8-2019

Modeling Electrostatics and Geometrical Quantities in Molecular Biophysics Using a Gaussian-Based Model of Atoms

Arghya Chakravorty
Clemson University, arghyac@g.clemson.edu

MODELING ELECTROSTATICS AND GEOMETRICAL QUANTITIES IN MOLECULAR BIOPHYSICS USING A GAUSSIAN-BASED MODEL OF ATOMS

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Physics

---

by
Arghya Chakravorty
August 2019

Accepted by:
Dr. Emil Alexov, Committee Chair
Dr. Feng Ding
Dr. Hugo Sanabria
Dr. Bradley S. Meyer

ABSTRACT

Electrostatic and geometric factors are critical to modeling the interactions and solvation effects of biomolecules in the aqueous environments of biological cells as they respectively influence the polar and non-polar components of the associated free energies. Conventional protocols use a hard-sphere model of atoms to devise and study the underlying thermodynamics. But this traditional model tends to overlook some of the important biophysical aspects at the cost of oversimplification of the representation of the solute-solvent environments. Here an alternative and physically appealing model of atoms – a Gaussian-based model, is presented which replaces the hard-sphere model with a smooth density-based description of atoms. This dissertation explains the derivation of a physically appealing dielectric distribution from the Gaussian schematic to model the electrostatics of biomolecules using the implicit-solvent/Poisson-Boltzmann (PB) formalism. It also demonstrates the advantages of using it for computing geometric properties of a molecule such as its volume and surface area (SA) for estimating non-polar portions of the free energy. While highlighting the qualitative importance of the Gaussian-based model, it offers

conceptual proofs towards its validity through computational investigations of explicit solvent simulations. It also reports the key features of the Gaussian-based model, which impart to it the capacity of accurately capturing the crucial biophysical factors that characterize biomolecular properties, namely – the effect of intrinsic conformational flexibility and salt distribution. The non-triviality of these factors and their portrayal through the Gaussian models are meticulously discussed. A major theme of this work is the implementation of the Gaussian model of dielectric distribution and volume/SA estimation into the PB solver package called *Delphi*. These developments illustrate the manner in which the utility of *Delphi* has been expanded and its reputation as a popular tool for modeling solvation effects with appreciable time-efficacy and accuracy has been enhanced.

DEDICATION

Dedicated to

my cat (Maxwell),

my lovely sister (Annie),

the summer of 2018 in Manchester, U.K.

and

the fine group of scientists who have made me want to be like them!

ACKNOWLEDGMENTS

I express my heartfelt thanks to my supervisor, Prof. Emil Alexov, for the advisor that he has been. I acknowledge the unbiased and selfless support he has shown towards me, which has allowed me to grow in my career path. I thank him for *putting up* with my demands – the kinds of projects I wished to work on and the kinds I didn't. I am grateful for the moral, logistic and most importantly the emotional support he has consistently provided me on various occasions. I wish to thank him for all of this and more and very importantly, for letting me freely collaborate internally as well as externally.

I am thankful to the members of my doctoral committee – Dr. Feng Ding for his overall knowledgeability about the subject matter and his help in certain discussions on various occasions, Dr. Hugo Sanabria for his amicability and the time he invested in educating me and others on some crucial topics of the field and Dr. Bradley Meyer for being an awesome, experienced and often joyful mentor overall (star-formation codes, summer school and football). Their individual advices have been priceless. I also thank them for adjusting through some emergencies I had to encounter in this time.

I am also deeply grateful to my collaborators – Dr. Shan Zhao from University of Alabama for inspiring me to orient my thoughts in more "mathematical" sense, Dr. Emilio Gallicchio from Brooklyn College for his time, ideas and patience while training me and Dr. Richard Henchman from the University of Manchester for the quality time we spent discussing scientific and other non-scientific topics and for helping me get into extended topics of research and thinking like a chemist!

I am also extremely fortunate to have had the chance to hone my skill as a computer programmer, while working on *Delphi*. It would have been very difficult had I not had the support of Dr. Chuan Li, Dr. Lin Li, Dr. Zhe Jia and Dr. Lin Wang. They have walked me through the struggles of coding a big program and inspired me to learn more and more. I will also thank Dr. Shailesh Panday in this respect as well. My lab mates certainly cannot go unnoticed. I am thankful for their presence in general, for the times we spent discussing science or otherwise, criticizing each other, learning together, uniting as a team to oppose or support an idea and most importantly, venting! I am thankful to Yunhui Peng, Swagata Pahari, Mihiri Hewa and Mahesh Koirala in that regard. They have helped me through thick and thins by staying together.

I am also joyously thankful for the friends I have made through this time – Abhishek Desai, Monsur Islam, Bishwambhar Sengupta, Judhajit Roy and Vaidehi Paliya. Their stance beside me has been of the greatest emotional support. In fact, I cannot thank them enough! I am also thankful to the batch of people who I have had the chance to interact with and share good memories in the classes we took together. With that I also thank the faculty of the PandA at Clemson, especially – Dr. Catalina Marinescu, Dr. Dieter Hartmann and Dr. Antony Valentini for igniting in me the interest to pursue fundamental understandings of physics. In that spirit, I will also thank Dr. Timo Heister and Dr. Fei Xue from the Department of Mathematics for the amazing training I received in their classes.

I am most certainly grateful to the resources provided by Clemson University in the form library services and computing services, especially the Palmetto computing cluster where most of my work was done. I was also fortunate to have received help from the office staff of PandA – Amanda, Lori, Risé, Celeste and Debra, who have never hesitated no matter how trivial the job was.

I will also acknowledge the sources of my research assistantship during my time as a PhD candidate here - NSF DMS/Mathematical Biology, grant number 1812597 and NIH, R01GM093937.

Not the least, I am extremely thankful to my parents and my lovely sister, Annie, for what they have done in order to get me here. Their countless sacrifices can never be compensated for. I very certainly cannot forget the important role of Amelia Abbott and her parents, Ed and Charlotte Abbott, who have been a constant source of assurance through this time. Had it not been for Amelia and my dear cat, Maxwell, I would have been more devastated than usual!

There is no way I can cover all the elements that have shaped the invaluable experience I have gained in the past 5 years. For better or worse, I am fortunate to have come across all of them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

This chapter introduces the fundamental concepts and tools that form the basis of the work presented in this dissertation. It discusses the significance of electrostatics in molecular biology and presents a detailed description of a theoretical formalism, called the Poisson-Boltzmann (PB) model, which faithfully describes the electrostatics of simple as well as complex biomolecular systems. It also lays the essential groundwork required for understanding the *modus operandi* of ***Delphi,*** a popular tool for solving the PB equation (PBE) to study the electrostatics of biomolecular systems. This tool is the platform on which all of the novelties characterizing this work are implemented. The last sections of this chapter highlight the use and range of application of the PB model in conjunction with *Delphi* in order to emphasize on the impact of this work.

1.1 Electrostatics in Molecular Biology

Charged groups are omnipresent in living cells by virtue of the biological molecules that comprise them. Of the 20 naturally occurring Amino acids (AA), 5 of

them tend to carry a net non-zero charge in physiological conditions. All the 5 nucleotide bases that make up a nucleic acid (DNA/RNA) carry a net negative charge. Lipid molecules, which are the primary ingredient of the cell and cell organelle membranes, also feature an inhomogeneous distribution of charged groups which eventually lead to their amphiphilic properties. These general facts allude to the ubiquitous role of electrostatic forces and the essential interactions driven by them.

But the mere presence of charged groups does not suffice in explaining role that electrostatics interactions play in conducing the observable stability and integrity of living cells. It is rather amazing that the typical values of charges (~1e or $1e^{-19}$ C) and masses (1 *a.m.u. or* $1e^{-27}$ kg) carried by the atoms in these groups and their placement within Angstrom range distances do *not* lead to an instantaneous disintegration of the cells because of the high velocities they may acquire as a result of the forces they feel. Simple classical mechanics will easily show that these velocities might be orders of magnitudes higher than the velocity of light! This is a clear indication of the presence of other factors that control the effects electrostatic interactions and preserve the basis of life as we know it.

A key element to structural and functional integrity of living cells is the presence of solvent (typically water) which tune the electrostatic interactions occurring at the atomic and cellular levels. 65-90% of the cell mass is water and owing to its high polarizability, it acts as high dielectric constant medium that "screens" the electrostatic forces and the resultant shielding paves way for other forces to contribute at par with the otherwise dominant electrostatic forces. In addition, the it also provides a physical matrix for diffusional motion of "freely" moving (bio) molecules before they "sense" the presence of nearby molecules though electrostatic forces. In fact, when taken out of the water phase and placed in a different environment such as vacuum, air, alcohol etc., these macromolecules are almost always rendered dysfunctional[1]. Therefore, when studying macromolecular property, any model of should account for the presence of water and its effects on the relevant processes.

