Supporting Information:

# Reproducing the Ensemble Average Polar Solvation Energy of a Protein from a Single Structure: Gaussian-Based Smooth Dielectric Function for Macromolecular Modeling

**Arghya Chakravorty[1], Zhe Jia[1], Lin Li[1], Shan Zhao[2] and Emil Alexov[1*]**

**[1]Computational Biophysics and Bioinformatics, Department of Physics and Astronomy, Clemson University, Clemson, South Carolina 29634, USA.**

**[2]Departement of Mathematics, College of Arts and Sciences, University of Alabama, Tuscaloosa, Alabama 35487, USA.**

## Table of Contents

## PDB IDs used in the work:

Proteins with the following 74 PDB IDs were used for this work.

1AHO 1C75 1CBN 1G6X 1IQZ 1IUA 1J0P 1L9L 1M1Q 1MC2 1NWZ 1OK0 1TG0 1TQG 1VB0 1VBW 1W0N 1X6X 1X8Q 1XMK 1ZUU 1ZZK 2FDN 2FMA 2FWH 2H5C 2IDQ 2NLS 2O9S 2PNE 2XOD 2XOM 3AGN 3E4G 3FSA 3GOE 3IP0 3KFF 3LL2 3LZT 3O5Q 3PUC 3UI4 3V1A 3VOR 3WCQ 3WDN 3WGE 3X2L 3X32 3ZR8 3ZSJ 3ZZP 4A02 4ACJ 4AQO 4EIC 4G78 4GA2 4HGU 4HS1 4MZC 4NPD 4O6U 4O8H 4TKB 4WEE 4XDX 5CMT 5HB7 5IG6 5JUG 5L87 5TIF

## Simulation parameters for GROMACS (v5.0.5):

### Parameters for Energy minimization:

Energy minimization (EM) in explicit water, GBIS or in vacuo were performed using 10000, 5000 and 5000 steepest descent (SD) steps, respectively. Minimization was terminated when the maximum force went below 100 KJ/mol/nm. A cut-off of 1.2 nm was used for the non-bonded forces for minimization in explicit water but they were revoked for GBIS and in vacuo minimizations. For the explicit water systems, periodic boundary conditions (PBC) with particle mesh-Ewald summation (PME) were also used to account for the long-range electrostatic calculations. For the other two minimizations, none of these were invoked.

### Parameters for MD simulation:

All the MD simulations (equilibration and production phases) were carried out in explicit water environments. The same cut-offs along with PBC and PME were continued into these steps. The equilibration for all the proteins (in explicit water environments) started with constant volume-temperature (NVT) equilibration and then was followed by constant pressure-temperature (NPT) equilibration and finally production phase. For each protein, 3 independent MD simulations were carried out, which means that three different initial velocities (by different random seeds) were used. Velocity rescaling was used to maintain constant temperature (300K) and Parrinello-Rahman barostat was used to maintain constant pressure (1 atm). Harmonic restraints, with force constants of 1000 KJ/mol/nm$^2$, were imposed on the protein heavy atoms for the NVT/NPT and the first 10ns of the production phase. Only the last 10ns of the production phase was used for sampling conformations. The other ancillary values were kept at their default values suggested by GROMACS. All the motion along covalent bonds in the system were constrained using the LINCS algorithm. A description and the appropriate references for these parameters can be found in this website (http://manual.gromacs.org/online/mdp_opt.html).

## Spherical Models:

To show that the reaction field energy computed by Delphi delivers the polar component of the solvation energy, we use it on simple spherical models and compare the results with the analytical Born formula of solvation energy. A spherical cavity containing a point charge/dipole and immersed in water (implicit solvent) is modelled. We examined the trend of polar solvation energy, $\Delta G_{polar}^{solv}$, obtained from Delphi, as a function of the net charge of the point source, the radius and dielectric constant of the cavity, the total dipole moment and the placement of the dipoles inside the cavity. **Figure S1** illustrates the observed outcomes. **Figures S1a and S1b** correspond to a point charge in the cavity.

In **Figures S1a**, one can observe the unchanging $\Delta G_{polar}^{solv}$ value as a function of the internal dielectric constant of the cavity when its radius is fixed (5 Å). This energy, which quantifies the electrostatic part of the transfer energy of the cavity system from vacuum (dielectric=1.0) to water (dielectric=80.0), however, depends on the absolute value of the net charge. We also tested the negative values of these charge values and their results coincided with that obtained from the corresponding positive values; hence the label "Absolute net charge" in **Figure S1a**. These findings are in exact agreement with the predictions of Born model, which expresses the polar solvation energy in the following functional form:

$$\Delta G_{polar}^{solv} = \frac{-q^2}{2R}\left(\frac{1}{\varepsilon_1} - \frac{1}{\varepsilon_2}\right) \qquad (S1)$$

Here, 'ε1' and 'ε2' are the dielectric constants of the media (vacuum and water here) across which a spherical solute of charge 'q' and radius 'R' is being transferred. It is crucial that one appreciates how the energy does not depend on the dielectric constant of the cavity itself.

