Communications in Information and Systems
Volume 22, Number 2, 187–221, 2022

Physics-guided multiple regression analysis for calculating electrostatic free energies of proteins in different reference states

Tania Hazra and Shan Zhao*

An implicit solvent modeling problem is studied in this work, i.e., by calculating the electrostatic free energy between water and a new reference state, how to recover the original solvation free energy between water and vacuum states. Such a recovery is considered for the super-Gaussian Poisson-Boltzmann (PB) model [T. Hazra, S. Ahmed-Ullah, S. Wang, E. Alexov, and S. Zhao, *Journal of Mathematical Biology*, (2019) 79:631–672], which is a heterogeneous dielectric model to mimic the conformational changes of a macromolecule. Nevertheless, while the dielectric function should physically decrease in the vacuum state as it leaves the macromolecular region, the super-Gaussian dielectric function has an inflation over the narrow band of the solute-solvent boundary. To avoid such a non-monotonicity issue, a new reference state with a large enough dielectric value is employed in the super-Gaussian PB model. Based on the electrostatic free energy calculated using this new reference state, a multiple regression model is developed in this paper to estimate the original free energy. The proposed regression model is built physically by accounting for the contribution of each individual atom explicitly, which is modeled via the analytical result of the Kirkwood sphere. Moreover, a regression analysis is conducted for four simple physical descriptors that are related to electrostatic interactions between solute and solvent, i.e., the total number of atoms, the total charge, and the area and volume of the solvent excluded surface (SES). By using a data set of 74 proteins, the dependence of these four descriptors is analyzed. Numerical results indicate that the multiple regression model performs well in estimating the electrostatic free energies.

Keywords and phrases: Poisson-Boltzmann equation, electrostatic free energy, regression analysis, Kirkwood sphere, molecular surface, Gaussian dielectric model.

*Corresponding author. ORCID: 0000-0002-3023-2107

1. Introduction

The electrostatic analysis is indispensable for studying various important biological processes at the atomistic level, which involve charged objects such as proteins, DNAs and RNAs, immersed in an aquatic environment with mobile ions. As an implicit solvent model, the Poisson Boltzmann Equation (PBE) [15, 2, 3] has been widely used to simulate electrostatic interactions between the solute macromolecular and the surrounding solvent molecules. One common application of the PBE is to calculate the electrostatic free energy or polar solvation energy, which is defined as the polar energy released when a solute is dissolved in a solvent. Traditionally, the polar solvation energy is calculated as the difference in the electrostatic energy between two reference states, i.e., the water state and vacuum state.

This paper is concerned with an interesting modeling problem, i.e., simulating the electrostatic free energy by choosing the base reference state to be different from the vacuum. Such a problem has been mentioned in a review article [3]. The only known study in the literature is presented in Ref. [24], in which the dependence of polar solvation energy on dielectric constants of two reference states is represented via an empirical formula. Motivated by [24], a multiple regression will be conducted in this paper when the underlying PBE model becomes more complicated than that in [24].

In the classical Poisson-Boltzmann (PB) model [15, 2, 3], the PBE assumes a two-dielectric setting, i.e., a lower dielectric constant ϵ_m is assigned to the molecular region, and a higher dielectric constant ϵ_s is used for the water subdomain. The solute-solvent boundary, in this case, is a sharp interface with a dielectric jump, and its shape is commonly modeled as a molecular surface of the macromolecule. The fitting formula proposed in [24] is for the two-dielectric PBE, and is able to approximate the electrostatic free energy for different combinations of (ϵ_m, ϵ_s) from the PBE calculation on a single set of (ϵ_m, ϵ_s) values.

Many improved PB models have been developed in the literature, including diffuse interface PB models [1, 4, 5, 8, 30, 9, 28] and the heterogeneous dielectric PB models [16, 17, 6, 13, 19]. These models all feature a smooth solute-solvent boundary, i.e., the dielectric function changes smoothly from the protein to the water region over a narrow transition band. See Fig. 1 (a) for an illustration. The modeling consideration here takes into account the physical definition of dielectric coefficient and the atomistic nature of the solute-solvent system. Physically, the dielectric coefficient of water molecules or dipoles is determined by the polarizability of the dipole in responding to

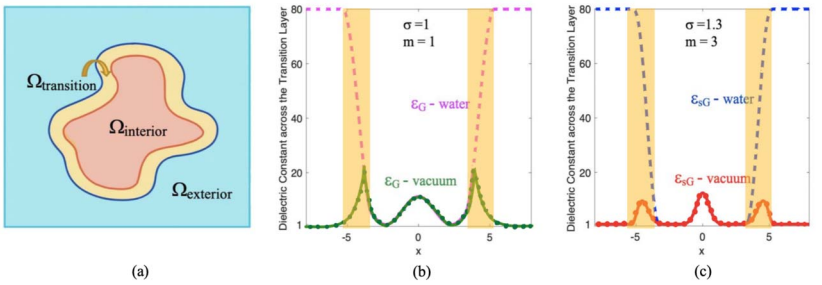

Figure 1: (a) A transition layer is assumed to define a smooth solute-solvent boundary. (b) Gaussian dielectric distributions in water and vacuum states. (c) Super-Gaussian dielectric distributions in water and vacuum states.

the electrostatic field. Such a polarizability will increases from the macromolecule interior to the water region, but should not undergo a sharp jump [7]. This is the main reason for adopting a smooth solute-solvent boundary in diffuse interface PB models [1, 4, 5, 8, 30, 9, 28], in which the water and protein regions are still treated as homogeneous dielectric media.

Besides the diffuse interface PB models, heterogeneous dielectric models, such as Gaussian PB model [16, 17] and super-Gaussian PB model [13], have been developed to construct inhomogeneous dielectric distributions to mimic the effect of the conformational changes of the macromolecule on the solvation free energy. These improved PB models perform better than the traditional two dielectric PB model in solvation free energy calculation for small molecules [16, 17]. Moreover, the Gaussian dielectric PB model provides a better prediction of the pKa's of ionizable groups against thousand experimentally measured pKa's in various proteins [26, 27]. An attractive feature of Gaussian and super-Gaussian PB models is that ensemble average electrostatic free energy could be captured by using a single structure [6, 19]. This is much more efficient than the usual ensemble calculation that involves thousands of steps of PBE computation, together with molecular dynamics or Monte Carlo simulations.

The Gaussian PB model [16, 17] and super-Gaussian PB model [13] share some similarity, but define the smooth solute-solvent boundary in different styles. In particular, when considering a protein immersed in the water or the so-called water state, the same type of density function $g(\vec{r})$ is defined in both models to describe an atom-specific heterogeneity throughout the domain. The Gaussian dielectric function $\epsilon_G(\vec{r})$ [16, 17] is simply defined as a linear combination of $g(\vec{r})$ and $1-g(\vec{r})$, with combination coefficients being

ϵ_m and ϵ_s. In the super-Gaussian model [13], a level set or surface function $S(\vec{r})$ is introduced to represent three regions: molecular subdomain, solute-solvent boundary, and water subdomain, see Fig. 1 (a). The super-Gaussian dielectric function $\epsilon_{sG}(\vec{r})$ is constructed through two linear combinations and its maximal value inside the protein is an adjustable parameter ϵ_{max}. Illustrations of the dielectric functions of two models are shown in Fig. 1. It is seen that they are close to each other in the water state.

For the vacuum state, the dielectric functions of Gaussian PB model [16, 17] and super-Gaussian PB model [13] are significantly different, see Fig. 1. In the Gaussian PB model [16, 17], the dielectric function in the vacuum state is generated by truncating that function in the water state. To this end, a surface cut boundary is first identified, say at $\epsilon = 20$. In the first version [16, 17], $\epsilon_G(\vec{r})$ is kept to be the same inside the surface cut boundary, while for outside $\epsilon_G(\vec{r}) = \epsilon_s = 1$. Consequently, $\epsilon_G(\vec{r})$ is discontinuous at the surface cut boundary. An improved version has been developed in [6], in which an exponentially decaying function is adopted outside the surface cut boundary, so that $\epsilon_G(\vec{r})$ is C^0 continuous, but not C^1 continuous, see Fig. 1 (b). From mathematical point of view, the low regularity of $\epsilon_G(\vec{r})$ will introduce additional difficulties in numerical solution and analysis of the partial differential equation (PDE). On the contrary, the dielectric function of the super-Gaussian model is at least C^2 continuous in both water and vacuum states [13]. This is achieved by defining both states via one dielectric model and the switching of two states is realized through one parameter.

We will focus only on the super-Gaussian PB model in this work. For the vacuum state, it can be seen from Fig. 1 (c) that $\epsilon_{sG}(\vec{r})$ has an inflation or a concavity change across the solute-solvent boundary. The height of this inflation is controlled by the value assigned to ϵ_{max} [13]. In some sense, such an inflation contradicts our physical intuition, i.e., the dielectric function in the vacuum state should decay monotonically outside the protein. We will not attempt to modify the super-Gaussian PB model in this study. Instead, we will consider a different modeling problem: compute the electrostatic free energy by choosing the base reference state to be different from the vacuum state. In particular, we can carefully choose the dielectric constant of the new reference state so that $\epsilon_{sG}(\vec{r})$ is monotonic outside the protein.

For the super-Gaussian PB model, the planned study allows us to bypass the non-monotonicity issue. We note that our modeling problem has a more general physical meaning: by solving solvation free energy between water state and a new reference state, how to recover the energy between water and vacuum states. To support this physical perspective, some reference

media which have low dielectric constants such as benzene, cyclohexane, trichloroethylene, etc. can be considered. In this paper, we propose a multiple regression model to retrieve the solvation free energy which was lost due to the dielectric inflation across the solute-solvent boundary and validate it for the super-Gaussian model. We note that the proposed multiple regression model can be applied to two-dielectric and Gaussian PB models to study the same type of problems.

The rest of the paper is structured with the following sections. Section 2 revisits the super-Gaussian model and the solvation free energy calculation. The time and space discretization techniques are included in that section. A physics-guided regression analysis will be conducted in Section 3 for retrieving the original solvation free energy, based on the energy calculations with different reference states. The proposed multiple regression model will be validated in Section 4 and Section 5. Finally, the paper is ended with a brief discussion. Two appendices will be presented. In the first one, a rigorous derivation of the electrostatic free energy formula for a Kirkwood sphere with a two-dielectric setting is offered. This study involves many different notations for dielectric functions and constants. A nomenclature of them is provided in Appendix B.

