist, J. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 267.
(9) Warshel, A. *Biochemistry* **1981**, *20*, 3167.
(10) Linderstrom-Lang, K. *C. R. Travl. Lab. Carlsberg* **1924**, *15*, 1.
(11) Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333.
(12) Tanford, C.; Roxby, R. *Biochemistry* **1972**, 2192.
(13) Warshel, A.; Russell, S. T.; Churg, A. K. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 4785.
(14) Lee, F. S.; Chu, Z. T.; Warshel, A. *J. Comput. Chem.* **1993**, *14*, 161.
(15) States, D. J.; Karplus, M. *J. Mol. Biol.* **1987**, *197*, 122.
(16) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.
(17) Russell, S. T.; Warshel, A. *J. Mol. Biol.* **1985**, *185*, 389.
(18) Warshel, A.; Naray-Szabo, G.; Sussman, F.; Hwang, J. K. *Biochemistry* **1989**, *28*, 3629.
(19) King, G.; Warshel, A. *J. Chem. Phys.* **1989**, *91*, 3647.
(20) Coalson, R. D.; Duncan, A. *J. Phys. Chem.* **1996**, *100*, 2612.
(21) The PDLD model has attracted criticism from some workers who suprisingly had little difficulties in accepting macroscopic models as realistic representations of macromolecules while sometimes finding the physics of simplified molecular models hard to accept. Some of this criticism[22] has occasionally involved "selective citations" (e.g. referring readers to the incorrect arguments of ref 23, whose deficiencies should have been quite clear (see in footnote 56 in ref 24). The criticism has continued even recently,[25] when it was incorrectly suggested that the PDLD model of ref 16 did not evaluate self-energies and argued that this model is in fact macroscopic in nature and that it is physically equivalent to the use of a finite-difference grid within DC methods. These assertions are incorrect both in presentation of facts and in interpretation of the physics of dipolar solvents. A grid of dipoles with finite spacing or the equivalent system of polar molecules on a lattice is not equivalent to the numerical grid used in evaluating the electrostatic potential in DC approaches (see ref 26), although as the lattice spacing is reduced to unrealistically small values, the average behavior of *simple* dipole lattices predictably approaches that of macroscopic models.
(22) Davis, M. E.; McCammon, J. A. *Chem. Rev.* **1990**, *90*, 509.
(23) Krishtalik, L. I. *Mol. Biol. (Moscow)* **1984**, *18*, 892.
(24) Yadav, A.; Jackson, R. M.; Holbrook, J. J.; Warshel, A. *J. Am. Chem. Soc.* **1991**, *113*, 4800.
(25) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578.
(26) The DC grid tries to represent the average electric potential in a system by assuming a continuum description of matter. The PDLD, on the other hand, retains a microscopic dipolar representation, where the field varies enormously between grid points (see Figure 1 of ref 13 and Figure 9.20 in Purcell's classical book[27]). No such variation of fields exists in the continuum formulations or in the DC treatments. It may be useful in this respect to consider an example of how a microscopic grid of dipoles is used in analytically solvable cases to obtain continuum results (e.g. ref 28). Since dipolar models provide a more "realistic" description, they can be reduced to a continuum description, but the reverse is not possible. Macroscopic models involve the use of dielectric constants, while no such scaling is used in the PDLD model. The use of the dipolar model has facilitated a consistent description of protein electrostatics quite early and is probably the reason why the PDLD model did not fall into all the traps that plagued macroscopic treatments of proteins; for example looking at the microscopic interactions led immediately to the explicit consideration of the protein permanent dipoles rather than to futile attempts of representing them by a low dielectric constant.
(27) Purcell, E. M. *Electricity and Magnetism*; McGraw-Hill: New York, 1965.
(28) Jackson, J. D. *Classical Electrodynamics*; Wiley: New York, 1975.
(29) Warshel, A. *J. Phys. Chem.* **1979**, *83*, 1640.
(30) Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1991**, *113*, 8305.
(31) Rashin, A. A. *J. Phys. Chem.* **1990**, *94*, 1725.
(32) The accuracy of DC pK_a calculations in *solutions* has been contrasted recently[33] with the accuracy of PDLD calculations of pK_a's *in proteins*[17] rather than with PDLD calculations of pK_a's in solutions (which were readily available, e.g. ref 29). Obviously, it is straightforward to obtain perfect pK_a results in solutions with adjustable Born radii, but the real challenge is to obtain reliable results for pK_a's in proteins where the van der Waals or Born radius cannot be adjusted to reproduce pK_a's of identical groups in different environments.
(33) Lim, C.; Bashford, D.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 5610.
(34) Warwicker, J.; Watson, H. C. *J. Mol. Bio.* **1982**, *157*, 671.