At the same time, as one can infer from **Figure S1b**, if the charge is fixed, $\Delta G_{polar}^{solv}$ is again independent of the cavity dielectric but depends on its radius. However, the radius of the cavity and the $\Delta G_{polar}^{solv}$ are inversely proportional. This also agrees with the Born approximation in **equation S1**.

With dipoles replacing the point charges, the solvation properties change (**Figures S1c and S1d**). With the same kind of treatment, we now observe that the dipole moment of the charge configuration inside the cavity causes its internal dielectric constant to affect $\Delta G^{solv}_{polar}$ (**Figure S1c).** At the same time, the net neutrality of the dipole doesn't equate it to a neutral point source (the non-zero values of $\Delta G^{solv}_{polar}$). In fact, the placement of the dipole inside the cavity also plays a role. From **figure S1d**, we add that the radius of the cavity with a fixed dipole is not a constant of its internal dielectric value. The observations from the dipole orientated cases clearly indicate that the Born formula for solvation energy is only relevant to simple point charges and when charge configurations get complex, other factors need to be considered.

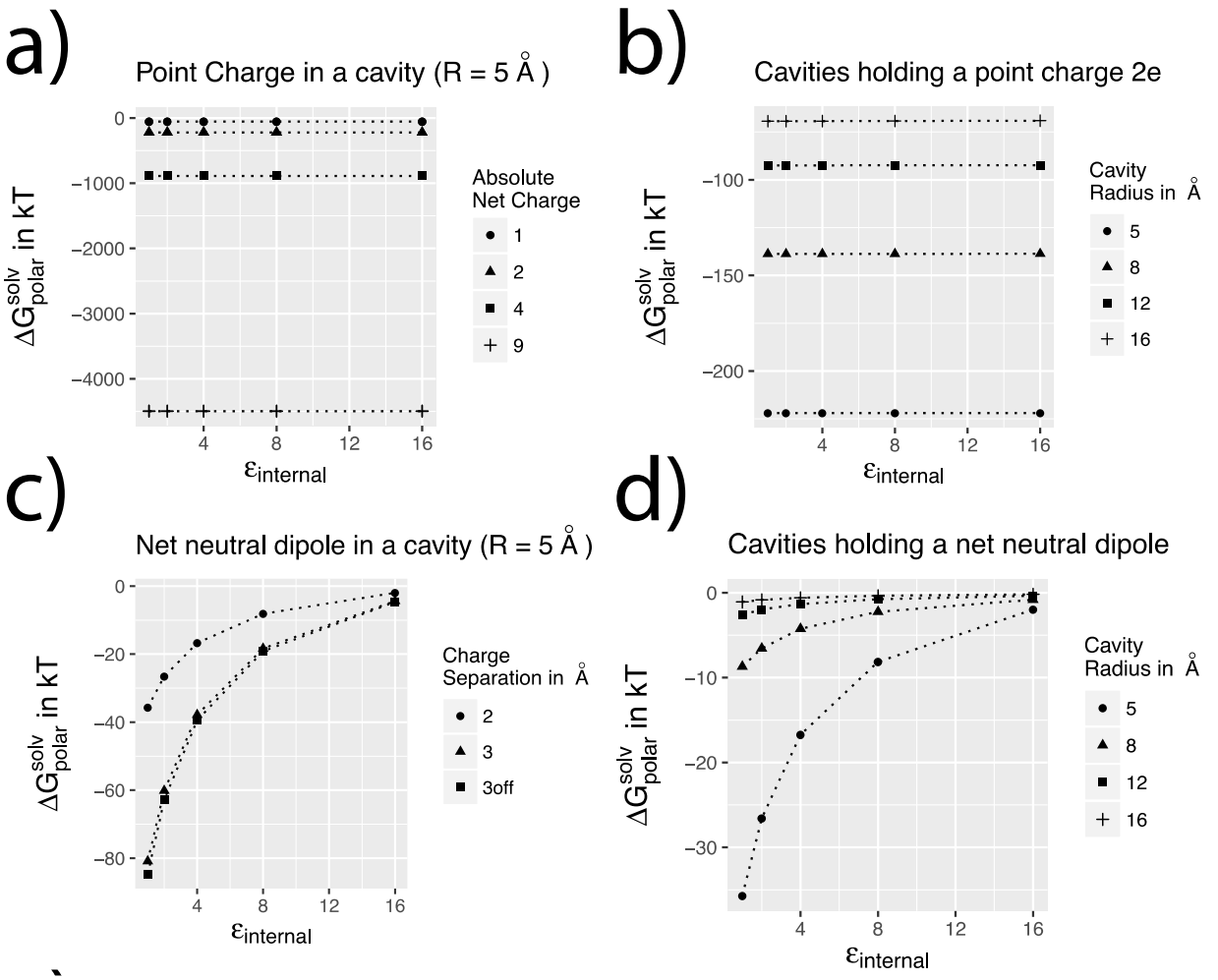


**Figure S 1**: Polar solvation energy (reaction field energy) of spherical models computed from Delphi. The top panel (S1a and S1b) show the energies for a spherical cavity containing a point charge of different magnitudes. The bottom panel (S1c and S1d) show the same for spherical cavities containing dipoles. These trends are compared with the analytical Born formula for solvation energy of solvation of simple ions.

## Schematic of the Gaussian-based smooth dielectric function with exponential decay function.

Two-fold modifications were made in the method of its implementation in DelPhi.