Electrostatic solvation effects are critical to a wide array of phenomena observed in molecular biology. They are quintessential to processes that scale from atomic to mesoscopic to macroscopic levels. Those pertaining to atomic levels are best explained using quantum mechanics. But most of the effects of interest to molecular biology occur at length scales that can be described using classical

mechanics. On account of their long-ranged nature, electrostatic and solvation effects guide intramolecular interactions, like the interaction of secondary structure elements and domains that build up biomolecules, and play a vital role in influencing properties such as the stability of protein structural folds[2, 3], protein assemblies[4] and bound complexes[5-8]. It is also instrumental to interaction across different media and phases that constitute biomolecular systems[9], surface charge-charge interactions[10] and molecular recognition[11] and optimal orientation for binding[6]. Electrostatic forces are also integral to pH dependent functional or stability changes of biomolecules which is evident from the fact that alterations of native water phase characteristics such as pH, salt concentration and presence of other molecules, can also cause complete unfolding and abolishment of macromolecular interactions[12-16]. At the mesoscopic level, it is critical to the shelf-life of colloidal mixtures comprised of suspended particles as it plays alongside the Brownian motion occurring therein[17, 18] and formation of gels[19]. At the macroscopic level, the significance of electrostatics in the dynamics of complex fluids and soft matter have been meticulously studied and used for modeling them, e.g. electro-osmotic effects[20].

1.2 Electrostatics through computational models: The indispensable role of solvent.

Most of our understanding of the importance of electrostatics and their machinery have come from collaborative experimental and computational studies. Computational studies, in general, have helped dissect the origins and the basis of the electrogenic properties observed in biomolecules while ingenious experimental observations have provided a benchmark for them. With improvements in both the fields, our understanding about the cellular level processes have only gotten deeper and much more refined.

Computational models of solvated biomolecular systems can be classified into many types. A conventional routine of classification is the method through which the solvent is represented. This is because the method of treatment of solvent is a key determinant of the tractability and accuracy of a model. Solvents are far more numerous than solute atoms in a typical setup which is justified by the fact that they comprise a major volume of any system. Along with the mobile ions (solvated in the pure solvent), the solvent therefore contributes with a significantly large

number of degrees of freedom compared to the solute. The contribution is made towards the energetics of solvent-solvent as well as solute-solvent interactions.

In the explicit solvent model (see **Figure** 1.2.1)**,** all of the solvent degrees of freedom are accounted for alongside the solute degrees of freedom. The solute and solvent (with ions), are both represented with atomistic level of details and each atom exists in its own right. Though descriptively accurate, the consideration of all the solvent atoms bears an expensive computational overhead and therefore limits the ability of the model to sample information that are more relevant to the topic of study. For instance, oftentimes it is the behavior of the solute that is of interest than and the depiction of the solute in a "sea" of explicitly represented solvent atoms only adds to the total cost of the simulation.

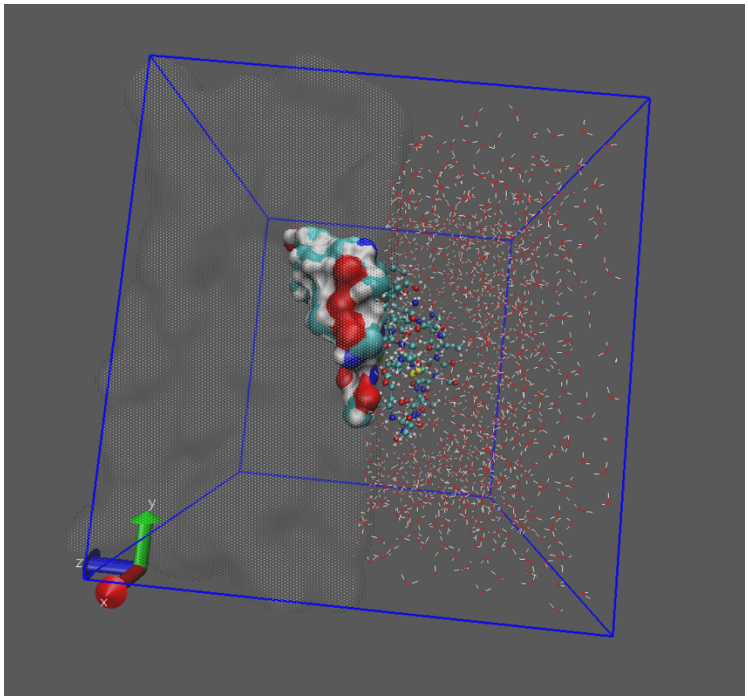


Figure *1.2.1: Explicit versus Implicit solvent models. A cartoon representation showing a solvated protein system with the left half of the box represented using a continuum approach (implicit solvent) and right half represented using explicit solvent model.*

Thermodynamically, the extra set of details can be unnecessary as most of the solvent molecules are indistinguishable and it would suffice to consider their average effect only. This forms the basis of the implicit solvent models which is a mean-field approach of representing solvated biomolecular systems. The contributions of the solvent degrees of freedom are "integrated out" and their presence is only implicitly accounted for through a mean-force and probability distribution (see Ref [21]). Typically, this is done by representing the solvent region

of the system as a medium with some dielectric value. In parallel, the solute region is represented as a cavity of its own size, with a different dielectric value. The resultant dielectric continuity of the individual phases allows for time-inexpensive computation of energies and forces in the system, in particular, the electrostatic energies and forces. Its long-ranged nature is well approximated in this framework while dramatically reducing the larger number of computational operations associated with its calculation in the explicit solvent framework. Therefore, implicit solvent models have been indispensable to continuum electrostatic models of studying solvation effects. Despite this simplicity of solvent's representation, the implicit solvent models have faithfully been able to reproduce results obtained using explicit solvent models and therefore provide a precise and time efficient alternative to the latter[22].

The widely used formalisms of continuum electrostatics are the Poisson-Boltzmann equation (PBE) model (see Review article[23]), Generalized Born (GB) model (see Review article[24]) and polarizable continuum model[25, 26]. Each of these formalisms carry their own specialties and limitations. While the polarizable continuum models are most effectively applicable for smaller compounds due to its quantum mechanics based roots, the GB model is well suited for integration with

high-throughput protocols like the classical molecular dynamics (MD) simulations[27, 28]. The PBE formalism isn't well suited for integration with high-throughput protocols and is best used with static structures. But it works utterly well with systems of any size and geometric peculiarity and is treated as a benchmark for other heuristic models, e.g. the GB model[29].

1.3 Poisson-Boltzmann (PB) formalism of continuum electrostatics

Poisson-Boltzmann formalism of continuum electrostatics model combines the Poisson equation for solving the potential distribution in space with the Boltzmann law which dictates the distribution of implicit charges in it. The Poisson equation (given in equation 1), relates the displacement vector field with the charge density of the system and can be formally derived from Maxwell's first equation (or the Gauss' law).

$$\vec{\nabla} \cdot \vec{D} = 4\pi\rho(\vec{r})$$

$$-\vec{\nabla} \cdot \left(\epsilon(\vec{r})\nabla\phi(\vec{r})\right) = 4\pi\rho(\vec{r}) \quad (1)$$

The electric displacement field vector ($\vec{D}$) is proportional to the local electric field ($\vec{E}$) via the local dielectric value (ϵ), i.e. $\vec{D} = \epsilon\,\vec{E}$, and the electric field in turn

is the negative gradient of the electrostatic potential ϕ. This potential ϕ is the result of the charges present in the system and is the unknown which is solved for.

The charges (ρ) in a solvated system arise from two sources – the fixed solute charges and the charges of the mobile ions whose effects are implicitly considered through the solvent phase. This yields the following:

$$\begin{aligned}\rho(\vec{r}) &= \rho_{fixed}(\vec{r}) + \rho_{mobile}(\vec{r}) \\ &= \sum_{i}^{N} q_i \delta(\vec{r} - \vec{r_i}) + \sum^{I} e z_i c_i^{\circ} e^{-\beta e z_i \phi(\vec{r})}\end{aligned} \tag{2}$$

The solute atomic charges (a total of N) are denoted by q_n and the solute's charge density is given by the first summation on the right-hand side. The second term expresses the charge density due to the mobile ionic species present in the system (a total of 'I' with respective valence z_i and bulk concentration c_i°). This expression originates from the canonical probability of finding an electrolyte/mobile ion at a position $\vec{r}$, on account of its potential energy ($ez_i\phi(\vec{r})$) there, which is guided by the Boltzmann law. The above equations collectively express a precursor to the differential form of the PB equation.