(35) Rogers, N. K. *Prog. Biophys. Mol. Biol.* **1986**, *48*, 37.
(36) Gilson, M.; Honig, B. *Biopolymers* **1986**, *25*, 2097.
(37) Bashford, D.; Karplus, M. *Biochemistry* **1990**, *29*, 10219.
(38) Yang, A. S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. *Proteins: Struct., Funct., Genet.* **1993**, *15*, 252.
(39) McPherson, P. H.; Schonfeld, M.; Paddock, M. L.; Okamura, M. Y. *Biochemistry* **1994**, *33*, 1181.
(40) Warshel, A.; Sussman, F.; King, G. *Biochemistry* **1986**, *25*, 8368.
(41) Merz, Ke. M. *J. Am. Chem. Soc.* **1991**, *113*, 3572.
(42) Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1992**, *97*, 3100.
(43) Buono, G. S.; Figueirido, F. E.; Levy, R. M. *Proteins: Struct., Funct., Genet.* **1994**, *20*, 85.
(44) King, G.; Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1991**, *95*, 4366.
(45) Dipolar models and any other explicit representation of molecular systems have to reproduce the compensating effect of opposing electrostatic interaction such as charge−charge and solvation effects by a balance of large contributions of opposite magnitude. On the other hand, macroscopic models obtain such effects by assuming the dielectric constant of the given medium rather than obtaining it microscopically. Obviously, it is more challenging to obtain precise results from microscopic models due to well-known convergence problems, but the chances of having the incorrect physics in nonhomogenous systems are much larger when one uses macroscopic models, since the correct continuum representation of a given microscopic system is far from obvious.
(46) Jorgensen, W. L.; Briggs, J. M. *J. Am. Chem. Soc.* **1989**, *111*, 4190.
(47) Lancaster, C. R. D.; Michel, H.; Honig, B.; Gunner, M. R. *Biophys. J.* **1996**, *70*, 2469.
(48) Beroza, P.; Okamura, M. Y.; Feher, G. *Proc. Natl. Acad. U.S.A.* **1991**, *88*, 5804.
(49) Warshel, A. In *Methods in Enzymology*; Packer, L., Ed.; Academic Press, Inc.: London, 1986; Vol. 127, p 578.
(50) Warshel, A. *Photochem. Photobiol.* **1979**, *30*, 285.
(51) Daggett, V.; Kollman, P. A.; Kuntz, I. D. *Biopolymers* **1991**, *31*, 1115.
(52) Muegge, I.; Qi, P. X.; Wand, A. J.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 825.
(53) King, G.; Warshel, A. *J. Chem. Phys.* **1990**, *93*, 8682.
(54) Papazyan, A.; Warshel, A. *Submitted.*
(55) Alden, R. G.; Parson, W. W.; Chu, Z. T.; Warshel, A.*J. Am. Chem. Soc.* **1995**, *117*, 12284.
(56) The assumption that dipolar models might be incorrect continues even now, and a referee of this paper has suggested that somehow we are now forced to abandon the PDLD and to use eq 22. We had not anticipated that such a notion still exists when we move to eq 22, and we thus would like to stress that (a) the Langevin function gives somewhat better results than those obtained by eq 22 for highly charged systems; (b) any model with sufficiently large α_L gives very similar solvation energies, as should be clear from the Born equation; and (c) many versions of our approach still include the Langevin function, including those with a publicly available source (ref 57), and it will probably be useful if those who have problems with dipolar models would try a computer program with dipoles around a charge and realize what the properties of such models are.
(57) The LD model has been used in a recent work (J. Florian and A. Warshel, submitted) that provides a reliable way of obtaining solvation free energies from the Gaussian program package. This program can be obtained from the authors.
(58) Smith, P. E.; van Gunsteren, W. F. In *Computer Simulation of Biomolecular Systems*; van Gunsteren, W. F.; Weiner, P. K.; Wilkinson, A. J.; Eds.; Escom: Leiden, 1993; Vol. 2; p 182.
(59) Langen, R.; Brayer, G. D.; Berghuis, A. M.; McLendon, G.; Sherman, F.; Warshel, A. *J. Mol. Biol.* **1992**, *224*, 589.
(60) Sharp, K. *J. Comput. Chem.* **1991**, *12*, 454.
(61) Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solutions*; John Wiley & Sons: New York, 1991.
(62) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395.
(63) Warshel, A. *J. Phys. Chem.* **1982**, *86*, 2218.
(64) Hwang, J. K.; Warshel, A. *Biochemistry* **1987**, *26*, 2669.
(65) Levy, R. M.; Belhadj, M.; Kitchen, D. B. *J. Chem. Phys.* **1991**, *95*, 3627.