(i) The "surface" separating the solute phase from the external medium (medium-2) when computing the reaction field energy is drawn not based on a dielectric value but on the atomic density value ($\rho_{SF}$). This is done to fix the solute "volume" regardless of the '$\varepsilon_{ref}$' value of the internal reference dielectric constant since the '$\varepsilon_{ref}$' can influence the position of the dielectric-based surface but not that of a density-based surface (**Equation 2** in the paper).

(ii) A smoother transition from this rather discontinuous "surface" to the external region for medium-2 was incorporated using an exponential function. Setting medium-2 as vacuum ($\varepsilon_2$ = 1), the smoothing term in the following form allows the smooth exponential decay.

$$if\ \rho_{in}(\boldsymbol{r}) \geq\ \rho_{SF}\colon \varepsilon'^{(\mathbf{r})} =\ \varepsilon(\mathbf{r})$$

$$if\ \rho_{in}(\boldsymbol{r}) <\ \rho_{SF}\colon \varepsilon'(\mathbf{r}) =\ 1 + (\varepsilon(\mathbf{r}) -\ 1)e^{-(\rho_{in}(\boldsymbol{r})-\rho_{SF})(\varepsilon_{ref}-\varepsilon_{solv})}$$

Here, $\varepsilon'$(r) is the dielectric value of a 3D point when the solute is present in vacuum and $\varepsilon$(r) is the dielectric value assigned to that point when the protein's presence in solvent was modelled (refer to **equation 2** in the paper). This form ensures that far away from the surface, the dielectric value is close to 1 and near the surface, it is close to the value that corresponds to $\rho_{SF}$.

The schematic for these modifications can be visualized in **Figure S2**.