$$\vec{\nabla} \cdot \left(\epsilon(\vec{r})\, \vec{\nabla}\, \phi(\vec{r})\right)$$

$$= -4\pi \sum_{n}^{N} q_n \delta(\vec{r} - \overrightarrow{r_n})$$

$$- 4\pi \sum^{I} e z_i c_i^{\circ} e^{-\beta e z_i \phi(\vec{r})} \qquad (3)$$

The above expression is further resolved. Ionic species in a solvent are rendered by the salt in it and salts are ionic compounds made up of cation(s) and anion(s) bound by an ionic bond. In the solution, the screening of the electrostatic forces between the ions due to the solvent causes them to disperse. As a result, the mobile ion species, which come in charge pairs because of the salt's electro-neutrality, contribute equally to the ionic charge density term. Due to opposite polarities, the summation of that exponential term can be decomposed into two different terms which account for the positive and negative charged ionic species (I^+/I^-).

$$4\pi \sum^{I} ez_i c_i^{\circ} e^{-\beta e z_i \phi(\vec{r})}$$

$$= 4\pi \sum^{I^+} |ez_i| c_i^{\circ} e^{-\beta e z_i \phi(\vec{r})}$$

$$- 4\pi \sum^{I^-} |ez_i| c_i^{\circ} e^{+\beta e z_i \phi(\vec{r})}$$

$$= -8\pi \sum^{salt} \beta e^2 z_i^2 c_i^{\circ} \sinh(\phi(\vec{r})) \qquad (4)$$

The final outcome, with some rearrangements and introduction of new terms, is referred to as the Non-linear form of the PB equation or NLPB, which is a nonlinear

$$\vec{\nabla} \cdot \left(\epsilon(\vec{r})\, \vec{\nabla}\, \phi(\vec{r})\right) - \kappa_D^2 \sinh\left(\phi(\vec{r})\right) = -4\pi \sum_{n}^{N} q_n \delta(\vec{r} - \overrightarrow{r_n}) \qquad (5)$$

The term κ_D^2 is known as the modified Debye-Hückel parameter which disregards the solvent dielectric value that is typically used in it (see Ref[30]). κ_D is a function of the ion concentration and determines the Debye length ($l_D = \frac{1}{\kappa_D}$), a factor that indicates the rate of exponential decay of the electrostatic potential in

the solvent medium. With a larger salt concentration, Debye length is lowered meaning the screening effect is higher and *vice versa*.

In the limit of lower salt concentration, the exponential term directing the implicit presence of the mobile ions can be approximated by a linear first order term given by the Taylor series expansion ($e^x \approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$). This produces the Debye-Hückel equation, also known as the linear form of the PB equation or LPB. In this work, LPB is the default form unless otherwise is stated.

$$\vec{\nabla} \cdot \left(\epsilon(\vec{r})\, \vec{\nabla}\, \phi(\vec{r})\right) - \kappa_D^2 \phi(\vec{r}) = -4\pi \sum_n^N q_n \delta(\vec{r} - \vec{r_n}) \tag{6}$$

1.3.1 Domain of the PBE

The domain of the PBE refers to the 3D space in which it is solved. Though the solvent is theoretically depicted as an infinitely spread structure-less medium, a boundary is used to bound the entire box and "limit" the extent of the solvent region for computational convenience.

The boundary of the box also serves the purpose of providing the boundary conditions for solving this 2nd order differential equation. The boundary may contain

the scalar value of the potential on it (Dirichlet boundary condition) or its gradient, in terms of the electric fields (Neumann boundary condition). Computationally, the boundary conditions are used to initiate the iterative protocol of solving the PBE and after some defined convergence of the potential values in the domain, the final potential field is delivered. This is the basis of many of the computational packages concocted to solve PBE, including *Delphi*. More of its details are discussed in the later parts of this chapter.

1.3.2 Dielectric boundaries and solute-solvent interfaces

The volume contained by the boundary is conventionally divided into a solute/molecule region and the solvent region (with implicit mobile ions) and the distinction is made via the dielectric values assigned to these regions (see **Figure** 1.3.1). Solvent, presumably more polarizable, is assigned a higher value of dielectric, ϵ_{out} (e.g. 80 for water) and the solute is assigned a lower value, ϵ_{in} (typically in the range of 1 - 20). The strict surface separates the two media and its primary purpose is to identify what regions are accessible to the solvent. It also provides other auxiliary information about the system such as the regions accessible to implicit mobile ions (which are part of the solvent) and the region occupied by the solute

atoms which are explicitly represented as opposed to the solvent phase. The physical discontinuity introduced by the surface is routinely supplemented by some interface continuity conditions on it [31, 32]. These interface continuity conditions ensure that the potential at the interface between two media are continuous and so is the normal component of the electric displacement vector.

$$\phi(\vec{r})|_{\epsilon_1} = \phi(\vec{r})|_{\epsilon_2}, \qquad \epsilon_1 \vec{\nabla}\phi(\vec{r}) \cdot \hat{n} = \epsilon_2 \vec{\nabla}\phi(\vec{r}) \cdot \hat{n} \tag{7}$$

Though the solute-solvent interface, also known as the "*dielectric boundary*", provides a simplistic segregation of the different regions in the solvated system, it presents some conceptual and computational challenges.

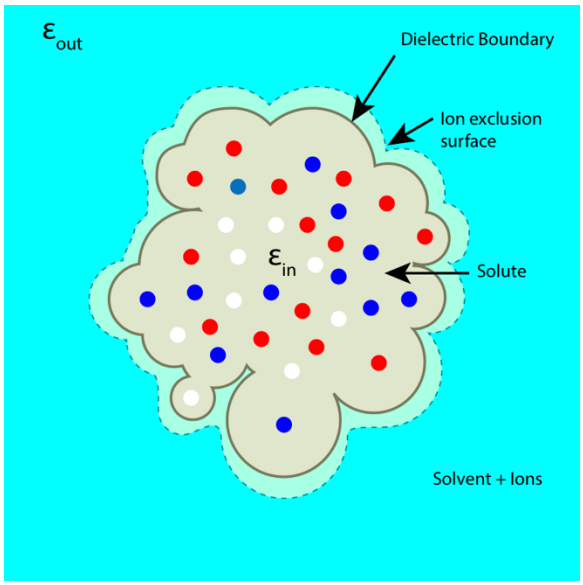


***Figure** 1.3.1: **Conventional representation of solvated system in the continuum electrostatics approach**. A cartoon illustration of the continuum electrostatic setup with implicit solvent model. The region colored in cyan represents that which is occupied by the solvent and electrolyte ions (with dielectric ϵ_{out}) and the region colored in gray represents that which is occupied the solute (with dielectric ϵ_{in}). The solid black boundary separating the two regions denotes the dielectric boundary. The dashed boundary that envelopes the solute and intrudes in to the solvent region is the ion-exclusion surface which separates the regions that are accessible and inaccessible to the electrolytes. The atoms comprising the solute retain their charges and they are colored arbitrarily to highlight the complexity and inhomogeneity of charge placement in the solute.*

A major conceptual difficulty is the lack of an exact and conclusive definition of the dielectric boundary. Dielectric boundary, whose purpose is to identify the regions with different polarizabilities and composition (e.g. solute region, ion-exclusion region and solvent region), are known to be used in various forms with varying justifications[33]. A routinely used surface is called the solvent-accessible surface (SAS) which defines a geometric surface around the solute and marks the regions strictly accessible/inaccessible to the solvent. The surface, in essence, is the locus of the center of a "solvent probe molecule" represented using a hard-sphere of some radius (1.4Å for water) as it probes the volume around the solute while constantly maintaining a tangential contact with at least one of the solute's atoms the size of their Van der Waals radius[34]. When the probe radius is set to 0, the

resultant surface is called the Van der Waals surface (VdWS). Another widely used surface definition is the molecular or the solvent-excluded surface (MS or SES; also called the Connolly surface) which is the union of the surfaces of tangential contact of the solvent probe and the solute atoms[35, 36]. Other surface definitions include van der Waals (VDW) surface, Gaussian surface, spline surface, geometric flow surface, blobby and skin surfaces[37].

On the other hand, a major computational challenge is its easy integration with high throughput molecular mechanics algorithms (see Ref. [38, 39]). Central to any molecular mechanics program is efficient and precise calculation of atomic forces. But solving PBE to garner these forces is nontrivial, especially when the shape and charge distribution of the solute is complicated (a feature prerogative of biomolecules). Popular numerical schemes like the Finite difference (FD) method[40, 41], Finite Element (FE) method[42] and Boundary element (BE) method[43] work by discretizing the space of interest and the final outcomes can be sensitive to the resolution of the discretization. The finesse or the resolution spawns a tradeoff between the accuracy of the outcome and the computation time, which if not addressed smartly, can make a protocol very inaccurate or extremely time intensive,

none of which resonate with the central idea of high-throughput molecular mechanics algorithms.