2. Implicit solvent model and energy calculation

In this section, we first briefly review the super-Gaussian Poisson Boltzmann (PB) model [13]. We then present how the polar solvation free energy is calculated.

2.1. Super-Gaussian Poisson-Boltzmann model

Consider a solute macromolecule such as a protein immersed in an aqueous solvent. Define a cubic domain $\Omega \subset \mathbb{R}^3$, comparatively larger than the macromolecule itself. In the super-Gaussian PB model [13], the domain Ω consists of three regions [Fig. 2(a)]: an interior domain Ω_i for solute, an exterior domain Ω_e for solvent, and a transition layer Ω_t in between Ω_i and Ω_e as a smooth solute-solvent boundary. Denote the boundary of Ω as $\partial\Omega$. The interface between Ω_i and Ω_t as Γ_i and the other interface Γ_e lies between Ω_t and Ω_e. The subdivisions of Ω can be characterized by a surface function $S(\vec{r})$ for $\vec{r} \in \Omega$ which equals to one and zero, respectively, in Ω_i and Ω_e. As $\vec{r}$ travels from interior to exterior zone, $S(\vec{r})$ decreases monotonically from one to zero [Fig. 2(b)], so that $S(\vec{r})$ is at least a C^2 continuous function over

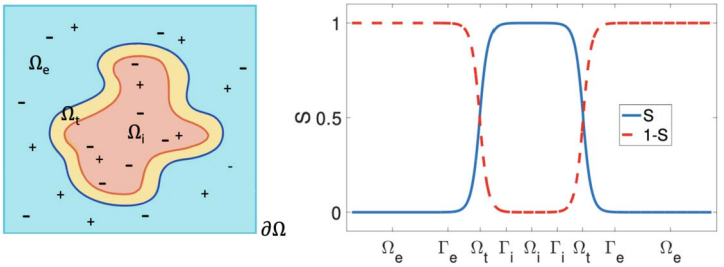

Figure 2: (a) The subdomain setting used in the super-Gaussian PB model. (b) The subdomains are characterized by a surface function S, which is plotted along a straight line.

the entire domain Ω. Following the original work [13], the surface function $S(\vec{r})$ is generated by the minimal molecular surface (MMS) model [4, 5, 23].

The electrostatic interaction of this solute-solvent system is governed by the nonlinear Poisson Boltzmann (PB) equation [15, 2, 3]

$$-\nabla\cdot(\epsilon_{sG}(\vec{r})\nabla u(\vec{r}))+(1-S(\vec{r}))\kappa^2\sinh(u(\vec{r}))=S(\vec{r})\rho_m(\vec{r}), \tag{1}$$

where $u(\vec{r})$ is the electrostatic potential and $\epsilon_{sG}(\vec{r})$ is the super-Gaussian dielectric distribution. The singular source term for N_m atoms is

$$\rho_m(\vec{r})=4\pi\frac{{e_c}^2}{k_BT}\sum_{j=1}^{N_m}q_j\delta(\vec{r}-\vec{r}_j). \tag{2}$$

On the outer boundary $\partial\Omega$, a modified Debye-Hückel boundary condition can be assumed

$$u(\vec{r})=\frac{{e_c}^2}{k_BT}\sum_{j=1}^{N_m}\left(\frac{q_j}{\epsilon_s r_j}\right)\frac{\exp\left[\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}}(a_j-r_j)\right]}{1+\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}}a_j}. \tag{3}$$

Here q_j is the partial charge, a_j is the radius and $\vec{r}_j$ is the center of the j^{th} atom. Also, r_j is defined as $|\vec{r}-\vec{r}_j|$. The other specifications of the parameters can be found in the appendix A. The singular source term is only defined within Ω_i with $S(\vec{r})=1$ there. Thus, we have $S(\vec{r})\rho_m(\vec{r})=\rho_m(\vec{r})$ in Eq. (1).

Regarding the dielectric models, Gaussian dielectric distribution [16, 17] is one of the smooth dielectric models which are designed to overcome the

discontinuity of classical two-dielectric PB model. The super-Gaussian dielectric model [13] is further introduced to address the low regularity issue of the Gaussian PB model in the vacuum state. The super-Gaussian density of the j^{th} atom is given as:

$$g_j^s(\vec{r}) = \exp\left[-\Big(\frac{|\vec{r}-\vec{r}_j|^2}{\sigma^2 R_j^2}\Big)^m\right]. \tag{4}$$

If $m = 1$, then the equation (4) presents the Gaussian density function. On the other hand, if m tends to infinity, a "hard sphere" density can be obtained, i.e., $g_j = 1$ inside the Van der Waals (VDW) ball and $g_j = 0$ otherwise. The steepness-quality of the higher order Gaussian function is controlled by two parameters, namely m: power (> 1) and σ: relative variance of the super-Gaussian function (4). In actual numerical experiments, $m = 2$ or 3 is commonly used.

To describe the entire protein, the total density function is given by:

$$g_0^s(\vec{r}) = 1 - \prod_{i=1}^{N_m}[1 - g_j^s(\vec{r})]. \tag{5}$$

Note that the total density function g_0^s can describe the density for overlapped region covered by multiple atoms. For instance, the product $g_i g_j$ accounts for the density of the overlap region due to the i^{th} and j^{th} atoms. As the dielectric treatment considered the macromolecular region as a heterogeneous medium, another parameter ϵ_{gap} was introduced. Physically, a preassigned constant ϵ_{gap} represents the maximum dielectric value of the cavity fluid inside a protein [18]. An appropriate value of ϵ_{gap} depends on the real protein system and it belongs to $(\epsilon_m, \epsilon_s]$. Now considering the presence of cavities inside the proteins, dielectric distribution within a protein region is calculated as

$$\epsilon_{in}(\vec{r}) = \epsilon_m g_0^s(\vec{r}) + \epsilon_{gap}[1 - g_0^s(\vec{r})] = \epsilon_m + (\epsilon_{gap} - \epsilon_m)\prod_{i=1}^{N_m}[1 - g_j^s(\vec{r})]. \tag{6}$$

The super-Gaussian dielectric distribution involved S with ϵ_{in} and $1 - S$ with ϵ_{out}:

$$\epsilon_{sG}(\vec{r}) = S(\vec{r})\epsilon_{in}(\vec{r}) + [1 - S(\vec{r})]\epsilon_{out}, \tag{7}$$

where ϵ_{out} can be water-phase, vacuum phase or any other solvent phase.

That means ϵ_{sG} can be seen as:

$$\epsilon_{sG}(\vec{r}) = \begin{cases} \epsilon_{in}, & \vec{r} \in \Omega_i \\ \epsilon_t, & \vec{r} \in \Omega_t \\ \epsilon_{out}, & \vec{r} \in \Omega_e. \end{cases} \quad \text{and} \quad \epsilon_t \in [\epsilon_m, \epsilon_{out}] \tag{8}$$

Here ϵ_t is controlled by the super-Gaussian density function (4), ϵ_{gap} and the surface function S. Since $S(\vec{r})$ changes from one to zero across Ω_t smoothly, $\epsilon_{sG}(\vec{r})$ changes from ϵ_m to ϵ_{out}. Therefore, $\epsilon_{sG}(\vec{r})$ is at least C^2 continuous on the entire domain Ω.

2.2. Numerical discretization of the Poisson-Boltzmann equation

Following [13], a pseudo-time approach is employed to solve the nonlinear PB equation (1). To this end, Eq. (1) is converted to a time dependent partial differential equation (PDE) by adding a pseudo-time derivative

$$\frac{\partial u}{\partial t} = \nabla \cdot (\epsilon_{sG} \nabla u) - (1 - S)\bar{\kappa}^2 \sinh(u) + \rho_m, \quad \text{in } \Omega, \tag{9}$$

with the same boundary condition (3). By using a trivial initial value $u = 0$, one numerically integrates (9) for a sufficiently long time period to steady state. The solution to the original nonlinear PB equation (1) is essentially recovered by the steady state solution of the pseudo-time dependent process (9).

One advantage of the pseudo-time approach [29, 10] is the analytical treatment of the PB nonlinear term $\sinh(u)$. In particular, one can split the time stepping of (9) into two subsystems in each time step. Then, the nonlinear subsystem can be analytically integrated so that the nonlinear instability is bypassed. For the linear subsystem, one actually solves a three-dimensional (3D) heat equation. A finite different spatial discretization together the alternating direction implicit (ADI) time stepping is employed for solving the 3D heat equation efficiently. In particular, by using the Douglas-Rachford type ADI scheme, one reduces the 3D linear system into independent one-dimensional (1D) systems in x, y, and z directions. The Thomas algorithm is employed to solve such 1D systems with tridiagonal structures. We refer to the original work [13] for more details.

2.3. Polar solvation free energy

The energy released when a solute macromolecule is dissolved in a solvent is known as the free energy of solvation. The polar component of solvation

free energy can be calculated in the PB model by computing the difference between electrostatic energy of the macromolecule in the water and vacuum states. In particular, for super-Gaussian PB model, the polar solvation free energy can be calculated as

(10)
$$\Delta G = \frac{1}{2}k_B T \int_\Omega \sum_{j=1}^{N_m} q_j \delta(\vec{r}-\vec{r}_j)(u(\vec{r})-u_0(\vec{r}))d\vec{r} = \frac{1}{2}k_B T \sum_{j=1}^{N_m} q_j (u(\vec{r}_j)-u_0(\vec{r}_j)),$$

where $u(\vec{r})$ is the solution of the PB equation (1) and $u_0(\vec{r})$ can be solved from the Poisson equation (11):

(11)
$$-\nabla \cdot (\epsilon_{sG}(\vec{r})\nabla u_0(\vec{r})) = \rho_m(\vec{r}),$$

where $\epsilon_{sG}(\vec{r})$ is obtained by taking $\epsilon_{out} = 1$. The boundary condition becomes

(12)
$$u_0(\vec{r}) = \frac{{e_c}^2}{k_B T} \sum_{j=1}^{N_m} \left(\frac{q_j}{\epsilon_{out} r_j}\right),$$

which is obtained by setting $\bar{\kappa} = 0$ in the modified Debye-Hückel boundary condition (3).