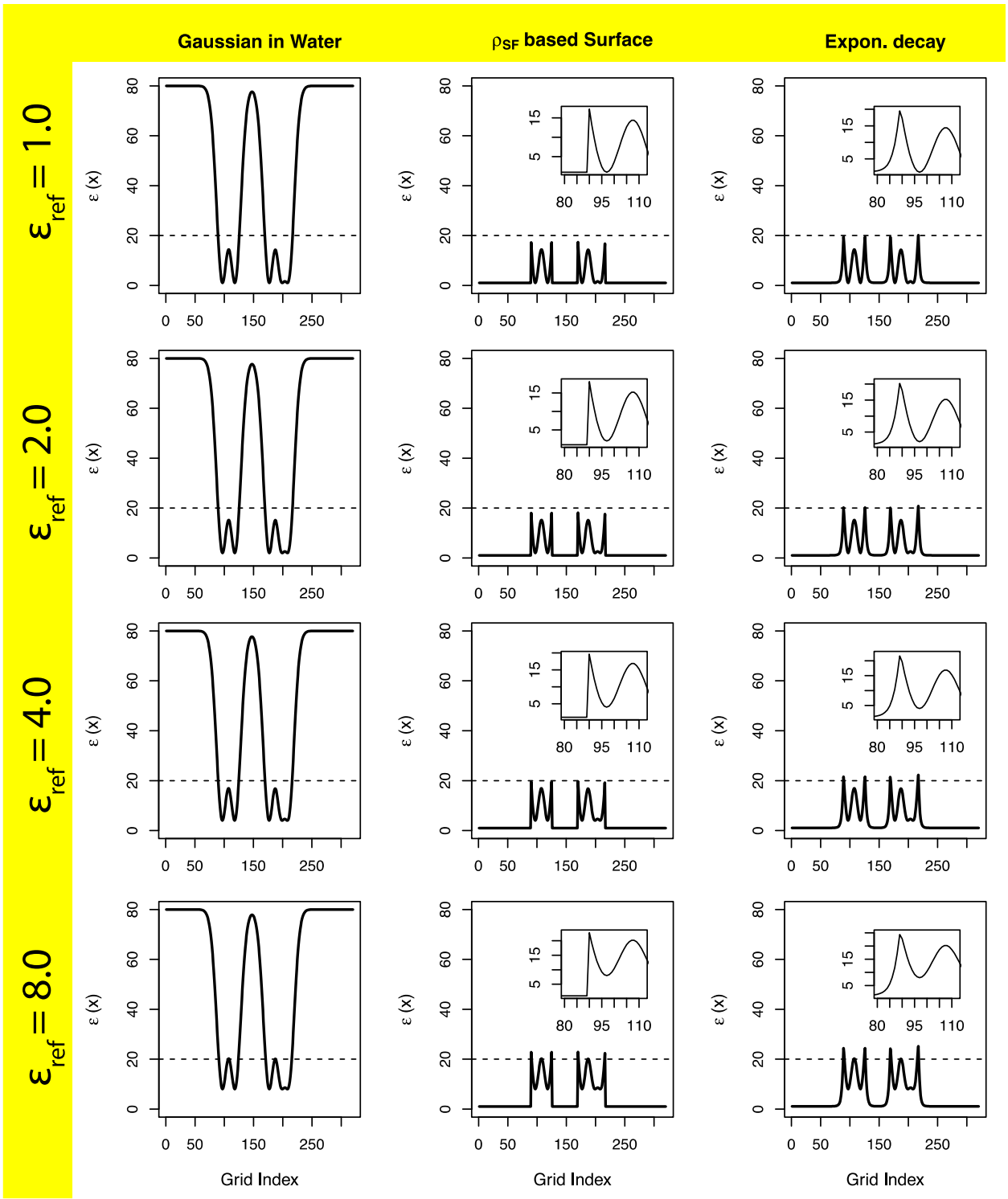


**Figure S 2**: The schematic illustrating the Gaussian-based smooth dielectric distribution of a 1D array of atoms placed arbitrarily. The schematic is shown for four different values of εref (1, 2, 4, 8). The left panel shows the distribution when the external medium is water (ε=80). The middle panel shows the distribution after demarcating a density-cutoff based "surface" that separates the solute from the medium-2 (vacuum here; ε=1) which is drawn when calculating solvation energy. The right panel shows the distribution that incorporates the exponential decay that allows a smoother transition of the dielectric from the "surface" to the external regions.

## Changing Polar solvation free energy with internal dielectric distribution

As one changes the protein dielectric ('$\varepsilon_{in}$' for the traditional 2-dielectric model or '$\varepsilon_{ref}$' for the Gaussian-based smooth dielectric model), the value of the $\Delta G_{polar}^{solv}$ changes. This change is inversely proportional. Due to the relatively simpler nature of the traditional dielectric model, it follows a $1/\varepsilon$ relationship. The analysis is done for structures minimized *in vacuo* and in solvent (GBIS/Explicit solvent).

The $\varepsilon_{in}$ values were set to 1, 2, 4 and 8 for the traditional dielectric model (TRAD) and the same values were used for $\varepsilon_{ref}$ of the Gaussian-based dielectric model (GAUSS). For the latter, sigma was equal to 0.93 and a density based "surface" was used to demarcate the protein region when computing the energies in vacuum (required in solvation energy calculation using Delphi). For more details, please see the preceding section or the METHODS in the main material. These calculations were applied to all the 74 proteins in our database and the resulting trends are illustrated in the form of boxplots. The results are shown in **Figure S3.**