The nontriviality in solving the PBE numerically has theoretical and computational bottlenecks, which can often be coupled. Theoretically, the "discontinuity" in the dielectric distribution introduces "singularities", cases where the numerical schemes used to solve the PBE breaks down. This dielectric discontinuity has other physical issues which add to the overall nontriviality of solving PBEs. For example, the dielectric separation requires that physical forces, besides the Coulombic force, also be calculated in order to correctly describe the solute-solvent interactions and account for the effects of "dielectric stress" on the surface atoms [44-47]. Numerically, the outcome of force calculation can depend on the method adopted for solving PBE and the definition of the dielectric boundary[48, 49]. On top of that, any small change in the macromolecular conformation can alter the dielectric border between macromolecule(s) and water phase which can lead to additional instability[39, 45, 50-52].

Regardless, the convenient separation of the solute and solvent regions into strictly disjoint zones overlook very vital properties of the solvent in regions local to the interface. For instance, they overlook the physical nature of interactions between

macromolecule and water and the ability of water molecules to mediate binding based on its location around the macromolecule (e.g. Ref [53]). This also overlooks the fact that the hydrophobic surface patches or cavities are naturally not very hydrated, while the hydrophilic patches are [54-56]. Therefore, a physically sound protocol that delivers a dielectric surface should not only account for the geometry but also consider the physio-chemical properties of a macromolecular surface.

Recently, the matched interface and boundary (MIB) method was introduced [57, 58]. The method rigorously enforces the solution and flux continuity conditions at the biomolecule-solvent dielectric [59, 60]. Similarly, the Variational implicit solvent method (VISM) was proposed to account for differential hydration depending on the physicochemical and structural characteristics of the biomolecule[61].

Besides accounting for the physio-chemical properties of the macromolecule-water interface, it is equally important to consider the fact that biomolecules do not stay "frozen" in their environment. This is the typical approach of applying PBE or other implicit solvent formalisms, in that they operate with one conformation at one time. Molecular flexibility continuously updates local interactions of solvent-exposed atoms with the solvent and other solute atoms. In addition, a strict boundary between the two phases (solute-solvent phases) is not

physically sound by virtue of the constant motion and interaction of the solute-solvent atoms [62]. Inspired by these challenges, a solvation model known as a Gaussian-based smooth dielectric distribution model [63] was developed whose design is motivated by the Gaussian model of atoms[64]. This dissertation is built around this model and extensively discusses its mathematical formalism, integration with PBE and its applications.

1.4 Total Solvation Energy in continuum electrostatic models

The total energy of a solvated system can be described using a simple decomposition into various energy terms, each with their own physical underpinnings and contributions. Since energy is a state function, it has been conceptually and numerically advantageous to define the total energy as the sum of energies stemming from polar and non-polar solute-solvent interactions and gas-phase potential energy terms stemming from the intra-solute interactions. Although it's not realistic, it provides a concrete basis for the decomposition. The total energy of the system, thus is equal to, the work done to place the solute from the gas-phase (or vacuum) into

the bulk of the solvent which requires creating a cavity in the bulk the size of the solute and allowing intra-solute and solute-solvent interaction through short and long-range forces. **Figure** 1.4.1 presents a visual cue for these processes and illustrates the thermodynamic cycle that describes the various energy terms involved in the process.

The solvation energy (ΔG_{solv}) is defined as the energy of transferring the solute into the solvent which invites contributions from the polar and non-polar effects ($\Delta G_{solv} = \Delta G_{solv}^{polar} + \Delta G_{solv}^{np}$). The polar effect originates from the interaction of the solute charges with the polarizability induced in the solvent as a result. The non-polar effect originates from the rearrangement of the solvent's structure in the vicinity of the cavity created by the solute's presence and the short-range dispersion forces acting between them. Respectively, the energy terms are referred to as the polar and the non-polar parts of the total solvation energy[65].

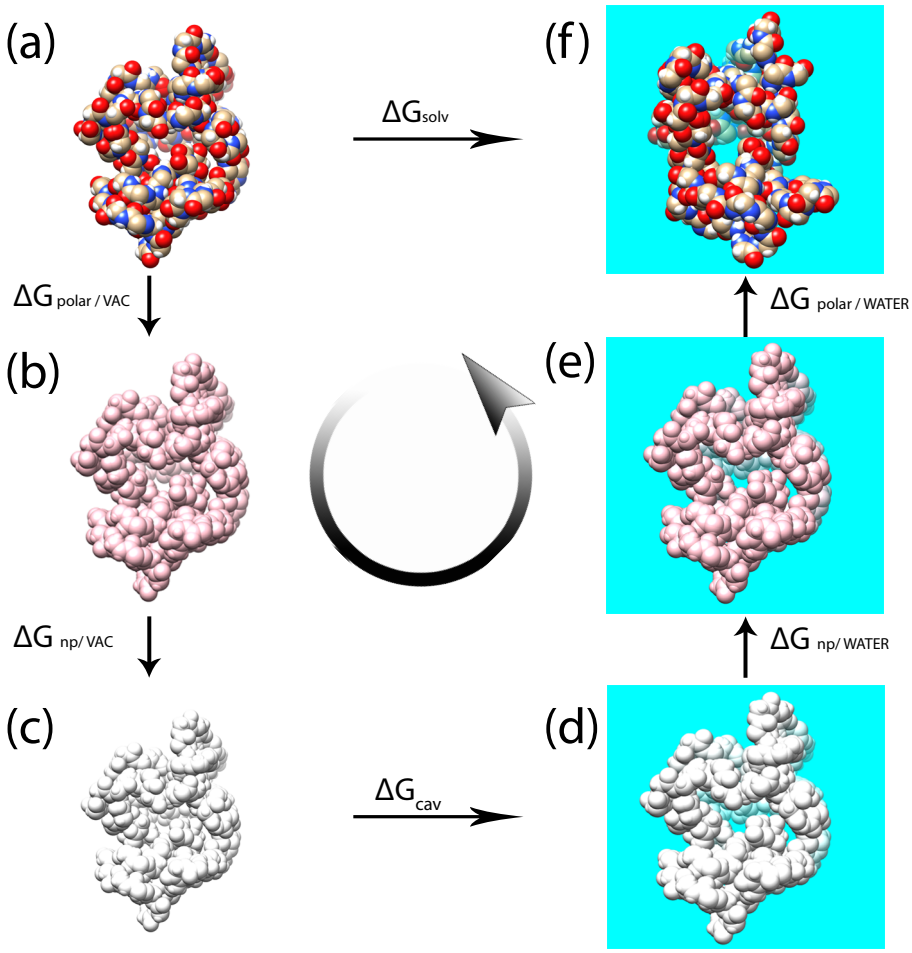


***Figure** 1.4.1: **Thermodynamic cycle of solvation**. The solvation energy is the sum of the various polar and non-polar components which originate from several transformations, which when put together, emulate the transfer of a solute from one medium to another. The cycle illustrates a series of these unrealistic transformation whose energies can be used to obtain the total solvation energy of the solute in question. The theory is based on the state-function nature of free energy. The exact transfer is shown in the top panel. But this is equivalent to following steps (a) through (f) in an anticlockwise manner.*

A different variation of this decomposition is also used; typically in the so-called *Variational implicit solvent models*[66]. The formulation invokes a strict role

of the dielectric boundary (Γ) to indicate solute-solvent separation and seek the point at which the total solvation energy is at its extremum. By deliberately discarding some of the energy terms from the functional of total solvation energy mentioned in Ref[66] (which are not extremely relevant to this introduction), the expression acquires a functional form:

$$\Delta G_{solv}[\Gamma] = \left(P \int_{\Gamma} dV + \int_{\Gamma} \gamma dS \right) + \Delta G_{solv}^{polar} \tag{8}$$

In the above expression, the terms in parentheses add up to yield the non-polar component of the solvation energy and it is clear that it depends on the solute's volume and surface area. In 5, these particular topics are addressed exclusively when the models of determining the non-polar solvation energy are discussed in the framework of *Delphi*'s finite difference setup.

In the subsections that follow, each of these terms are conceptually introduced and appropriate references for their detailed discussion, present in later chapters, are mentioned.

1.4.1 Polar Solvation energy

As mentioned above, the polar solvation energy is the work associated with allowing the solute charges to interact amongst themselves through Coulombic forces in a low dielectric medium and to interact with the field induced in the high dielectric solvent medium.