2.4. Estimating polar solvation free energy

As mentioned earlier, the dielectric function $\epsilon_{sG}(\vec{r})$ of the super-Gaussian model is at least C^2 continuous in both water and vacuum states. However, $\epsilon_{sG}(\vec{r})$ is not monotonic across the solute-solvent transition layer in the vacuum state. See for example Fig. 1(c). Physically, the dielectric function should decrease monotonically in the vacuum state whenever it leaves the molecular region, just like the Gaussian dielectric model ϵ_G [Fig. 1(b)]. To avoid such a non-monotonicity issue, we propose to calculate the electrostatic free energy by choosing the base reference state to be different from the vacuum. Note that the height of the inflation of $\epsilon_{sG}(\vec{r})$ in the solute-solvent boundary depends on the parameters ϵ_{gap} and ϵ_{out}. For each ϵ_{gap}, it is always possible to choose $\epsilon_{out} = \epsilon_{gap}$ such that $\epsilon_{sG}(\vec{r})$ does not experience concavity change outside the protein. The super-Gaussian PB model becomes "physical" in this new reference state, and the corresponding solvation free energy can be calculated by using the above mentioned formulas. With this calculated electrostatic free energy using a new reference state, this study aims to estimate the original solvation free energy. We note that a similar study has been carried out in [24]. Nevertheless, the empirical formula developed in [24]

may not be applicable to the present problem, because the super-Gaussian PB model is more complicated than the two-dielectric PB model used in [24].

3. Multiple regression model

In this section, we will develop a multiple regression model for calculating electrostatic free energy (EFE) for the super-Gaussian PB model. That means we will solve a physical problem: with EFE between water and reference states, how to recover the original energy between water and vacuum states. In other words, if we denote ΔG_{12} as the EFE of the water-vacuum case and ΔG_{13} as the EFE of the water-reference case, what is the best way to retrieve the original EFE.

3.1. Energy estimation for a Kirkwood sphere

In order to develop a physically meaningful estimation, we first investigate a simplified problem by considering a Kirkwood sphere. The electrostatic interactions are assumed to be governed by the linearized PB equation with a two-dielectric setting. In this case, one can analytically solve the solvation free energies with different states. Then, our energy estimation problem can be solved exactly.

Consider a Kirkwood sphere with radius being a and a point charge q at its center. Denote r as the distance from the center of the atom. We first introduce three different dielectric constant settings:
Setting 1: Molecule – Water

$$\epsilon_1 = \begin{cases} \epsilon_m = 1 & r < a, \\ \epsilon_W = 80 & r > a. \end{cases} \tag{13}$$

Setting 2: Molecule – Vacuum

$$\epsilon_2 = \begin{cases} \epsilon_m = 1 & r < a, \\ \epsilon_V = 1 & r > a. \end{cases} \tag{14}$$

Setting 3: Molecule – Reference Medium 1 (RM1)

$$\epsilon_3 = \begin{cases} \epsilon_m = 1 & r < a, \\ \epsilon_{RM1} = 8 & r > a. \end{cases} \tag{15}$$

In the reference medium, we choose $\epsilon_{RM1} = 8$, because ϵ_{gap} is often taken to be 8 in the Super-Gaussian simulations [13].

We aim to estimate ΔG_{12} based on ΔG_{13}, where ΔG_{12} denotes the EFE calculated based on the water (setting 1) and vacuum (setting 2) states, and ΔG_{13} is defined as the EFE calculated based on the water state (setting 1) and a reference state (setting 3). To this end, we will derive a generic formula for calculating the polar solvation free energy for the Kirkwood sphere. Similar studies have been carried out before, see for example [14, 11]. A self-contained description on the derivation of the generic energy formula (48) is presented in the Appendix A.

For one atom model, by the definition Eq. (10), the water-vacuum state EFE can be calculated as

$$\begin{aligned}\Delta G_{12} &= \frac{1}{2} K_B T \int q\delta(\vec{r})(u_W(\vec{r}) - u_V(\vec{r}))d\vec{r} \\ &= \frac{1}{2} q^2 e_c^2 \frac{1}{a} \left[\frac{1}{\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_W}} a + 1\right)\epsilon_W} - \frac{1}{\epsilon_V} \right].\end{aligned} \tag{16}$$

It is noteworthy that the EFE ΔG_{12} does not depend on the molecular region's dielectric constant ϵ_m. In the water state, we follow the setting 1 or Eq. (13), while in vacuum state, the dielectric distribution follows the setting 2 or Eq. (14). If we consider the solvation free energy for transforming the macromolecule from a low dielectric reference medium to the water, we have

$$\begin{aligned}\Delta G_{13} &= \frac{1}{2} K_B T \int q\delta(\vec{r})(u_W(\vec{r}) - u_{RM1}(\vec{r}))d\vec{r} \\ &= \frac{1}{2} q^2 e_c^2 \frac{1}{a} \left[\frac{1}{\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_W}} a + 1\right)\epsilon_W} - \frac{1}{\epsilon_{RM1}} \right].\end{aligned} \tag{17}$$

As before, for the water and reference media, we use the setting 1 and 3 respectively.

By comparing Eqs. (16) and (17), we propose to estimate ΔG_{12} based on ΔG_{13} by using the formula

$$\Delta G_{12} = \Delta G_{13} + \frac{q^2 e_c^2}{2a} \left(\frac{1}{\epsilon_{RM1}} - \frac{1}{\epsilon_V} \right), \tag{18}$$

for the Kirkwood sphere model. We note that the estimation formula (18) is exact for the one atom model with a two-dielectric setting. Moreover, all parameters involved in (18) are known.

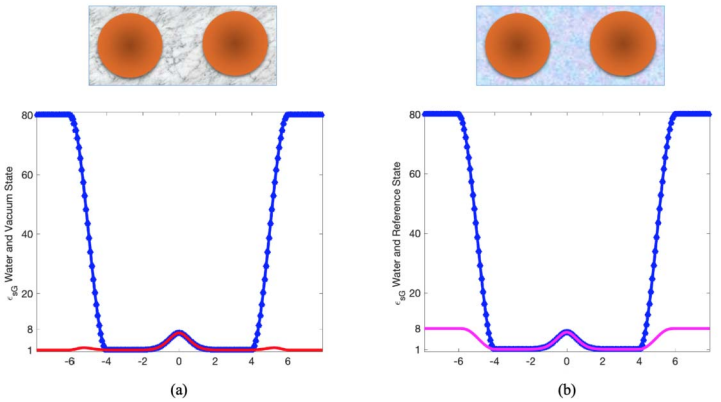

Figure 3: (a) Super-Gaussian dielectric in water (blue) and vacuum (red) state. (b) Super-Gaussian dielectric in water (blue) and reference medium (bright pink) state.

3.2. Energy estimation for the super-Gaussian PB model

We next consider the estimation of EFE for proteins in the framework of the super-Gaussian PB model. Similarly, we define ΔG_{12} as the EFE between the water and vacuum states, while ΔG_{13} is between the water state and a reference state. In particular, the dielectric functions of the water, vacuum, and reference states are all calculated by Eq. (7) with $\epsilon_{out} = 80$, 1, and ϵ_{gap}, respectively. We note that with $\epsilon_{out} = \epsilon_{gap}$, ϵ_{sG} increases monotonically in the solute-solvent boundary (see Fig. 3), so that the super-Gaussian PB model becomes "physical" in this new reference state. Thus, when we apply the super-Gaussian PB model to analyze electrostatics of real proteins, it is preferred that with the calculated ΔG_{13} value, one can directly estimate the original EFE ΔG_{12}. A regression formula will be developed for this purpose. Moreover, the accuracy of this formula will be assessed by calculated ΔG_{12} values via the super-Gaussian PB model.

Consider a macromolecule consisting of N_m atoms, having centers at $\vec{r}_j$, radii a_j and charges q_j. The regression formula to be developed is built based on the physical law for Kirkwood sphere. From the modeling point of view, there are several differences between the present system and the Kirkwood sphere, such as a heterogeneous dielectric profile Vs. a piecewise dielectric constant, and a nonlinear PB equation Vs. a linearized PB equation. Moreover, with multiple atoms or charges, the electrostatic interaction among atoms is unavoidable now. Nevertheless, our hypothesis is that the

physical law underlying the Kirkwood sphere could provide at least a low order approximation to the desired EFE. In particular, we assume the EFE between water and vacuum states can be expressed as

$$\Delta G_{12} = \frac{1}{2} q^2 e_c^2 \sum_{j=1}^{N_m} \frac{1}{a_j} \left[\frac{1}{\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_W}} a_j + 1 \right) \epsilon_W} - \frac{1}{\epsilon_V} \right] + A, \tag{19}$$

which is obviously generalized from Eq. (16). The polar solvation free energy of a single atom with other atoms being neglected represents one term in the summation of Eq. (19). Thus, the superposition of such terms is the collective energy of all atoms. The energy due to the electrostatic interactions between the atoms is expressed as a correction term, say A, which is unknown. Likewise, for the water and reference state, the EFE can be calculated using the formula:

$$\Delta G_{13} = \frac{1}{2} q^2 e_c^2 \sum_{j=1}^{N_m} \frac{1}{a_j} \left[\frac{1}{\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_W}} a_j + 1 \right) \epsilon_W} - \frac{1}{\epsilon_{RM1}} \right] + B, \tag{20}$$

where B is an unknown correction term too.

Motivated by the Kirkwood result, we assume that ΔG_{12} can be estimated by ΔG_{13} in an additive manner. For this reason, we examine the difference between two energies,

$$\begin{aligned} \Delta G_{12} - \Delta G_{13} &= \frac{1}{2} q^2 e_c^2 \sum_{j=1}^{N_m} \frac{1}{a_j} \left[\frac{1}{\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_W}} a_j + 1 \right) \epsilon_W} - \frac{1}{\epsilon_V} \right] + A \\ &\quad - \frac{1}{2} q^2 e_c^2 \sum_{j=1}^{N_m} \frac{1}{a_j} \left[\frac{1}{\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_W}} a_j + 1 \right) \epsilon_W} - \frac{1}{\epsilon_{RM1}} \right] - B \\ &= \left(\frac{1}{\epsilon_{RM1}} - \frac{1}{\epsilon_V} \right) \sum_{j=1}^{N_m} \frac{q_j^2 e_c^2}{2a_j} + \underbrace{A - B}_{\text{Unknown}} \end{aligned} \tag{21}$$

So, the equation (21) has two components: the first part with $C = \sum_{j=1}^{N_m} \frac{q_j^2 e_c^2}{2a_j}$ is known for a given protein but the second part $A - B$, the combined

correction term, is unknown. A regression analysis will be conducted in the next two sections to estimate $A - B$.