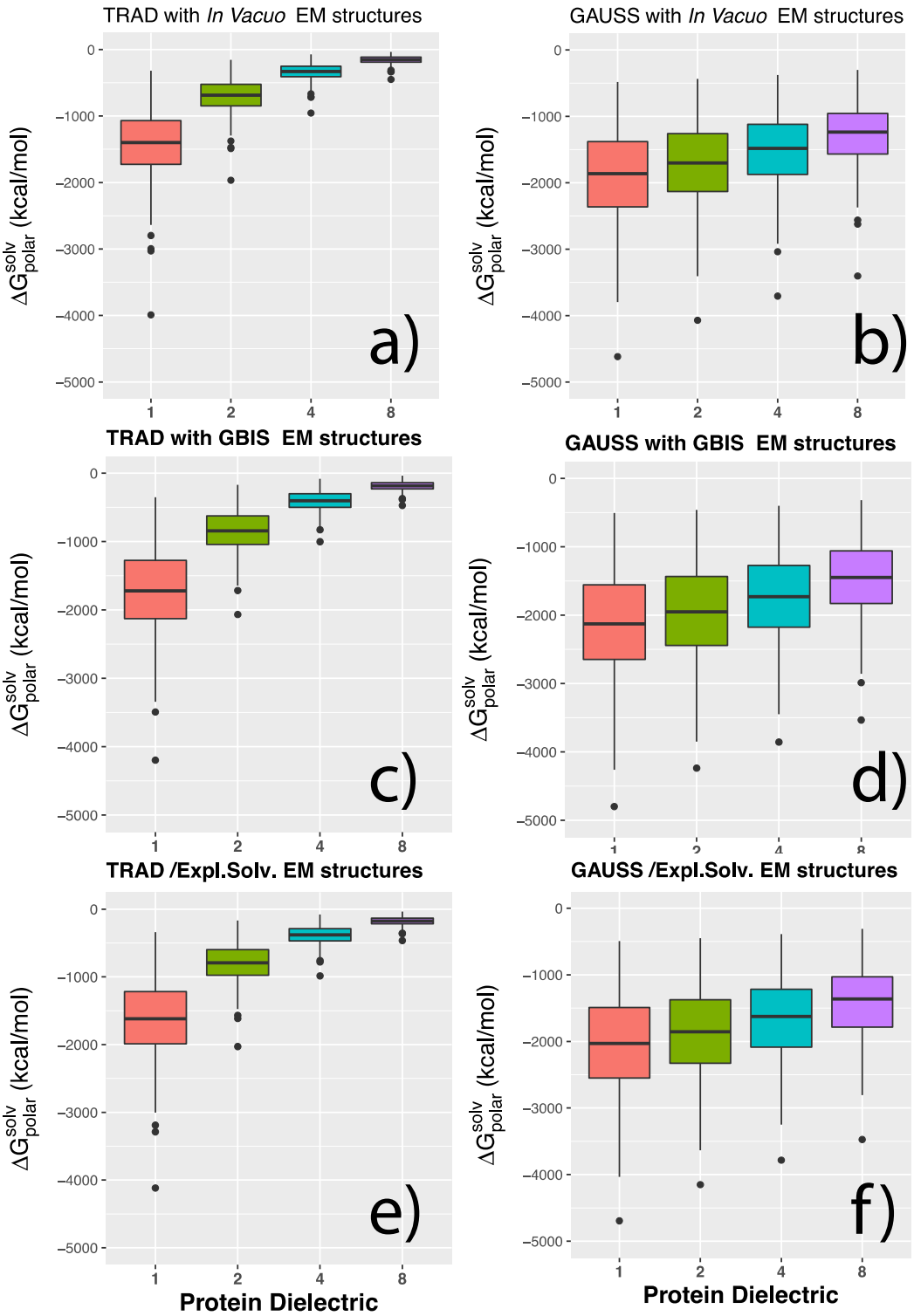


**Figure S 3:** Boxplots showing the distribution of $\Delta G^{solv}_{polar}$ from 74 proteins for each of the internal dielectric values (1,2,4,8) when applied using the traditional dielectric model (left) and the Gaussian-based dielectric model (right). The top panel (a, b) show the trend for *in vacuo* minimized structures and the middle (c, d) and bottom (e, f) show that for structures minimized in GBIS and explicit solvent, respectively.