In effect, it is equal to the difference of the *total electrostatic energy* of the solute in the two media across which it is transferred. Conventionally, it is understood that the solute is transferred from the gas-phase (equivalent to vacuum) into a solvent (typically an aqueous solvent). In the solvent, the total electrostatic energy is the sum of the electrostatic interaction energy of the solute charges with solvent (and mobile ions) and with other solute charges. In this arrangement, the total electrostatic energy in the continuum solvent framework can be decomposed into a coulombic term (G_{coul}), arising from the interaction amongst the solute charges occurring in a lower dielectric medium and a reaction-field term (G_{rf}), which signifies the interaction of the solute with the polarized solvent medium. When implicit ions are included in the equation, a further contribution is made to the

reaction field term. By taking the difference of the total electrostatic energy in the two setups (zero in the gas-phase), the polar solvation energy can be expressed as

$$\Delta G_{solv}^{polar} = G_{coul} + G_{rf} \tag{9}$$

In the limit of lower magnitude of solute charges, linear form of PBE can be invoked and the total electrostatic energy in that case amounts to the sum of the total electrostatic potential energy of individual solute atoms. The total potential "felt" by an atom is the sum of those arising from the Coulombic fields of other solute atoms (inside the dielectric cavity) and those arising due to the reaction field induced in the solvent due to the solute charges.

$$\Delta G_{solv}^{polar} = \frac{1}{2}\sum_{i}^{N} q_i \phi_i = \frac{1}{2}\sum_{i}^{N} q_i (\phi_{coul,i} + \phi_{rf,i}) \tag{10}$$

In the PBE formalism, the above method is popularly used to determine the polar component of the solvation energy. In fact, this formula is integral to *Delphi*'s energy calculations and is used throughout this work, unless otherwise is stated.

1.4.2 Non-polar solvation energy

The non-polar term, as mentioned, indicates the work done to create a cavity amidst the bulk solvent which can fit the solute atoms. This invites a relationship of ΔG_{solv}^{np} with the volume and surface area (SA) of the solute, both geometrical estimators of its size, which intends to capture the repulsion of the water molecules from the relatively less polar solute as the cavity is created. This is also known as the *hydrophobic effect*.

But drawbacks of this simple linear non-polar model have come under light and they have led to refinement of the non-polar energy models to include an additional dispersion energy term (the term deliberately discarded in equation 8)[67]. The dispersion term is meant to capture the effect of turning on the Van der Waals (vdW) interactions between the hypothetically uncharged solute and the solvent once the cavity is created. The addition of this term has been shown to improve the quality of predictions made by non-polar energy models[68, 69].

1.5 Delphi: A PBE solver package

Delphi is a popularly used computer program that solves the PBE to deliver the potential distribution in the space of a solvated biomolecular system. The program, currently written in C++, is available for download (latest being version 8.0+) at no cost from *http://compbio.clemson.edu/delphi*[70, 71]. The software is designed to operate on Linux and Mac operating systems with new features for MPI and OpenMP parallelization which can be used for investigating larger mesoscopic systems like the virus assembly.

1.5.1 Finite Difference representation

Delphi uses a finite difference method to solve the PBE[40, 41]. Since its inception, it was designed to use this technique to deliver electrostatic potential distribution in space by discretizing it (*a.k.a* its domain). The space contains the solute (protein, DNA, RNA and small molecules) approximately at its center by default and a volume of solvent continuum surrounding it. The electrostatic potential is used to determine different energy terms (per the user's request) but most

commonly it is used to obtain the polar of the electrostatic component of the solvation energy (e.g. equation 10).

To enable finite difference PB calculation, the analytical PBE is manipulated to fit into the discretized space for which the derivation is well explained in some of the first works on *Delphi*[40, 41]. Essentially, the integral form of Gauss's law (equation 1) is obtained by volume integration of the gradient and the charge density term and the resultant bounding surface integral leads to the following expression which can be solved iteratively.

$$\phi_0 = \left[\frac{\left(\sum_{i=1}^{6} \epsilon_i \phi_i\right) + \frac{4\pi q_0}{h}}{\left(\sum_{i=1}^{6} \epsilon_i\right) + (\kappa_D h)^2}\right] \tag{11}$$

The purpose of discretization is to provide points in 3D space at which the potential (ϕ_0) and other quantities will be determined to yield a field. In *Delphi*'s glossary, these points are referred to as the *grid points.* The discretization is regular and isotropic, which results in a cubical box with sides long enough to accommodate the solute in question with a layer of solvent's continuum around it. This box is and will also be referred to as the computational box. The computational box is the union of non-overlapping cubes of side length h which indicate the resolution of the

discretization. With a total of M grid points along each direction (*aka grid size*), the total side length of the computational box equals $\boldsymbol{Mh}$. A factor called *scale* is defined as $\boldsymbol{1/h}$ which tells the number of *grid points* placed per Angstrom. The charges in the solute are "projected" onto the nearest grid points in space. The total charge on a grid point then amounts to $\boldsymbol{q_0}$. The space dependent dielectric value is specified between two *grid points*, i.e. at *midpoints*. Midpoints are likewise located at regular distances (h) and a distance of $\boldsymbol{h/2}$ from the nearest two grid points. Ultimately, the 2nd order nature of PBE and leads to the above expression (equation 11) where the potential $\boldsymbol{\phi_0}$ on a grid point with a total charge $\boldsymbol{q_0}$ depends on the potential of the 6 of the nearest neighbors $\boldsymbol{[\phi_i]_{i=1}^6}$ and the dielectric values at the midpoints that connect them with it $\boldsymbol{[\epsilon]_{i=1}^6}$. **Figure** 1.5.1 provides an illustration for a single cube and its vicinity and shows how they are juxtaposed to form a larger computational box.

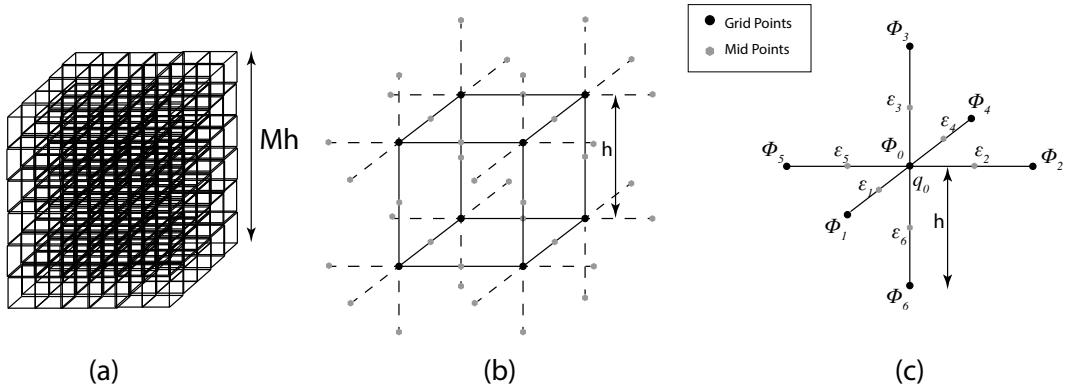


***Figure** 1.5.1: **Discretization of the space by Delphi.** The figures illustrate how the domain of the PBE is discretized into regularly sized cubes by Delphi in order to solve the PBE suing the numerical finite difference method. (a) The computational box with M grid points per side. (b) An expanded view of a cube with side length 'h' showing the positioning of the grid points and the mid-points. The solid black line outlines the cube's edges and the dashed lines represent the edges of the neighboring cubes (applicable for the non-boundary cubes only) (c) Points where an arbitrary charge (q_0), the dielectric values (ϵ_i) and electrostatic potentials (ϕ_i) are assigned.*

Typically, the user provides what is called a *perfil* that suggests the size of the box, in that the value of this parameter (in %) suggests the upper bound of the volume the solute must occupy in the larger cubical box. There are other parameters that the user can provide to supply similar information, but they are left out of this particular discussion to stay within the intended scope. For more details, one is referred to the *Delphi*'s user manual[1].

1.5.2 Overview of Delphi's workflow

The workflow of *Delphi* entails a sequential operation of various computational classes. A general outline of the program, with all of its classes is

[1] See *http://compbio.clemson.edu/downloadDir/delphi/delphi_manual8.pdf* .

shown in **Figure** 1.5.2. Each of these classes have their special functions and are all connected through a global data container. The data container is populated with default as well as user provided values for all the variables that are used during a *Delphi* run.

The work presented by means of this dissertation were mostly focused on modifying the '*Space'* class and the '*Energy'* class of the *Delphi* program. The contents in Chapter 2, 3 and 4 have been implemented through changes in the '*Space'* class and the contents in Chapter 5 have been implemented through changes in, both, '*Space'* and '*Energy'* classes. At times, the term '*module'* will be used to denote a '*class'*. Needless to say, the addition of new features also involved some changes in the '*IO*' class and other minor alterations in other classes, but they are not discussed here for the sake of brevity and relevance.