4. Regression model – a test case

Our final goal is to recover the electrostatic free energy (EFE) ΔG_{12} via the EFE ΔG_{13} based on the reference medium 1. To test the proposed regression model, we will consider a similar problem in this section, i.e., using ΔG_{13} to retrieve the EFE ΔG_{14} considering macromolecules in water and another solvent, say reference medium 2. We call that dielectric set-up as setting 4. For instance, for the Kirkwood sphere, the setting 4 can be defined as the follows.

Setting 4: Molecule – Reference Medium 2 (RM2)

$$\epsilon_4 = \begin{cases} \epsilon_m = 1 & r < a, \\ \epsilon_{RM2} = 12 & r > a. \end{cases} \tag{22}$$

A multiple regression model will be developed and validated for energy estimation. The benchmark is conducted via the super-Gaussian PB numerical solutions for real proteins. Here we will follow a similar multi-atom structure solvation free energy calculation as described in (21):

$$\Delta G_{14} - \Delta G_{13} = \Big(\frac{1}{\epsilon_{RM1}} - \frac{1}{\epsilon_{RM2}}\Big) \sum_{j=1}^{N_m} \frac{q_j^2 e_c^2}{2a_j} + \underbrace{A - B}_{\text{Unknown}} \tag{23}$$

4.1. Date set and numerical setup

To create an appropriate regression model, first we considered a data set of 74 proteins studied in [6], which gives a representative sample of proteins in the protein databank. The EFE values for both ΔG_{14} and ΔG_{13} were calculated numerically using the super-Gaussian PB Solver [13]. For the super-Gaussian density function, we considered $m = 3$ and $\sigma = 1.3$ in (4). For both ΔG_{14} and ΔG_{13}, the water state follows the setting 1 Eq. (13). Now EFE ΔG_{14} is calculated considering the macromolecule being immersed first into water Eq. (13), and then in the second reference medium Eq. (22). On the other hand, ΔG_{13} is immersed into the first reference medium, setting 3 Eq. (15). The following discretization parameters are chosen in the super-Gaussian PB model [13]: spacing $\Delta x = \Delta y = \Delta z = 0.5$Å, pseudo-time increment $\Delta t = 0.01$ for solving the NPB equation and $\Delta t = 0.1$ for the solvent case, the stopping time $T_e = 10^4 \times \Delta t$ with the tolerance level 10^{-3}.

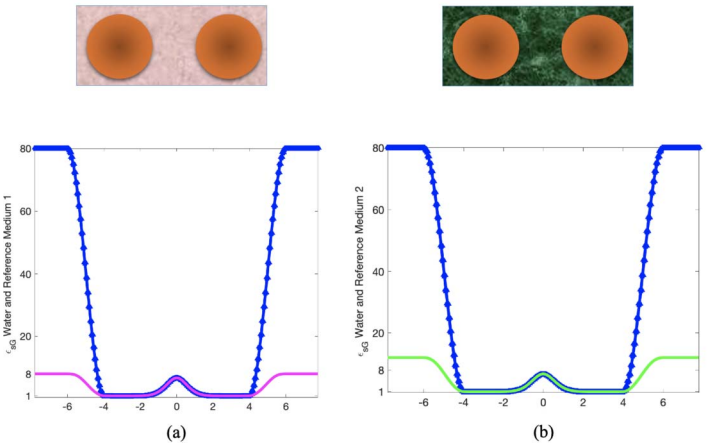

Figure 4: (a) Super-Gaussian dielectric in water (blue) and reference medium 1 (bright pink) state. (b) Super-Gaussian dielectric in water (blue) and reference medium 2 (green) state.

4.2. A multiple regression model

In this paper, we assume that the unknown part of Eq. (23), i.e., $A - B$, can be predicted by a function of several protein specific variables such as (atomic) partial charges q_j, number of atoms N_m of a protein, SES (solvent excluded surface) volume and SES area. That means the equation (23) can be written as:

$$\Delta G_{14} - \Delta G_{13} = C\Big(\frac{1}{\epsilon_{RM1}} - \frac{1}{\epsilon_{RM2}}\Big) + f(\text{Charge, Number of atoms, SES Volume, SES Area}). \tag{24}$$

For simplicity, we will denote $\tilde{Y} = \Delta G_{14} - \Delta G_{13} - C\Big(\frac{1}{\epsilon_{RM1}} - \frac{1}{\epsilon_{RM2}}\Big)$. So, Eq. (24) can be rewritten as

$$\tilde{Y} = f(\text{Charge, Number of atoms, SES Volume, SES Area}). \tag{25}$$

For the four features considered in (24), the charges are obviously related to electrostatic interactions. The number of atoms determines the size of the protein, and thus affects the solvation free energy. Moreover, the electrostatic interactions between solute and solvent will also be influenced by the volume

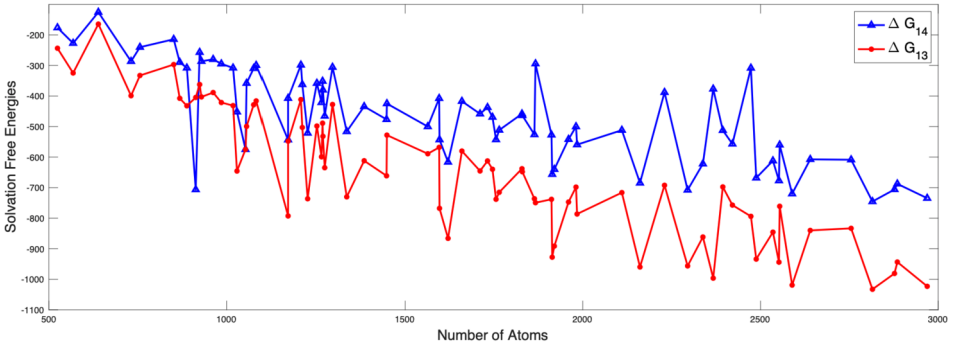

Figure 5: The line graphs of polar solvation free energies ΔG_{14} (water-RM2 states) and ΔG_{13} (water-RM1 states).

and area of SES [25]. Thus, the selection of four features in (24) is driven by physical considerations.

Using the calculated EFE values of ΔG_{14} and ΔG_{13} for 74 proteins, we next investigate the dependence of the function f on four features in Eq. (24). We first plot ΔG_{14} and ΔG_{13} against the total number of atoms N_m in Fig. 5. Since the reference media have comparatively close dielectric constants ($\epsilon_{RM1} = 8$ and $\epsilon_{RM2} = 12$), there is a consistent pattern in ΔG_{13} and ΔG_{14}, while several outliers are pretty obvious. To see the correlation in more details, we calculate $\tilde{Y} = \Delta G_{14} - \Delta G_{13} - C\Big(\frac{1}{\epsilon_{RM1}} - \frac{1}{\epsilon_{RM2}}\Big)$ for 74 proteins, and plot $\tilde{Y}$ against the total number of atoms N_m in Fig. 6 (b). A linear trend is clearly seen. As the number of atom increases, the quantity $\tilde{Y}$ becomes smaller. Thus, in our regression model, we will assume $\tilde{Y}$ depends on N_m in a linear manner.

We next consider the SES volume and area. We note that as a surface-free model, the super-Gaussian PB model [13] does not explicitly define a molecular surface separating the solute and solvent. Instead, the MSMS software developed in [22] is employed in this work to calculate the SES volume and area for 74 proteins. The MSMS package provides an efficient generation of the SES based on a reduced surface, and has been adopted in many PB models, as well as in visualization softwares, for molecular modeling. Through a similar analysis, we have concluded that $\tilde{Y}$ also depends on the SES volume and area in a linear manner. In particular, the correlation coefficient of $\tilde{Y}$ with N_m, SES area and SES volume are found to be -0.94, -0.87 and -0.67, respectively.

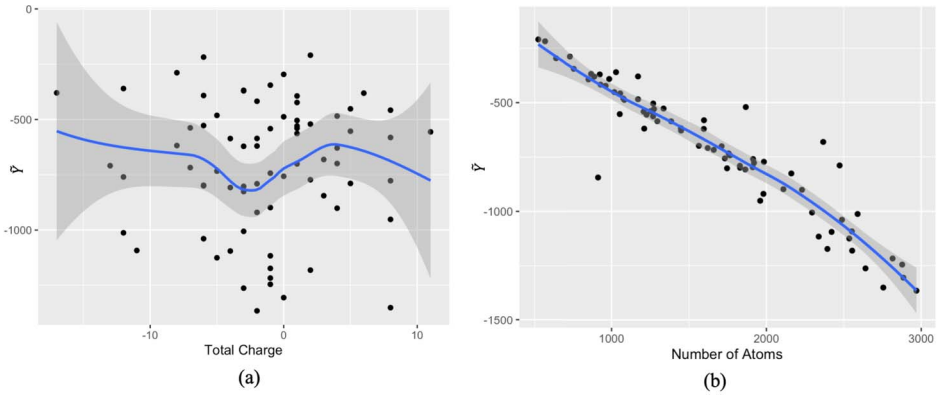

Figure 6: Trend of $\tilde{Y} = \Delta G_{14} - \Delta G_{13} - C\Big(\dfrac{1}{\epsilon_{RM1}} - \dfrac{1}{\epsilon_{RM2}}\Big)$ with respect to (a) the total partial charge Q and (b) total number of atoms in a protein.

The partial charges carried by a protein can be regarded as a four-dimensional data, i.e., for each charge, we know its location in terms of x, y, and z coordinates, and charge values (negative or positive). Neglecting all geometrical information, we consider only one feature for the charges in this paper, i.e., the total charge $Q = \sum_{j=1}^{N_m} q_j$. The plot of $\tilde{Y}$ against Q is shown in Fig. 6 (a), which obviously does not display any pattern. This basically means that the feature Q is too abstract, and cannot capture the actual charge profile of each protein. Nevertheless, a more advanced charge descriptor will inevitably becomes much more complicated in regression analysis. Thus, for simplicity, we will consider the dependence of $\tilde{Y}$ on the total charge Q. We note that in a similar study [24], the electrostatic free energy of the two-dielectric PB model is assumed to be dependent on $|Q|^{0.65}$. There was no physical justification why the power 0.65 is chosen. For the present data set with 74 proteins, we have plotted $\tilde{Y}$ with respect to $|Q|$ or $|Q|^{0.65}$, and the resulting graphs are similar to Fig. 6 (a). Since all such figures do not manifest a clear pattern, we will simply use Q, instead of $|Q|$ or $|Q|^{0.65}$, for our regression analysis.