(66) Åqvist, J.; Medina, C.; Samuelsson, J.-E. *Protein Eng.* **1994**, *7*, 385.
(67) Lee, F. S.; Chu, Z. T.; Bolger, M. B.; Warshel, A. *Protein Eng.* **1992**, *5*, 215.
(68) Warshel, A.; Hwang, J. K. *J. Chem. Phys.* **1986**, *84*, 4938.

(69) Warshel, A.; Chu, Z. T.; Parson, W. W. *Science* **1989**, *246*, 112.
(70) Brooks, C. L.; Karplus, M. *Biopolymers* **1985**, *24*, 843.
(71) Gilson, M. K. *Curr. Opin. Struct. Biol.* **1995**, *5*, 216.
(72) Klein, B. J.; Pack, G. R. *Biopolymers* **1983**, *22*, 2331.
(73) Yang, A. S.; Honig, B. *J. Mol. Biol.* **1993**, *231*, 459.
(74) Ramanadham, M.; Sieker, L. C.; Jensen, L. H. *Acta Crystallogr. Sect. A* **1981**, *37c*, 33.
(75) Wilson, K. P.; Malcolm, B. A.; Matthews, B. W. *J. Biol. Chem.* **1992**, *267*, 10842.
(76) Cutler, R. L; Davies, A. M.; Creighton, S.; Warshel, A.; Moore, G. R.; Smith, M.; Mauk, A. G. *Biochemistry* **1989**, *28*, 3188.
(77) Thoma, J. A. *J. Theor. Biol.* **1974**, *44*, 305.
(78) Hwang, J. K.; Warshel, A. *Nature* **1988**, *334*, 270.
(79) Simonson, T.; Perahia, D. *J. Am. Chem. Soc.* **1995**, 117, 7987.
(80) Simonson, T.; Perahia, D. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 1082.
(81) Smith, P. E.; Brunne, R. M.; Mark, A. E. *J. Phys. Chem.* **1993**, *97*, 2009.
(82) A recent study[83] has suggested that the "disturbing consequences" of using large ϵ_p may eventually disappear if the effect of protein dynamics will be taken into account and a more realistic small value of ϵ_p could be used again. This assertion has several problems. It reflects the assumption that ϵ_p is equal to the "macroscopic" $\bar{\epsilon}$ of the protein. This is not true since ϵ_p simply represents the part of the protein polarity that is not *explicitly* modeled. Furthermore, the "protein relaxation" should have been included in the of the protein in a continuum description (as $\bar{\epsilon}$ measures the "fluctuating" part of polarity). Increasing the explicitness of a model should lead to smaller ϵ_p (eventually to $\epsilon_p = 1$ when the whole system is treated explicitly), but that has nothing to do with the ability of a continuum description to reconcile the microscopic solvation in a heterogenous environment with the macroscopic $\bar{\epsilon}$.
(83) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *J. Mol. Biol.* **1994**, *238*, 415.
(84) Gunner, M. R.; Alexov, E.; Torres, E.; Lipovaca, S. *J. Biol. Inorg. Chem.* **1997**, *2*, 126.
(85) Warshel, A.; Papazyan, A.; Muegge, I. *J. Biol. Inorg. Chem.* **1997**, *2*, 143.
(86) A case in point is the discussion in ref 37 where the importance of the self-energy term was downplayed. There it was suggested that our early estimates[17] of $\Delta pK_a \approx 30$ kcal/mol for moving an ionizable group to the center of a hypothetical nonpolar protein largely overestimated the actual effect and that the calculation was done for a "small group". This overlooks the simple fact that the Born radius used in our consideration is the proper Born radius that reproduces the solvation free energy of Asp^- (a typical rather than a "small" ionizable group) in solution exactly. Thus, any correct estimate of the energetics of moving an ionizable acid to the center of a protein would reproduce our results, confirming the primary importance of the solvation term in the energetics of proteins.
(87) You, T. J.; Bashford, D. *Biophys. J.* **1995**, *69*, 1721.
(88) Allewell, N. M.; Oberoi, H. *Methods Enzymol.* **1991**, *202*, 3.
(89) Jensen, G. M.; Warshel, A.; Stephens, P. J. *Biochemistry* **1994**, *33*, 10911.
(90) Stevens, P. J.; Jollie, D. R.; Warshel, A. *Chem. Rev.* **1996**, *96*, 2491.
(91) Muegge, I.; Schweins, T.; Langen, R.; Warshel, A. *Structure* **1996**, *4*, 475.
(92) Gibas, C. J.; Subramaniam, S. *Biophys. J.* **1996**, *71*, 138.
(93) Kuramitsu, S.; Hamaguchi, K. *J. Biochem.* **1980**, *87*, 1215.