## Quantitative comparison of the $U_{Coul}$ and $\Delta G_{Solv}$ of differently minimized structures.

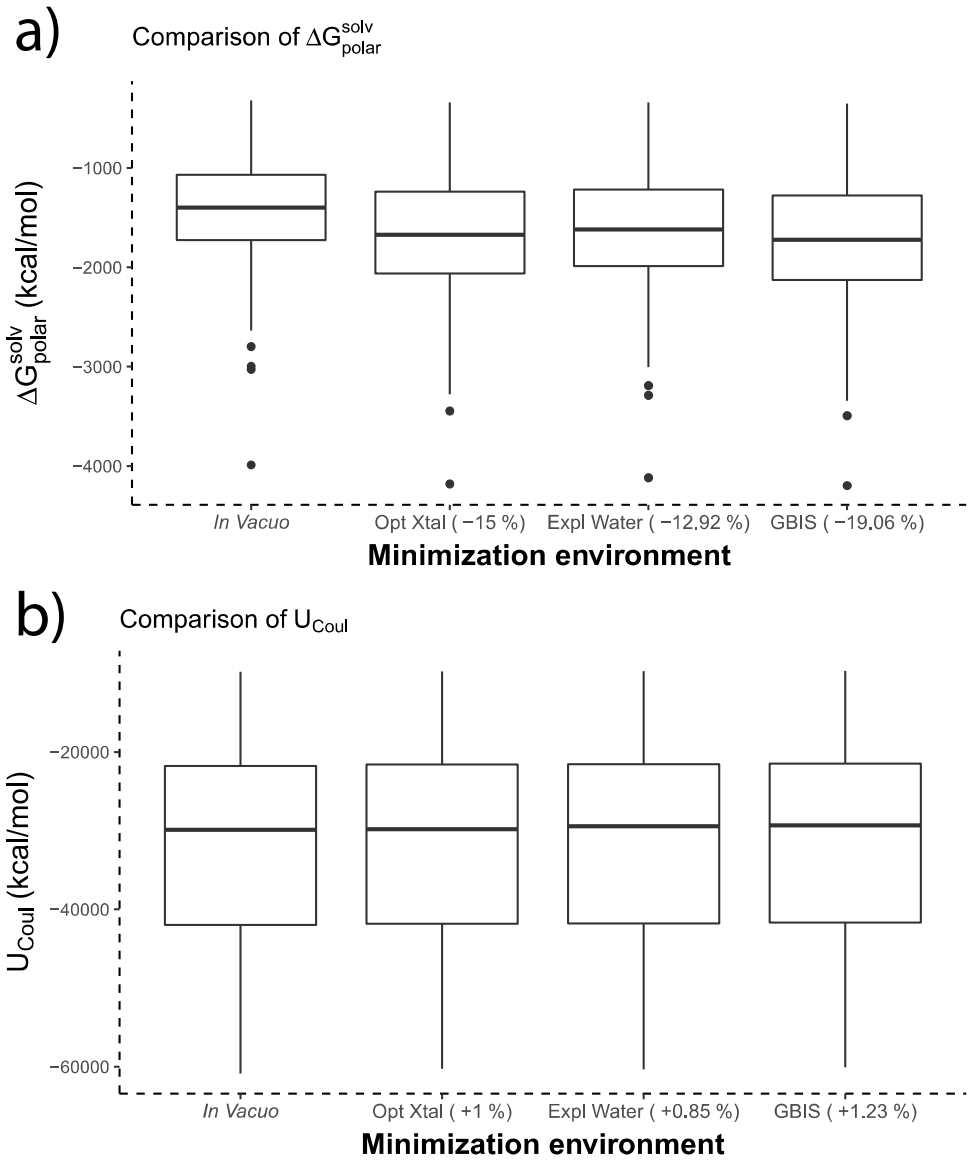


**Figure S 4:** Comparison of $\Delta G^{solv}_{polar}$ and $U_{coul}$ of the 74 proteins minimized in different environments. The percent difference of the mean of the corresponding energy components w.r.t to that of the in vacuo minimized structure are also noted in parentheses. Legend: "***In Vacuo***" – structure minimized in vacuum, "**Opt Xtal**" – optimized crystal structure, "**Expl Water**" – Explicit solvent (TIP3P) and "**GBIS**" – Generalized Born Implicit solvent.

For the optimized crystal structures and those minimized in vacuum and solvent, the polar solvation energy ($\Delta G_{polar}^{solv}$) and Columbic energy $U_{Coul}$, were compared to seek a trend for the comparison. In the main article, the comparison is shown in **Equation 3, 4**. To provide a quantitative perspective to the comparison**, Figure S4** illustrates the distribution of these energy terms for structures minimized differently and indicates the corresponding average % difference of the means w.r.t. the *in vacuo* minimized structures, i.e. for a given minimization environment X and energy component E, the % difference value is given by:

$$\%\ difference = \frac{\big(E(In\ Vacuo) - E(X)\big)}{E(In\ Vacuo)} * 100$$

**Figure S4a** corresponds to the comparison of $\Delta G_{polar}^{solv}$ (computed using Traditional dielectric model with solute dielectric = 1 and solvent dielectric = 80). **Figure S4b** corresponds to the same for $U_{Coul}$. Though the percentage difference values are not comparable, the opposite trends of $U_{Coul}$ and $\Delta G_{polar}^{solv}$ is evident. The different orders of the % difference is mainly due to the difference in the order of magnitudes of the values of these energy components; $U_{Coul}$ values are at least an order larger than the $\Delta G_{polar}^{solv}$values.