In the plot of $\tilde{Y}$ against Q in Fig. 6 (a), a moving average trend is generated. It can be seen that the trend is roughly anti-symmetric with respect to $Q = 0$, and is experienced at least three concavity changes. Obviously, it is impossible to capture this trend with a linear function. A more realistic model is to assume $\tilde{Y}$ as a polynomial function of Q. Moreover, instead of

fixing the degree of such a polynomial, we will explore the best of fit by considering the degree up to four. With the trend shown in Fig. 6 (a), it seems unnecessary to employ an even higher degree.

In summary, we propose a multiple regression model $\tilde{P}_i$ for approximating $\tilde{Y}$:

(26)

$$\tilde{P}_i = \tilde{b}_0 + \tilde{b}_1 N_m + \sum_{k=1}^{i} \tilde{b}_{k+1} Q^k + \tilde{b}_{i+2}(SESV) + \tilde{b}_{i+3}(SESA), \quad i = 1, 2, 3, 4.$$

Here the polynomial degree for the total charge Q is denoted as i, which could be 1, 2, 3, and 4. Based on the calculated $\tilde{Y}$ values for 74 proteins [6], the least square fitting is conducted to determine the coefficients $\tilde{b}_i$. The results are presented in Table 1 for four regression models, i.e., $\tilde{P}_1$, $\tilde{P}_2$, $\tilde{P}_3$, and $\tilde{P}_4$.

Table 1: Recovery of ΔG_{14}: Least-square fitted coefficients for four regression models $\tilde{P}_i$s

$\tilde{P}_i$	Intercept	N_m	Q	Q^2	Q^3	Q^4	SES Vol	SES Area
$\tilde{P}_1$	-67.58	-0.51	-2.35	-	-	-	- 0.003	0.044
$\tilde{P}_2$	-77.79	-0.51	-0.92	0.33	-	-	-0.003	0.044
$\tilde{P}_3$	-74.75	-0.51	1.71	- 0.055	-0.035	-	-0.002	0.046
$\tilde{P}_4$	-77.84	-0.51	4.39	0.47	-0.081	-0.004	-0.003	0.046

Observing the coefficients, we can say that Q^i and N_m have more impact on the regression model than the SES generated components. The weight of the total charge is changed from model to model but N_m has a consistent contribution to $\tilde{P}_i$. Moreover, the present data analysis shows that SESA influences the prediction model more than SESV. From physical point of view, the total partial charge and the volume of a protein remain to be constants as the protein is immersed into a solvent. During the solvation process, the conformation of the protein may change, which affects the SES area. Moreover, the EFE could be determined by the induced surface charges [20, 21], which are defined on the molecular surface. Due to these factors, the SES area is more important than the SES volume in an energy prediction model.

4.3. Validation of the multiple regression model

We next validate the regression model $\tilde{P}_i$ defined by Eq. (26), with the fitting coefficients given in Table 1. To assess the estimation accuracy, we

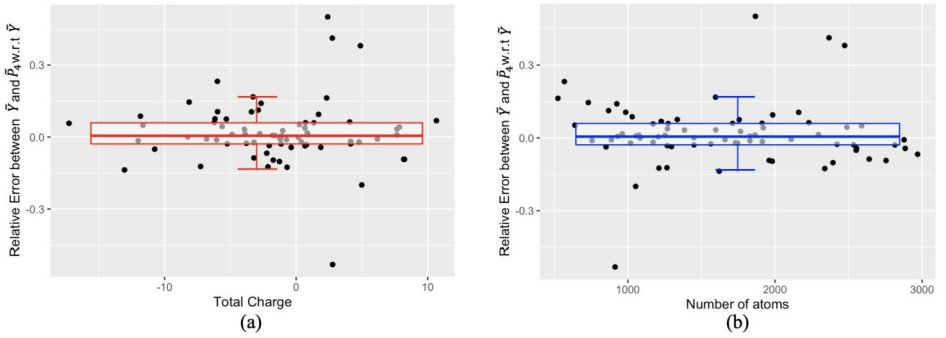

Figure 7: Recovery of ΔG_{14}: Plots of the relative error $(\tilde{P}_4 - \tilde{Y})/\tilde{Y}$ against (a) the total charge Q and (b) total number of atoms N_m.

could compare estimate $\tilde{P}_i$ value against the actual $\tilde{Y}$ value for each protein. Alternatively, we could also compute the recovered EFE

$$\Delta G_{14}^R = \Delta G_{13} + C\Big(\frac{1}{\epsilon_{RM1}} - \frac{1}{\epsilon_{RM2}}\Big) + \tilde{P}_i, \tag{27}$$

where the superscript denotes the recovered value. Then we can benchmark ΔG_{14}^R with respect to the actual ΔG_{14}.

Table 2: Model $\tilde{P}_i$'s validation result

$\tilde{P}_i$	Multiple R-Squared for $\tilde{P}_i$	Correlation Coeff between ΔG_{14}^R and ΔG_{14}
$\tilde{P}_1$	0.8942	0.803
$\tilde{P}_2$	0.8967	0.803
$\tilde{P}_3$	0.8982	0.806
$\tilde{P}_4$	0.8994	0.808

To quantify the energy estimation, the R-squared statistic is calculated in Table 2, which provides a measure of how well the model is fitting the actual data. Among the multiple regression models, the model $\tilde{P}_4$ is the best fitting model for $\tilde{Y}$. In particular, $\tilde{P}_4$ accounts for roughly 89.94% of the variance found in the response variable $(\tilde{Y})$ that can be explained by the predictor variables (Q^i, N_m, $SESV$ and $SESA$). Correspondingly, the best Pearson correlation coefficient between ΔG_{14}^R and ΔG_{14} is achieved by the model $\tilde{P}_4$. We also note that all Pearson correlation coefficients are pretty high for the present fitting test.

We will next focus on the assessment of the $\tilde{P}_4$ model. To this end, the relative error $(\tilde{P}_4 - \tilde{Y})/\tilde{Y}$ is computed for each protein, and such an

error is plotted against the total charge Q and total number of atoms N_m in Fig. 7. Two box-plots are drawn in Fig. 7 against different scatter plots as a measure of relative standing. In particular, the upper (third quartile) and lower (first quartile) edges are generated by enclosing 50% of relative errors between them, while one quarter of errors is above the upper edge and another quarter is below the lower edge of the boxes. It can be seen that half of the relative errors is within the inter-quartile range [-0.0285,0.0595]. Moreover, the median value of all relative errors is 0.005, which means that the relative errors stay mostly around zero.

For each box shown in Fig. 7, two horizontal bars are plotted at locations automatically determined by the upper and lower edges of the box. Data points beyond such two bars can be regarded as outliers statistically. It is seen that there are six outliers in the relative error boxplot: two of them (PDB id: 3ZR8 and 1VB0) are below the bottom bar and four of them (PDB id: 4HGU, 4O8H, 4GA2, and 5IG6) are above the top bar. Here the bottom bar is $Q_1 - 1.5 \times IQR$ and the top bar is $Q_3 + 1.5 \times IQR$ where Q_1 and Q_3 are the first and third quartiles, respectively. Also, IQR stands for inter-quartile range. In all cases in Fig. 7, $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR] = [-0.14, 0.17]$. The existence of outliers depends on the choice of the protein-set and accordingly, it affects the regression model. Now, the outliers are not size(N_m)-dependent but there might be some relation with the total charge. There are two turns around $Q = -6$ and $Q = 4$ in Fig. 6 and the outliers exist around those critical charges.

This test case of energy estimation (recovering ΔG_{14}) establishes that the regression model (26) is a reasonable construction with the global variables including total charge, number of atoms, solvent excluded area and volume corresponding to a particular protein. In the next section, we will construct a similar model to recover ΔG_{12}.

5. A regression model for energy estimation

To create a regression model similar to equation (26), we will maintain the same numerical structure except the EFE pairs. Now we will try to recover the EFE ΔG_{12} by using ΔG_{13}. A replica of the model (23) can be formed as:

$$\Delta G_{12} - \Delta G_{13} = C\Big(\frac{1}{\epsilon_{RM1}} - \frac{1}{\epsilon_V}\Big) + f(\text{Charge, Number of atoms, SES Volume, SES Area}). \tag{28}$$

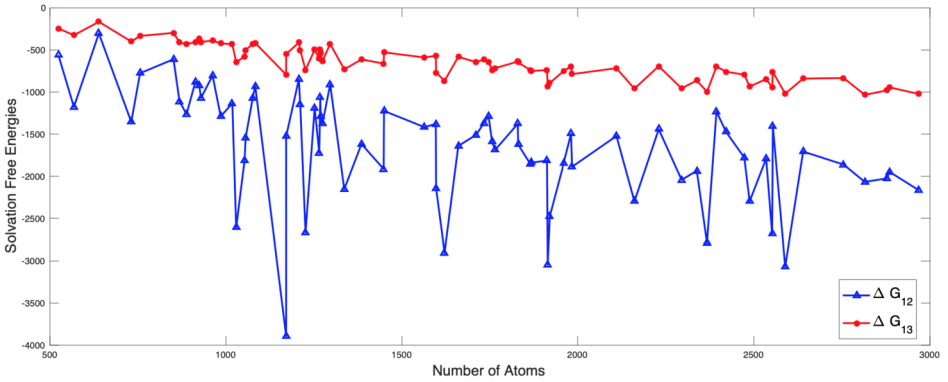

Figure 8: The line graphs of polar solvation free energies ΔG_{12} (water-vacuum states) and ΔG_{13} (water-RM1 states).

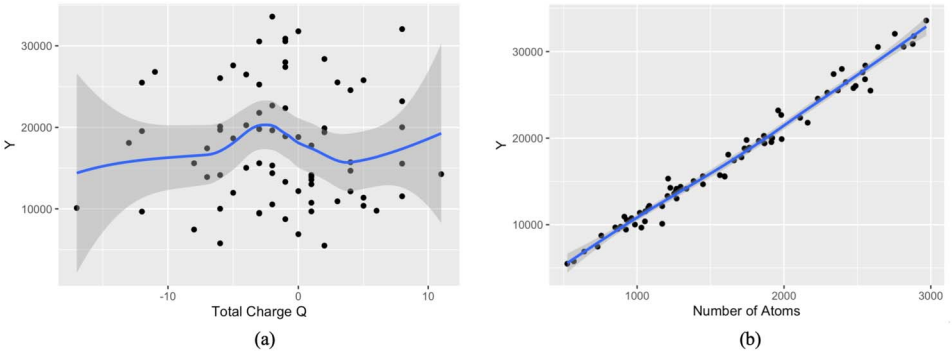

Figure 9: Trend of $Y = \Delta G_{12} - \Delta G_{13} - C\left(\frac{1}{\epsilon_{RM1}} - \frac{1}{\epsilon_V}\right)$ with respect to (a) the total partial charge Q and (b) total number of atoms in a protein.