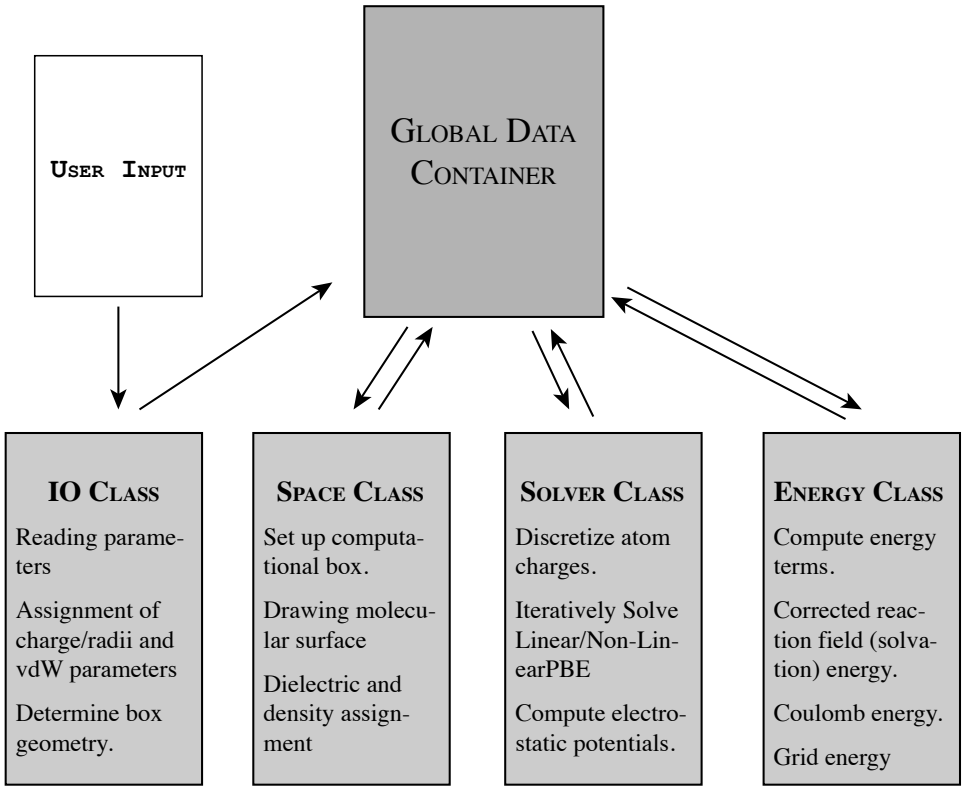


***Figure** 1.5.2: **Delphi's workflow**. The schematic presents the order of the functions/operations performed by Delphi in order to solve the PBE. Based on the 'User input', the IO class updates default global values of Delphi (stored in the Global Data Container). This triggers the Space, Solver and Energy classes to follow in this order as they receive parameters from the data container, use them and update them for the next order of execution. The arrows in the schematic indicate the direction of the flow of data in the form of various variables used throughout the run of Delphi.*

1.5.3 Applications of Delphi through packages and webservers

Delphi's widespread popularity is a testimony of its contribution to the scientific studies that based on understanding the role of electrostatics in molecular biophysics. As emphasized so far, electrostatics is a crucial element of molecular biophysics studies and with the variety of applications that *Delphi* has been involved in, it only shows the range of topics that come under its ambit.

Delphi is used by the community worldwide, directly and indirectly. Direct use of *Delphi* is possible through its standalone version (downloadable from *http://compbio.clemson.edu/delphi*) and through its web-server (*http://compbio.clemson.edu/sapp/delphi_webserver*). Indirect use of *Delphi* is possible through other packages and webservers that operate on *Delphi*. *DelphiPka*[72] is a tool that uses *Delphi* internally to determine the protonation state of the polar residues in a protein (or nucleotide bases of DNA/RNA) at a given pH value. pH determined protonation states are crucial to understanding the pH-dependent binding affinities of molecules and their mutants[14, 73]. *DelphiPka's* services are also made available through an exclusive webserver[74]. On top of this, three other webservers make use of *Delphi* to predict the effect of missense mutations

on the binding affinity of protein-protein complexes (SAAMBE)[75, 76], protein-DNA complexes (SAMPDI)[77] and the folding free energy of a protein (SAAFEC)[78].

Delphi has also been enhanced with features that allow the computation of the surface potentials around a biomolecule[79]. This feature is useful in predicting the ζ-potential of proteins and other biomolecules. This bears an additional advantage since this potential is typically inferred from the electrophoretic mobility of the solute/particle in question spawned by an external field due to its presence in an electrolyte buffer. The relationship is based on numerous classical and rather ideal assumptions about the structure of the molecule and typically none of them are valid for proteins by virtue of its complicated geometric shape and inhomogeneous charge distributions. With *Delphi*'s PBE solving capacity, this module can deliver the average and the distribution of electrostatic potential at any distance from the surface of a molecule of any shape and charge distribution. The role of explicit non-specific surface bound ions, whose positions were predicted using another *Delphi*-based web-tool called BION[80], is very critical to the surface potential distribution and its average value.

Another use of *Delphi*'s machinery is made by *DelphiForce*, a package[81] and a webserver[82] that predicts the force on a group(s) of atoms due to the charges from another group(s) of atoms. These force calculations differ from the typical qE Coulombic forces in that the forces are computed from the gradient of the potential which in turn is computed by considering the presence of varying dielectric environments. *DelphiForce* has been used in past works to observe the role of electrostatic forces on the mechanism of translocation of motor transport proteins on the microtubules inside the cells[83]. It has also proved handy in determining the effects of mutations in the proteins involved in this mechanism by means of changes incurred in the electrostatic forces at play[84].

With new features of *Delphi* underway, it is expected that its utility as a robust and versatile computational package will expand.

1.6 Summary

This introductory chapter presents the fundamental concepts used in the rest of this dissertation. First, the importance of studying electrostatic forces and interaction in molecular biophysics is discussed with emphasis on the role of solvent and solvation effects. Then the computational schemes currently available for such

studies are presented and their distinctions are clarified. The relevance of the Poisson-Boltzmann formalism to this work is stated and its key concepts are presented in order to lay the groundwork for the developments presented here. Finally, the implementation of the PBE formalism in *Delphi*, a C++ based PBE solver, is presented.

1.7 Outline of the dissertation

Following this introductory chapter, the dissertation is laid out to present the conceptual basis and the applications and implementation of the Gaussian-based model of dielectric distribution and computing molecular volume and surface area.

1. Chapter 2 outlines the Gaussian description of solute atoms and emphasizes its use to obtain an inhomogeneous space-dependent dielectric distribution with no dielectric boundary between the solute and the solvent regions. It presents the qualitative aspects of the Gaussian-based dielectric distribution model for solving PBE and attempts to illustrate its physically appealing characteristics.
2. In Chapter 3, the focus is shifted towards demonstrating results from explicit solvent MD simulations that enforce the idea behind an

inhomogeneous dielectric distribution and how Gaussian-model can effectively capture it.

3. In Chapter 4, the ability of the Gaussian model to reproduce ensemble averaged polar solvation effects is presented. The chapter highlights the features of the Gaussian dielectric model that enable it to do so with fair success and how this feature provides it an edge over the traditional 2-dielectric setup.
4. In Chapter 5, the integration of the Gaussian-based atomic model with *Delphi*'s finite difference platform is meticulously discussed to show its use for computing non-polar components of the solvation free energy.
5. Chapter 6 presents the concluding remarks and is drafted as a compendium of the contents of this dissertation.

2 GAUSSIAN-BASED MODEL OF ATOMS AND DERIVATION OF A SMOOTH DIELECTRIC FUNCTION

In this chapter, the Gaussian based model of atoms is introduced by means of its mathematical formulation. The use of a modified version of this model to derive an inhomogeneous and continuous dielectric distribution is presented. This model delivers a smooth transition of dielectric properties from the macromolecular interior to the solvent phase, eliminating any unphysical surface separating the two phases. Using various examples of macromolecular binding, its utility is demonstrated and comparisons with the conventional 2-dielectric model are illustrated. Some additional abilities of this model, viz. to account for the effect of electrolytes in the solution and to render the distribution profile of water across a lipid membrane, are also showcased. The contents of the chapter resemble with that of a work previously published[85].

2.1 Gaussian model of atoms

The Gaussian model of atoms is inspired by the seminal work of Grant and Pickup[64]. The idea behind the Gaussian model of atoms is to represent each atom as an atom-centered Gaussian density function as opposed to a hard sphere. Some new conventions and symbolisms are adapted here to describe the model. An atom *'i'* with Van der Waals radius R_i and coordinate $\vec{r_i}$ is described in a Gaussian representation via a density function given by

$$g_i = p_i exp\ (-\alpha_i(\vec{r} - \vec{r_i})) \tag{12}$$

Argument α_i of Gaussian exponent function can be further expressed using a dimensionless parameter κ and height factor p_i such that

$$\alpha_i = \frac{\kappa}{R_i^2}$$

$$p_i = \frac{4\pi}{3}\left(\frac{\kappa}{\pi}\right)^{\frac{3}{2}} \tag{13}$$

The above relations ensure that the volume, obtained from the volume integral of this density function, equals the hard-sphere volume ($V_i = \frac{4}{3}\pi R_i^3$) of the atom.