## Fluctuations of all the salt bridges identified across the 74 proteins.

To illustrate that the salt-bridges (SBs) present across the 74 proteins (484 in total), their occupancies in the MD generated ensembles was calculated. Occupancy of an SB was defined as the number of frames that SB was closed (O-N Distance < 3.2 Ang) over the total number of frames (3000). This was expressed in percentage, and therefore, a SB with 100% occupancy was always closed in the ensemble and that with 0% occupancy in the ensemble was only present in the minimized structure but not in the ensemble. Any intermediate value must be understood accordingly.

For all the SBs, their occupancies were computed using the aforementioned O-N distance criterion and the resulting distribution of the occupancy values are shown as a histogram in **Figure S5.**

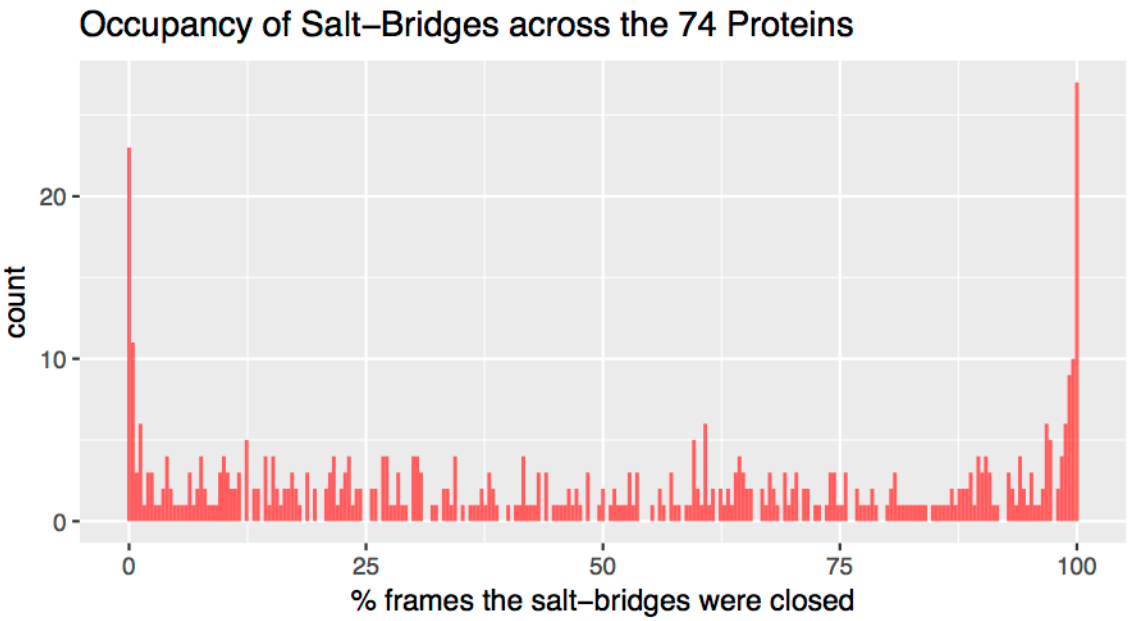


**Figure S 5:** Histogram showing the distribution of occupancies of all the salt-bridges that were identified across all the 74 proteins.

The histogram depicts that there are about 30 SBs that were always closed (occupancy ~ 100%) and around 25 of them never existed in the ensemble of their corresponding host protein. The rest of the SBs have variable occupancies ranging between 0-100%. This clearly indicates that in the MD generated ensembles, SBs were in general found to fluctuate between open and closed states (break and form respectively).

## Average Dielectric distribution using the Gaussian-based dielectric model.

The regions rich in the non-polar, polar and titratable residues across all the 74 proteins in our database was computed. The location of any residue was measured in terms of its Euclidean distance from the geometric center of the protein that contains it (host protein) normalized by the protein's radius of gyration ($R_{gyr}$). This was done to attain uniformity across the proteins which have variable sizes and geometry.

After applying the Gaussian-based model to the proteins using Delphi, the respective 'epsilon maps' were generated. The average dielectric at a radial distance from the center of a protein was calculated by identifying all the grid points that lie in a spherical shell of that radius and thickness 0.2 Ang and averaging the dielectric values on them.