Therefore, the regression model can be written as:

$$Y = f(\text{Charge, Number of atoms, SES Volume, SES Area}). \tag{29}$$

where $Y = \Delta G_{12} - \Delta G_{13} - C\left(\frac{1}{\epsilon_{RM1}} - \frac{1}{\epsilon_V}\right)$.

Calculating the EFE values of ΔG_{12} andΔG_{13} by using the super-Gaussian model, we first plot ΔG_{12} and ΔG_{13} in Fig. 8. The correlation between ΔG_{12} and ΔG_{13} can be seen, but the difference between them is much higher than that in the previous test case.

Next, we will calculate $Y = \Delta G_{12} - \Delta G_{13} - C\Big(\frac{1}{\epsilon_{RM1}} - \frac{1}{\epsilon_V}\Big)$ for the same set of 74 proteins. The behavior of Y against the total charge and the number of atoms presents an opposite trend. As the Fig. 8 shows $\Delta G_{12} < \Delta G_{13}$, the trend lines in Fig. 9 are flipped in comparison to those in Fig. 6. The number of atoms N_m shows a linear relation with Y whereas the total charge Q shows some nonlinear relation with Y. The correlation between Y and N_m is 0.99. In the prediction model, SES area and volume predictors are also employed. The Pearson correlation coefficient of Y with SES area and SES volume are found to be 0.92, and 0.70, respectively.

Regarding the total partial charge Q, we will follow the same choice with the degree of the Q-polynomial. Therefore, the regression models P_i following the same equation (26), consider N_m, an i^{th} degree polynomial in Q, SES area and volume. Here are the coefficients of four regression models:

Table 3: Recovery of ΔG_{12}: Least-square fitted coefficients for four regression models P_is

P_i	Intercept	N_m	Q	Q^2	Q^3	Q^4	SES Vol	SES Area
P_1	70.09	11.52	57.59	-	-	-	0.024	-0.29
P_2	251.86	11.48	32.06	-5.94	-	-	0.024	-0.29
P_3	204.91	11.55	-8.45	-1.64	0.54	-	0.020	-0.31
P_4	264.69	11.55	-60.35	-9.71	1.43	0.079	0.025	-0.32

Observing the coefficients, the intercepts have much higher value in comparison to Table 1, especially for P_2, P_3 and P_4. This happens because $|\Delta G_{12} - \Delta G_{13}| > |\Delta G_{14} - \Delta G_{13}|$. The predictors Q^i and N_m have more impact on the regression model than the SES generated components. Moreover, the present data analysis shows that SESA influences the prediction model more than SESV.

5.1. Validation of the multiple regression model

We next validate the regression models P_i defined by Eq. (26), with the fitting coefficients given in Table 3. To assess the estimation accuracy, we could compare estimate P_i value against the actual Y value for each protein. Here $Y = \Delta G_{12} - \Delta G_{13} - C\Big(\frac{1}{\epsilon_{RM1}} - \frac{1}{\epsilon_V}\Big)$. Since our final goal is to retrieve the EFE ΔG_{12}, we can compute the recovered EFE using the formula:

$$\Delta G_{12}^R = \Delta G_{13} + C\Big(\frac{1}{\epsilon_{RM1}} - \frac{1}{\epsilon_V}\Big) + P_i, \quad i = 1, 2, 3, 4 \tag{30}$$

Table 4: Model P_i's validation result

P_i	Multiple R-Squared for P_i	Correlation Coeff between ΔG_{12}^R and ΔG_{12}
P_1	0.9829	0.371
P_2	0.9842	0.509
P_3	0.9848	0.515
P_4	0.9855	0.522

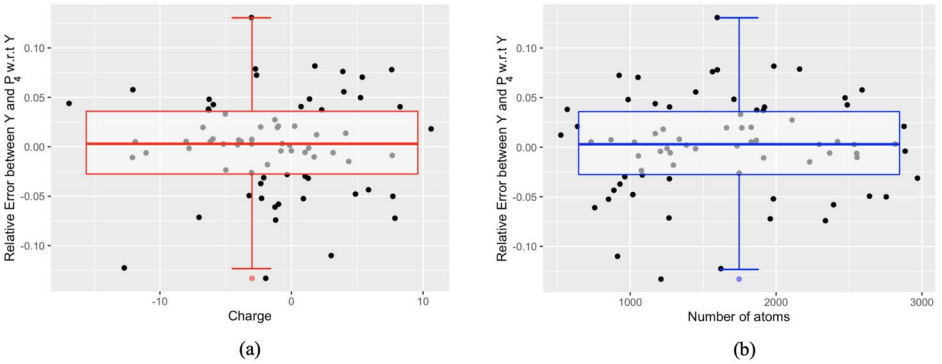

Figure 10: Recovery of ΔG_{12}: Plots of the relative error $(P_4 - Y)/Y$ against (a) the total charge Q and (b) total number of atoms N_m.

where the superscript denotes the recovered value. Then we can compare ΔG_{12}^R with respect to the actual ΔG_{12}.

Just like the test case in section 4, the R-squared statistic is calculated in Table 4. Among the multiple regression models, the model P_4 is the best fitting model for Y again. In particular, P_4 accounts for roughly 98.55% of the variance found in the response variable (Y) that can be explained by the predictor variables (Q^i, N_m, $SESV$ and $SESA$) whereas correlation coefficient between ΔG_{12}^R and ΔG_{12} is 0.522.

If we compare Table 2 and Table 4, we see that the multiple R-squared is improved for ΔG_{12}^R model but Pearson's correlation coefficient is impaired. We believe this is due to the underlying super-Gaussian PB model. In particular, the super-Gaussian model produces reasonable energy values for RM1 and RM2, but in the vacuum state, the dielectric setting of the super-Gaussian model becomes unphysical. This affects the accuracy of ΔG_{12} calculated based on the vacuum state. Here, our assumption is that if the EFE ΔG_{12} is accurate enough, the Pearson correlation in Table 4 should be as good as that in between ΔG_{14}^R and ΔG_{14}, because the same regression model is employed.

We will continue the discussion on the model P_4. So far, the relative error $(P_4 - Y)/Y$ is computed for each protein, and such an error is plotted against the total charge Q and total number of atoms N_m in Fig. 10. In every case, the relative error of the prediction model P_4 with respect to Y stays within $[-0.13, 0.13]$. Moreover, the first, second (or median) and third quartiles are $-0.13, 0.003$ and 0.13 respectively. It is noteworthy that half of relative errors stay within $[-0.0275, 0.0359]$. In fact, 74% of the proteins displays at most 5% absolute relative error. We present the EFEs ΔG_{12} and ΔG_{12}^R in Table 5 corresponding to the relative errors in between the first and third quartiles portrayed in Fig. 10.

From the boxplots Fig. 10, it is seen that there is one outlier below the bottom horizontal bar, for which the present regression model fails, i.e., protein 3ZZP. With a total charge number -2 and number of atoms 1211, the protein 3ZZP has a relative error -0.133, which is the smallest among all 74 errors. This outlier has been detected by the system, and is automatically plotted again along the center line, giving a red dot and blue dot, respectively, for chart (a) and chart (b) in Fig. 10. On the other hand, there is a data point located on the top horizontal bar. This protein, i.e., 2IDQ, has a total charge number -3 and number of atoms 1596, whose relative error 0.131 is larger than all other errors. Even though 2IDQ is not counted as an outlier by the system, this protein is the only one out of 74 proteins that produces a positive recovered EFE ΔG_{12}^R, no matter which regression models P_i, for $i = 1, 2, 3, 4$, is used. This is more worrisome than 3ZZP for the present energy estimation. Thus, 2IDQ shall be regarded as an outlier too in the present study. The exact reason for the failure of 2IDQ is unknown. But it can be observed in Fig. 10 that all other relative errors are below 0.08, showing a big gap below 2IDQ. Thus, it is believed that 2IDQ has some certain features that are different from other proteins, but cannot be captured in the present regression. A better model will be explored in the future to avoid such an "unphysical" failure.

6. Conclusion

Motivated by a particular modeling issue of the super-Gaussian Poisson-Boltzmann (PB) model [13], this work concerns with a general modeling problem, i.e., estimating the electrostatic free energy by choosing the base state to be different from the vacuum state. In particular, by taking the dielectric constant of the new reference state to be ϵ_{gap}, one can avoid the non-monotonicity issue of the super-Gaussian PB model, so that the computation of the electrostatic free energy between the water state and reference

Table 5: Comparison of ΔG_{12} and ΔG_{12}^R

PDB ID	N_m	Q	ΔG_{12}	ΔG_{12}^R
3WDN	1914	-12	-3042.075061	-3253.348719
1TG0	1029	-12	-2611.139519	-2561.957145
4O6U	2552	-11	-2686.262865	-2852.487496
3WCQ	1448	-8	-1914.252593	-1938.214199
2FDN	731	-8	-1355.185727	-1315.194623
1TQG	1660	-7	-1640.593004	-1298.653407
2FWH	1830	-6	-1623.541645	-1518.306645
3GOE	1336	-6	-2153.565618	-2040.527032
1OK0	1076	-5	-1070.743789	-1351.878017
3IP0	2535	-5	-1790.548429	-1715.909294
1W0N	1756	-5	-1592.398371	-973.8559237
3X32	1385	-4	-1617.450867	-1586.969658
3X2L	2421	-4	-1472.218888	-1322.514651
2XOD	1864	-4	-1850.607524	-1711.895275
3LL2	1746	-3	-1287.760105	-1804.126202
2XOM	2295	-3	-2043.628016	-1972.403819
1ZUU	868	-3	-1110.367707	-1039.296041
4EIC	1296	-2	-909.4579074	-1169.233369
3FSA	1829	-2	-1376.802973	-982.8432334
1IUA	1207	-1	-849.3899532	-905.7186321
1X8Q	2815	-1	-2067.754522	-1970.935339
3WGE	1765	-1	-1685.750322	-1316.353232
3E4G	2877	-1	-2026.282006	-1378.979307
4TKB	2110	-1	-1522.031326	-909.1233775
2O9S	1083	0	-928.2530694	-1269.314727
1X6X	1732	0	-1372.905593	-1345.964193
1CBN	639	0	-303.1376225	-158.96884
4AQO	1274	1	-1376.773593	-1455.430707
1ZZK	1252	1	-1188.738167	-1206.167244
4A02	2554	2	-1401.469698	-1693.076462
2NLS	524	2	-561.5991996	-494.5326891
4GA2	2367	3	-2789.299641	-2940.374902
3ZSJ	2230	4	-1441.937554	-1805.619445
4XDX	1171	4	-1524.64128	-1358.106643
1VBW	1056	8	-1540.070188	-1641.030134
1L9L	1226	11	-2670.237417	-2412.235291

state becomes truly physical. Based on energies calculated using the reference state, a multiple regression model is proposed to recover the original energies. Based on the analytical results for the Kirkwood sphere, the energy recovery is conducted in an additive manner, and the contribution of each

individual atom is accounted for explicitly. To estimate the difference between two energies, four simple physical descriptors are adopted, including the total number of atoms, the total charge, and the volume and area of the solvent excluded surface (SES). Detailed analysis is conducted to establish the function form of these four dependent variables. The resulted multiple regression model provides a satisfactory recovery for calculating electrostatic free energies of a set of 74 proteins.