This Gaussian model can be conditioned to yield a probability function by removing the height factor and yielding a dimensionless factor bound between 0 and 1. As reported in Ref [86], the space-dependent function can be expressed as a function of scaling factor σ and atom-wise radius R_i.

$$\rho_i(\vec{r}) = exp\left(-\frac{|\vec{r} - \vec{r_i}|^2}{\sigma^2 R_i^2}\right) \forall\, i \in atoms \tag{14}$$

This can be used to obtain a collective probability function ($\rho_{mol}(\vec{r})$) which indicates the probability of finding a solute atom anywhere in the space of interest. In the specific context of solving for potentials, the space is the volume occupied by the computational box where PBE is to be solved. From the collective spatial probability, the space-wide dielectric distribution ($\epsilon(\vec{r})$) can be determined using a linear relationship with a reference solute dielectric (ϵ_{ref}) and the solvent dielectric values (ϵ_{out}).

$$\rho_{mol}(\vec{r}) = 1 - \prod_i (1 - \rho_i(\vec{r})) \tag{15}$$

$$\epsilon(\vec{r}) = \rho_{mol}(\vec{r})\epsilon_{ref} + \left(1 - \rho_{mol}(\vec{r})\right)\epsilon_{out} \tag{16}$$

The result is a smooth Gaussian-based dielectric function throughout the entire computational space. The necessity of such an approach is evident from the other works[87, 88], which show that the water molecules in the proximity of the macromolecule and inside its cavities, have different dielectric responses from those far out in the bulk region. Moreover, an inhomogeneous dielectric distribution in the region between the molecules also highlight how the long-range electrostatic interactions are affected in the process of recognition before binding[81]. The Gaussian-based smooth dielectric model has been implemented in *Delphi*[71, 89] Please also note that the solute dielectric is denoted by ϵ_{ref} in the context of the Gaussian dielectric model instead of ϵ_{in} but the two terms are synonymous. Both of them technically imply the same quantity which is the lowest dielectric value in the solvated system.

In **Figure** 2.1.1, the Gaussian based model of atoms derived from the atomic probability function is illustrated. The figure shows cases of a system with only 1 atom, 2 atoms and for a real protein (PDB: 2NVH). For comparison, the hard-sphere equivalents are also shown.

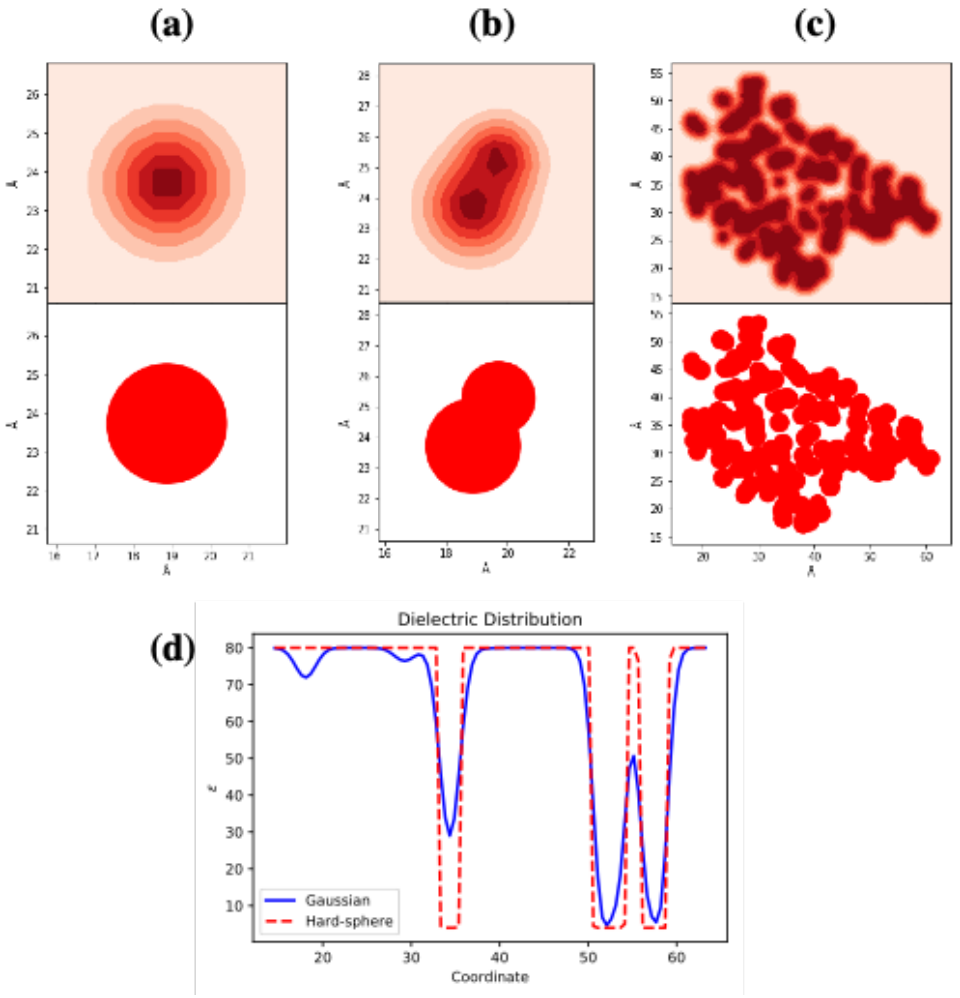


***Figure** 2.1.1: **Gaussian model of atoms vs hard-sphere model of atoms**. (a-c) From left to right, in the top panel, the hard-sphere representation of a one-atom system, a two-atom system and a real protein, IL-1β (PDB: 2NVH) is show. In the bottom panel, the equivalent Gaussian model representations are shown. (d) Dielectric distribution obtained using the Gaussian model (solid blue line) and the hard-sphere model (dashed red line), along an arbitrarily chosen axis, cutting through the slice of the protein in (c) is shown. The reference dielectric (ϵ_{ref}) is set at 4 and the solvent dielectric (ϵ_{out}) is 80.*

2.1.1 Super Gaussian model of atoms

A variant of the Gaussian model has recently been proposed, known as the Super-Gaussian model of atoms[90]. The authors of this work demonstrated the use of this model in conjunction with a third parameter, ϵ_{gap}, which defines an upper bound for the dielectric of the interstitial cavities in a protein's structure. Leaving this parameter aside, the formalism of the Super-Gaussian model is fundamentally similar to that of the Gaussian model of atom except for an additional power of the exponent. Retaining the symbolic references of equation 14, the atomic probability in the Super-Gaussian model can be expressed as:

$$\rho_i(\vec{r}) = exp\left(-\left(\frac{|\vec{r} - \vec{r_i}|^2}{\sigma^2 R_i^2}\right)^m\right) \forall\, i \in atoms \tag{17}$$

When the factor m equals 1, the expression becomes identical to our Gaussian model of atoms. For higher values (typically integers), it acquires a Super-Gaussian form. In the limit of higher m values, the expression begins to take the form of hard-sphere representation with abrupt transitions in ρ and ϵ. Regardless, with the above expression for ρ, equation 15 and equation 16 forms can be unambiguously applied. As mentioned above, the authors of Ref [90], have used an additional *gap epsilon* to

formulate the model. For this discussion, it has been left out as our approaches and motivations differ.

In **Figure** 2.1.2(a, b) the effects of m on the atomic probability function (ρ) and therefore on the dielectric distribution (ϵ) are shown. As m increases, the dielectric profile visually becomes sharper, though its mathematical continuity is infinitely preserved. As is evident, at $m=6$, the profile is highly similar to the profile obtained using the hard-sphere model shown in **Figure** 2.1.1(d). In **Figure** 2.1.2(c), the change in the visual appearance of the protein (PDB: 2NVH) due to the value of m is also illustrated.