**Figure S6** shows how the average dielectric constant obtained from the Gaussian-based dielectric model features at different regions of the proteins in terms of the population of the non-polar, polar and titratable residues. In addition, we also determined plotted the distribution of the salt-bridge forming titratable residues as a function of the normalized distance. All the calculations were done using the *in vacuo* minimized structures of the 74 proteins.

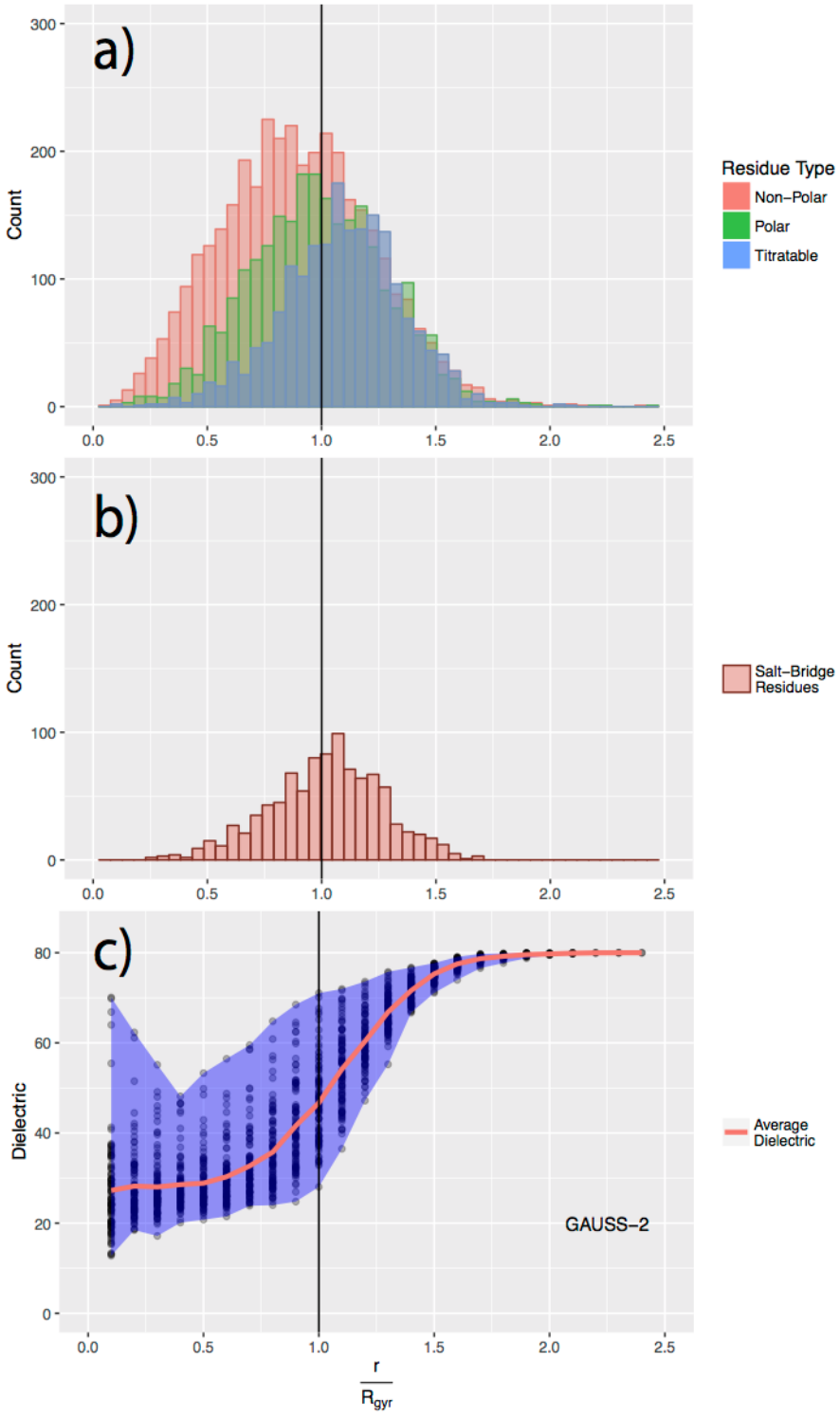


**Figure S 6:** Plots showing the average dielectric value (c) obtained from the Gaussian-based dielectric model features at different regions of the proteins in terms of the population of the non-polar, polar and titratable residues (a). In addition, the salt-bridge forming titratable residues are also considered (b).