Two tests are conducted to verify the proposed multiple regression model. In the first test, the energy estimation is considered by using reference medium 1 (RM1) to predict reference medium 2 (RM2), where both RM1 and RM2 are physical reference states. The Pearson correlation between actual and recovered electrostatic free energy (EFE) is about 0.8. In the second test, the EFE prediction based on RM1 is considered for the vacuum state. Nevertheless, because the dielectric setting in the vacuum state becomes unphysical, the actual energy calculated by the super-Gaussian model is quite inaccurate. Consequently, the Pearson correlation between actual and recovered EFE can only reach 0.52. Thus, such a low correlation coefficient does not mean that the regression model is ineffective. On the contrary, this indicates that the proposed model can detect inaccurate EFE calculations due to unphysical settings.

It is noted that four physical descriptors used in the present model are global features that are readily available or can be easily generated for proteins. It is expected that if local features involving structural details of proteins, such as distribution of charges, geometrical and topological information of molecular surfaces, are taken into the consideration, the prediction accuracy could be significantly improved. However, this is beyond the scope of this work. We also note that the simplicity of the present model allows it to be applied to other PB models for energy recovery. For example, in the the Gaussian dielectric PB model [6], the dielectric function is dumped to 1 by an exponential function for the vacuum state. Instead of doing that, one could set $\epsilon = 20$ outside the protein region or surface cut zone. Then, the present regression model can be applied to recover the electrostatic free energy by calculating the energy for the new reference state with $\epsilon = 20$.

The outliers of the present regression analysis have been carefully examined. In particular, there exists a protein that produces a positive or unphysical energy estimation. The exact cause of this failure will be investigated in the future to improve the regression model.

Appendix A. A Kirkwood sphere

The derivation of the electrostatic free energy formula for a Kirkwood sphere with a two-dielectric setting is provided in this appendix. To be self-contained, we first present the linearized Poisson-Boltzmann (PB) model for a general solute-solvent system. In the classical two-dielectric setting, the domain Ω is divided by a molecule surface Γ into two regions, namely the inner solute domain Ω_m and the outer solvent domain Ω_s such that $\Omega = \Omega_m \cup \Omega_s$ and $\Omega_m \cap \Omega_s = \Gamma$. Denote the boundary of Ω as $\partial\Omega$. For $\vec{r} \in \mathbb{R}^3$, the linearized PB equation in dimensionless form is given as [14]

$$-\nabla \cdot (\epsilon(\vec{r})\nabla u(\vec{r})) + \kappa^2 u(\vec{r}) = \rho_m(\vec{r}), \tag{31}$$

where u is the electrostatic potential, and the singular source term is also given by Eq. (2), in which N_m is the number of atoms, T is the temperature, k_B is the Boltzmann constant, e_c is the fundamental charge. Here q_j is the partial charge on the j^{th} atom of the macromolecule, located at its center $\vec{r}_j$, and a_j is the radius of this atom.

In the two-dielectric PB model, the dielectric function $\epsilon(\vec{r})$ is assumed to be a piecewise constant

$$\epsilon(\vec{r}) = \begin{cases} \epsilon_m, & \vec{r} \in \Omega_m \\ \epsilon_s, & \vec{r} \in \Omega_s. \end{cases} \tag{32}$$

The modified Debye-Hückel parameter κ is a piecewise constant. It vanishes in Ω_m, i.e., $\kappa = 0$, while in Ω_s, $\kappa = \bar{\kappa}$ where $\bar{\kappa}^2 = 8.486902807\mathring{A}^{-2} I_s$ and I_s is the ionic strength of the solvent. Over the dielectric interface Γ, the potential u satisfies two interface jump conditions [12]

$$[u] = 0, \quad \text{and} \quad \left[\epsilon \frac{\partial u}{\partial n}\right] = 0, \tag{33}$$

where the notation $[f] = f^+ - f^-$ represents the difference of the functional value across the interface Γ, and the directional derivative $\frac{\partial u}{\partial n}$ is along the outer normal direction of the interface Γ. On the boundary $\partial\Omega$, a numerical boundary condition (3) is usually assumed.

We next consider a one-atom system with Ω_m being a Kirkwood sphere with radius a. A partial charge q is assumed at the atom center, which is also the origin of the coordinate system. See Fig. 11. The interface Γ is simply a sphere. In order to derive an analytical solution, the entire space

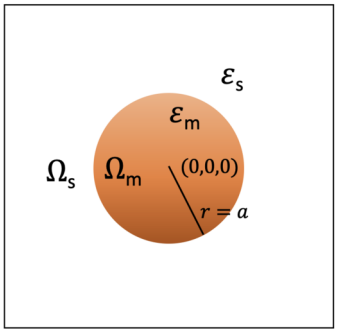

Figure 11: An illustration of the Kirkwood sphere with radius a and a charge q at the atom center $(0, 0, 0)$.

is considered with $\Omega = R^3$, and $\Omega_s = R^3 \backslash \Omega_m$. The linearized PB model for the Kirkwood sphere is given as

$$-\nabla \cdot (\epsilon(\vec{r})\nabla u(\vec{r})) + \kappa^2 u(\vec{r})) = 4\pi A q \delta(\vec{r} - \vec{0}), \tag{34}$$

$$\lim_{|\vec{r}|\to\infty} u(\vec{r}) = 0, \tag{35}$$

where $A = \dfrac{e_c^2}{k_B T}$. Denote the solution u inside and outside sphere as u^- and u^+, respectively. Across the interface Γ, the potential u satisfies the jump conditions (33). The boundary condition (35) is prescribed at the infinity.

A regularization formulation [12] will be adopted, which allows for an easier derivation of the analytical solution u. In particular, the potential u will be decomposed into two components, i.e., Coulomb component u_C and reaction-field component u_{RF}. The Coulomb component will capture the singularity of the system by satisfying a Poisson's equation

$$-\epsilon_m \Delta u_C(\vec{r}) = 4\pi A q \delta(\vec{r} - \vec{0}), \tag{36}$$

and the same boundary condition (35) at the infinity. In fact, u_C can be solved analytically, and is expressed in terms of a Green's function

$$u_C(\vec{r}) = G(\vec{r}) := \frac{Aq}{\epsilon_m |\vec{r}|} = \frac{Aq}{\epsilon_m r}, \tag{37}$$

where $r = |\vec{r}|$.

The reaction-field component u_{RF} will satisfy a source-free linearized PB equation [12]. In particular, one can recast such an equation into the spherical coordinate (r, θ, φ). Due to the rotational symmetry of the Kirkwood sphere model, all involved functions do not change in θ and φ directions. Thus, the PB equation can be reduced to an ordinary differential equation (ODE) in the radical direction r. Consequently, u_{RF} satisfies an ODE in the Ω_m

$$-\epsilon_m \frac{1}{r^2}\frac{d}{dr}\left(r^2 \frac{du_{RF}^-}{dr}\right) = 0, \quad r < a, \tag{38}$$

where we have added a superscript to u_{RF} to denote that this solution is within the sphere Ω_m. The general solution to this ODE is

$$u_{RF}^- = C_1 + \frac{C_2}{\epsilon_m r}.$$

Since u_C captures the singularity, u_{RF} should be bounded everywhere [12]. This rules out the second term, i.e., we have to take $C_2 = 0$, so that $u_{RF}^- = C_1$. By combining both Coulomb and reaction-field components, the electrostatic potential inside the sphere takes the form

$$u^- = C_1 + \frac{Aq}{\epsilon_m r}, \quad r < a, \tag{39}$$

where the constant C_1 is to be determined.

In Ω_s, the linearized PB equation (34) reduces to

$$-\epsilon_s \frac{1}{r^2}\frac{d}{dr}\left(r^2 \frac{d}{dr}u^+\right) + \bar{\kappa}^2 u^+ = 0, \quad r > a. \tag{40}$$

The general solution of this ODE is

$$u^+ = C_3 \frac{e^{-\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}}r}}{r} + C_4 \frac{e^{\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}}r}}{2\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}}r}.$$

As r goes to infinity, the boundary condition (35) requires that $u^+ = u$ approaches to zero. Thus, we have to take $C_4 = 0$. Consequently, the electrostatic potential outside the sphere takes the form

$$u^+ = C_3 \frac{e^{-\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}}r}}{r}, \quad r > a, \tag{41}$$

where the constant C_3 is to be determined.

To determine the unknown constants C_1 and C_3, one can substitute solutions (39) and (41) into the jump conditions (33) over the interface $r = a$

$$\begin{cases} u^-(a) & = u^+(a), \\ \epsilon_m \frac{\partial}{\partial r} u^-(a) & = \epsilon_s \frac{\partial}{\partial r} u^+(a), \end{cases} \tag{42}$$

where the normal direction is the radical direction. The continuity of potential gives rise to

$$C_1 + \frac{Aq}{\epsilon_m a} = C_3 \frac{e^{-\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} a}}{a}. \tag{43}$$

Since $\dfrac{du^-}{dr} = -\dfrac{Aq}{\epsilon_m r^2}$ and $\dfrac{du^+}{dr} = -C_3 \dfrac{e^{-\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} r}\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} r + 1\right)}{r^2}$, the continuity of flux yields

$$-\epsilon_m \frac{Aq}{\epsilon_m a^2} = -\epsilon_s C_3 \frac{e^{-\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} a}\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} a + 1\right)}{a^2}. \tag{44}$$

By solving equations (43) and (44) algebraically, we obtain

$$C_1 = \frac{Aq}{a\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} a + 1\right)\epsilon_s} - \frac{Aq}{a\epsilon_m}, \quad \text{and} \quad C_3 = \frac{Aqe^{\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} a}}{\epsilon_s\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} a + 1\right)}.$$

Therefore, the potential inside the Kirkwood sphere is

$$u^- = C_1 + \frac{Aq}{\epsilon_m r} = \frac{Aq}{a\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} a + 1\right)\epsilon_s} - \frac{Aq}{a\epsilon_m} + \frac{Aq}{\epsilon_m r}, \quad r < a, \tag{45}$$

while for the outside, it takes the form

$$u^+ = C_3 \frac{e^{-\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} r}}{r} = \frac{Aqe^{\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} a}}{\epsilon_s\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} a + 1\right)} \frac{e^{-\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} r}}{r} = \frac{Aqe^{\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}}(a-r)}}{\epsilon_s\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}} a + 1\right) r}, \quad r > a. \tag{46}$$

The electrostatic free energy of the Kirkwood sphere can be analytically

computed based on the potential values

$$\Delta G = \frac{1}{2} k_B T \int_{R^3} q\delta(\vec{r}-\vec{0})(u^-(\vec{r})-u_C(\vec{r}))d\vec{r} = \frac{1}{2} k_B T q(u^-(\vec{0})-u_C(\vec{0})), \tag{47}$$

where u^- is given by Eq. (45) and u_C is given by (37). Therefore,

$$\begin{aligned} \Delta G &= \frac{1}{2} qA \left(\frac{q}{a\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}}a+1\right)\epsilon_s} - \frac{q}{a\epsilon_m} \right) k_B T \\ &= \frac{1}{2} q^2 e_c^2 \frac{1}{a} \left(\frac{1}{\left(\sqrt{\frac{\bar{\kappa}^2}{\epsilon_s}}a+1\right)\epsilon_s} - \frac{1}{\epsilon_m} \right), \end{aligned} \tag{48}$$

where ΔG is in the unit kcal/mol.

Appendix B. Nomenclature

- ϵ_m: Dielectric constant in the molecular part
- ϵ_s: Dielectric constant in the solvent part
- ϵ_W: Dielectric constant in water state
- ϵ_V: Dielectric constant in the vacuum state
- ϵ_{RM1}: Dielectric constant in the reference medium 1
- ϵ_{RM2}: Dielectric constant in the reference medium 2
- ϵ_{sG}: Super-Gaussian dielectric model.
- ϵ_{in}: Dielectric constant inside a macromolecule (including atomic part and cavities)
- ϵ_{out}: Dielectric constant outside the molecular region (ϵ_{in}) which can be any of these $\{\epsilon_W, \epsilon_s, \epsilon_V, \epsilon_{RM1}, \epsilon_{RM2}\}$
- ϵ_{gap}: Dielectric constant of the cavity region in a protein
- ϵ_{max}: Maximum dielectric constant of the cavity region in a protein
- ϵ_1: Two dielectric model (molecule-water)
- ϵ_2: Two dielectric model (molecule-vacuum)
- ϵ_3: Two dielectric model (molecule-reference)

Acknowledgements

Hazra's research is partially supported by the Summer Research Grant awarded by Misericordia University, Dallas, PA 18612, USA, Summer 2020.

Zhao's research is partially supported by the National Science Foundation (NSF) of USA under grants DMS-1812930 and DMS-2110914.

References

[1] A. Abrashkin, D. Andelman, H. Orland, Dipolar Poisson-Boltzmann equation: ions and dipoles close to charge interface, *Physical Review letters*, **99**, 077801, (2007).

[2] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon, Electrostatics of nanosystems: application to microtubules and the ribosome, *Proceedings of the National Academy of Sciences*, **98**, 10037–10041, (2001).

[3] N. A. Baker, Poisson-Boltzmann methods for biomolecular electrostatics. *Methods Enzymol.* **383**, 94–118 (2004).

[4] P. Bates, G. W. Wei, S. Zhao, Minimal molecular surfaces and their applications, *J. Comput. Chem.*, **29**, 380–391, (2008).

[5] P. W. Bates, Z. Chen, Y. H. Sun, G. W. Wei, S. Zhao, Geometric and potential driving formation and evolution of biomolecular surfaces, *J. Math. Biol.*, **59**, 193–231, (2009). MR2507289

[6] A. Chakravorty, Z. Jia, L. Li, S. Zhao, E. Alexov, Reproducing the Ensemble Average Polar Solvation Energy of a Protein from a Single Structure: Gaussian-Based Smooth Dielectric Function for Macromolecular Modeling, *Journal of Chemical Theory and Computation*, **14**, 1020–1032, (2018).

[7] A. Chakravorty, S. Pandey, S. Pahari, S. Zhao, E. Alexov, Capturing the effects of explicit waters in implicit electrostatics modeling: Qualitative justification of Gaussian-based dielectric models in DelPhi. *Journal of Chemical Information and Modeling*, **60**, 2229–2246, (2020).

[8] L-T. Cheng, J. Dzubiella, J. A. McCammon, B. Li, Application of the level-set method to the solvation of nonpolar molecules. *J Chem Phys*, **127**, 084503, (2007).

[9] S. Dai, B. Li, J. Liu, Convergence of phase-field free energy and boundary force for molecular solvation, *Archive for Rational Mechanics and Analysis*, **227**, 105–147,(2018). MR3740372

[10] W. Geng, S. Zhao, Fully implicit ADI schemes for solving the nonlinear Poisson-Boltzmann equation, *Molecular Based Mathematical Biology*, **1**, 109–123, (2013).

[11] W. Geng, A boundary integral Poisson-Boltzmann solvers package for solvated bimolecular simulations, *Molecular Based Mathematical Biology*, **3**, 54–69, (2015). MR3372488

[12] W. Geng, S. Zhao, A two-component Matched Interface and Boundary (MIB) regularization for charge singularity in implicit solvation, Journal of Computational Physics, **351**, 25–39, (2017). MR3713413

[13] T. Hazra, S. Ahmed Ullah, S. Wang, E. Alexov, S. Zhao, A super-Gaussian Poisson–Boltzmann model for electrostatic free energy calculation: smooth dielectric distribution for protein cavities and in both water and vacuum states. *J. Math. Biol.*, **79**, 631–672 (2019). MR3982707

[14] M. J. Holst, Multilevel Methods for the Poisson-Boltzmann equation, University of Illinois at Urbana-Champaign, 1993 (Ph.D. thesis). MR2690378

[15] B. Honig, A. Nicholls, Classical electrostatics in biology and chemistry, *Science*, **268**, 1144–1149, (1995).

[16] L. Li, C. Li, Z. Zhang, E. Alexov, On the dielectric "constant" of proteins: smooth dielectric function for macromolecular modeling and its implementation in DelPhi, *J. Chem. Theory Comput.*, **9**, 2126–2136, (2013).

[17] L. Li, C. Li, E. Alexov, On the modeling of polar component of solvation energy using smooth Gaussian-based dielectric function, *J. Theory Comput. Chem.*, **13**, 1440002, (2014).

[18] J. Ng, T. Vora, V. Krishnamurthy, S-H. Chung, Estimating the dielectric constant of the channel protein and pore, *European Biophysics Journal*, **37**, 213–222, (2008).

[19] S. K. Panday, M. H. Shashikala, A. Chakravorty, S. Zhao, E. Alexov, Reproducing ensemble averaged electrostatics with super-Gaussian-based smooth dielectric function: application to electrostatic component of binding energy of protein complexes, *Communications in Information and Systems*, **19**, 405–423, (2019).

[20] W. Rocchia, E. Alexov, B. Honig, Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. *The Journal of Physical Chemistry B*, **105**, 6507–6514, (2001).

[21] W. Rocchia, S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera, B. Honig, Rapid grid-based construction of the molecular surface and the

use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *Journal of Computational Chemistry*, **23**, 128–137, (2002).

[22] M. Sanner, A. Olson, J. Spehner, Reduced surface: an efficient way to compute molecular surfaces, *Biopolymers*, **38**, 305–320, (1996).

[23] W. Tian, S. Zhao, A fast ADI algorithm for geometric flow equations in biomolecular surface generation, *Int. J. Numer. Meth. Biomed. Engrg.*, **30**, 490–516, (2014). MR3195017

[24] H. Tjong, H. X. Zhou, The dependence of electrostatic solvation energy on dielectric constants in Poisson-Boltzmann calculations, *J. Chem. Phys.*, **125**, 206101 (2006)

[25] C. Wang, D. Greene, L. Xiao, R. Qi, R. Luo, Recent Developments and Applications of the MMPBSA Method. *Frontiers in molecular biosciences*, **4**, Article 87, (2018).

[26] L. Wang, L. Li, E. Alexov, pKa predictions for proteins, RNAs and DNAs with the Gaussian dielectric function using DelPhiPKa, *Proteins*, (2015).

[27] L. Wang, M. Zhang, E. Alexov, DelPhiPKa Web Server: Predicting pKa of proteins, RNAs and DNAs, *Bioinformatics*, (2015).

[28] S. Wang, E. Alexov, S. Zhao, On regularization of charge singularities in solving the Poisson-Boltzmann equation with a smooth solute-solvent boundary, *Mathematical Biosciences and Engineering*, **18**, 1370–1405, (2021). MR4228983

[29] S. Zhao, Operator splitting ADI schemes for pseudo-time coupled nonlinear solvation simulations, *J. Comput. Phys.*, **257**, 1000–1021, (2014). MR3129570

[30] Y. Zhao, YY. Kwan, J. Che, B. Li, J. A. McCammon, (2013). Phase-field approach to implicit solvation of biomolecules with Coulomb-field approximation. *The Journal of Chemical Physics*, **139**, (2013).

Tania Hazra
Department of Mathematics
Misericordia University
Dallas, PA 18612
USA
E-mail address: thazra@misericordia.edu

Shan Zhao
Department of Mathematics
University of Alabama
Tuscaloosa, AL 35487
USA
E-mail address: szhao@ua.edu

Received July 20, 2021