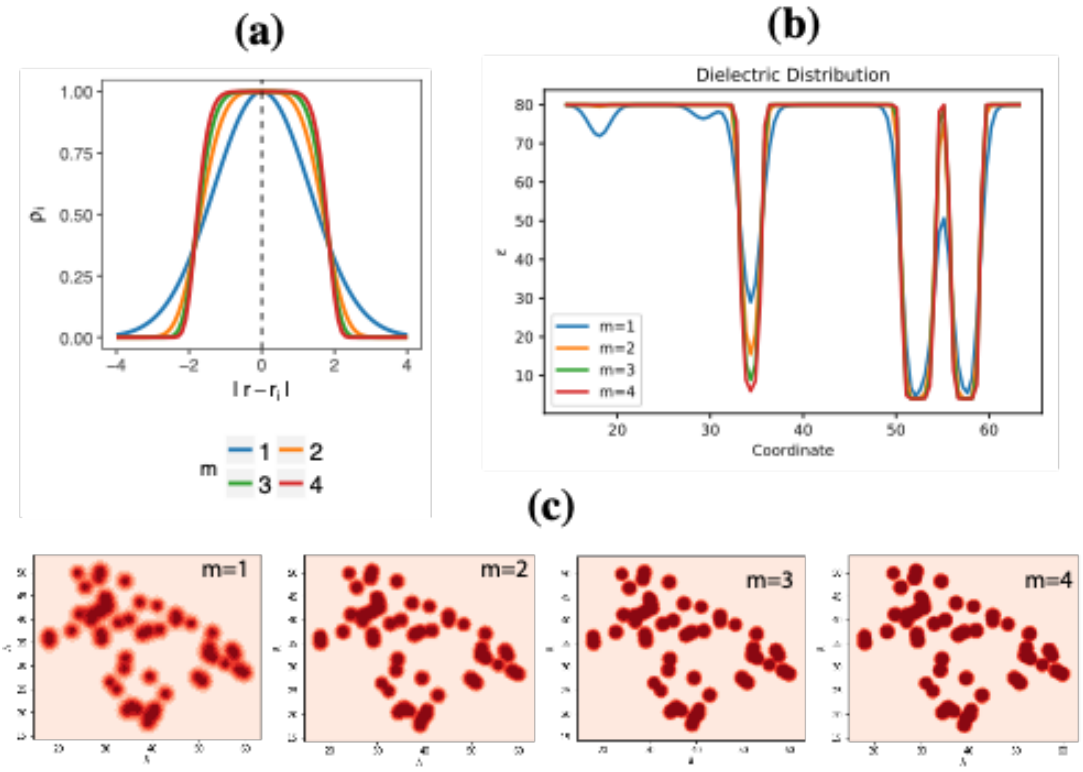


***Figure** 2.1.2: **Gaussian and Super-Gaussian forms**. (a) For a single atom of radius 2 (with σ=0.93), the profile of atomic probability function, $\rho_i(|r - r_i|)$, is shown for 'm' ranging from 1 through 4. As the value of m increases, the profile appears to take the form of a hard-sphere Heaviside function. (b) The dielectric distribution obtained with 'm' from 1 through 4, along an arbitrarily chosen axis, cutting through the slice of a protein (PDB: 2NVH). The reference dielectric (ϵ_{ref}) is set at 4 and the solvent dielectric (ϵ_{out}) is 80. (c) For m=1 through m=4, the Gaussian model's depiction for this protein is shown.*

2.2 Gaussian-dielectric model for protein-protein interactions (*Barnsase-Barstar)*

The *Barnase-Barstar* complex from *Bacillus amyloliquefaciens*, where Barnase (*Bn*) is an extracellular ribonuclease and Barstar (*Bs*) is its intracellular inhibitor, has been used extensively in previous studies (e.g. Ref [91-93]). An experimental study of their water mediated-interaction has reported that the water molecules (H_2O) crystallized at the interface have different B-factors[53]. The different B-factors have been attributed to the number of H-bonds these water molecules made with either or both monomers and their ability, henceforth, their ability to reorient and respond to local electrostatic field.

The Gaussian-based dielectric model is used to provide a description of the dielectric distribution at the interface of *Bn-Bs* complex (PDB: 1X1X) as its monomers are moved apart in space. For comparison, the same is done with the *traditional 2-dielectric* model or the *hard-sphere model*. The results are shown in **Figure** 2.2.1 for configurations where the monomer centers are moved apart by distances in the range of 0-10Å.

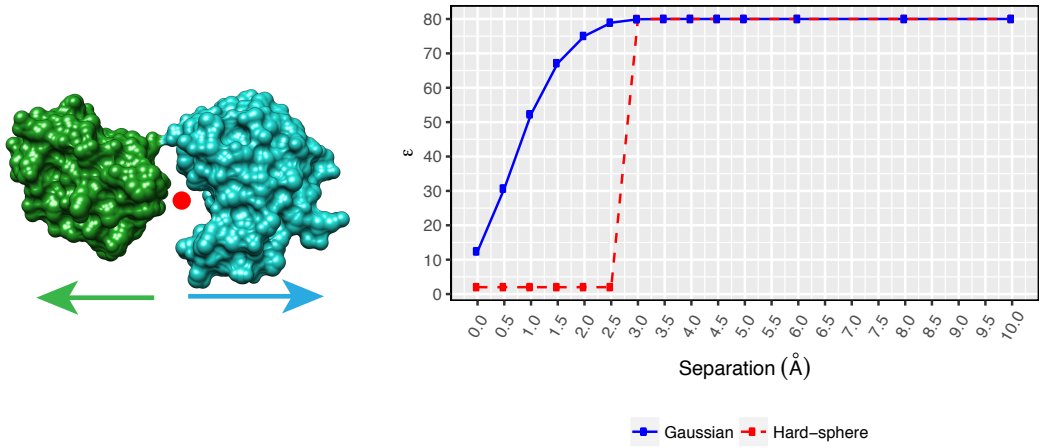


***Figure** 2.2.1: **Dielectric at the binding interface of a protein-protein complex**. For Barnase-Barstar protein complex (PDB 1X1X), the dielectric value at the center of the binding interface (marked by the red dot in the left **Figure**) is plotted as a function of the distance by which the monomers are separated in space along an arbitrary direction. Both, the Gaussian and the hard-sphere models are used to illustrate the profile and also highlight the difference between the two.*

One can appreciate the lack of sharp change in the dielectric achieved with the Gaussian model, suggesting a smooth change of dielectric constant value in the space between the monomers as they are moved apart. Even at very low separations, the space between the *Bn* and *Bs* exhibits a dielectric between ϵ_{ref} and ϵ_{out} but not identical to ϵ_{ref}. Such a trend depicts how the space between interfaces begin to gain higher dielectric constant mimicking the increased flexibility of interfacial residues upon separation and increased probability of water molecules to enter there.

This also resonates with the observation that the interfacial water molecules, when there is very little room between interfaces, have different mobility compared to the bulk water due to plausible interactions with the monomers.

2.3 Modeling salt contribution using the Gaussian-based model

The Gaussian-model was refined to appropriately model the presence of implicit electrolytes in the solvent phase. The presence of electrolytes in the PBE models is accounted for by their Boltzmann distribution, i.e. their concentration in the solvent phase is proportional to the Boltzmann factor corresponding to the electrostatic energy of an ion at some point in the solvent region. In the 2-dielectric setup, the salt is homogeneously accessible to all the region that lies outside the ion-exclusion surface. The smooth dielectric transition due to the Gaussian-based dielectric model eliminates the provision of a clearly demarcated solvent region which therefore, challenges its ability to incorporate the non-trivial effects of salt on binding [94, 95]. This issue has been investigated and solved in a recent work published by our lab [96]. The publication that presented this idea was also able to demonstrate

the ability of this feature to predict the effect of salt on the binding affinity, as shown earlier by Bertonati *et. al.* [95] using the conventional 2-dielectric model.

The solution to this problem was inspired by the fact that charges, which migrate to regions with different dielectric constants, sustain a (de)-solvation energy or a "penalty". In the Gaussian-based model, this penalty is expressed by the absolute value of the energy of transfer across two dielectric media of a centrosymmetric ion obtained using the Born formalism. In SI units, the expression acquires the following form

$$\Delta G_{penalty}(\vec{r}) = +\frac{N_A z_i^2 e^2}{8\pi\epsilon_0 r_0}\left(\frac{1}{\epsilon(\vec{r})} - \frac{1}{\epsilon_{out}}\right) \tag{18}$$

Here N_A is Avogadro's constant, e is the elementary charge and $\epsilon(\vec{r})$ – space-dependent dielectric at a calculated by Gaussian-based model. The penalty term influences an ion's ability (of charge $q_i = z_i e$ and radius r_0) to be present at some $\vec{r}$ in the solvent medium which when added to the electrostatic potential there ($-q_i\varphi(\vec{r})$) renders the following expression for PBE:

$$\vec{\nabla} \cdot [\epsilon(\vec{r})\nabla\phi(\vec{r})]$$

$$= -4\pi\left(\rho_{solute}(\vec{r}) + \sum_{i=1}^{N} q_i \, c_i^{\circ} exp\left(\frac{-q_i\phi(\vec{r}) + \Delta G_{penalty}(\vec{r})}{RT}\right)\right) \quad (19)$$

Quantities $\varphi(\vec{r})$, and $\rho_{solute}(\vec{r})$ are the electrostatic potential and charge density of a solute at $\vec{r}$, respectively; c_i° is the bulk ion concentration and T is the temperature.

Figure 2.3.1(a) shows the profile of the $\Delta G_{penalty}$ term as a function of the coordinate being probed. As one approaches the solute region (denoted by a rectangular slab of 4Å width), the penalty term increases non-linearly to the point that its presence inside the solute region drops to zero following the Boltzmann distribution.

Figure 2.3.1(b) presents a visual description of the distribution of salt obtained by the modified PBE formalism in a plane from the computational box that contains the *Barnase-Barstar* complex. It can be seen that ions can propagate inside

the binding interface if there are small cavities allowing for transient ions to come in.

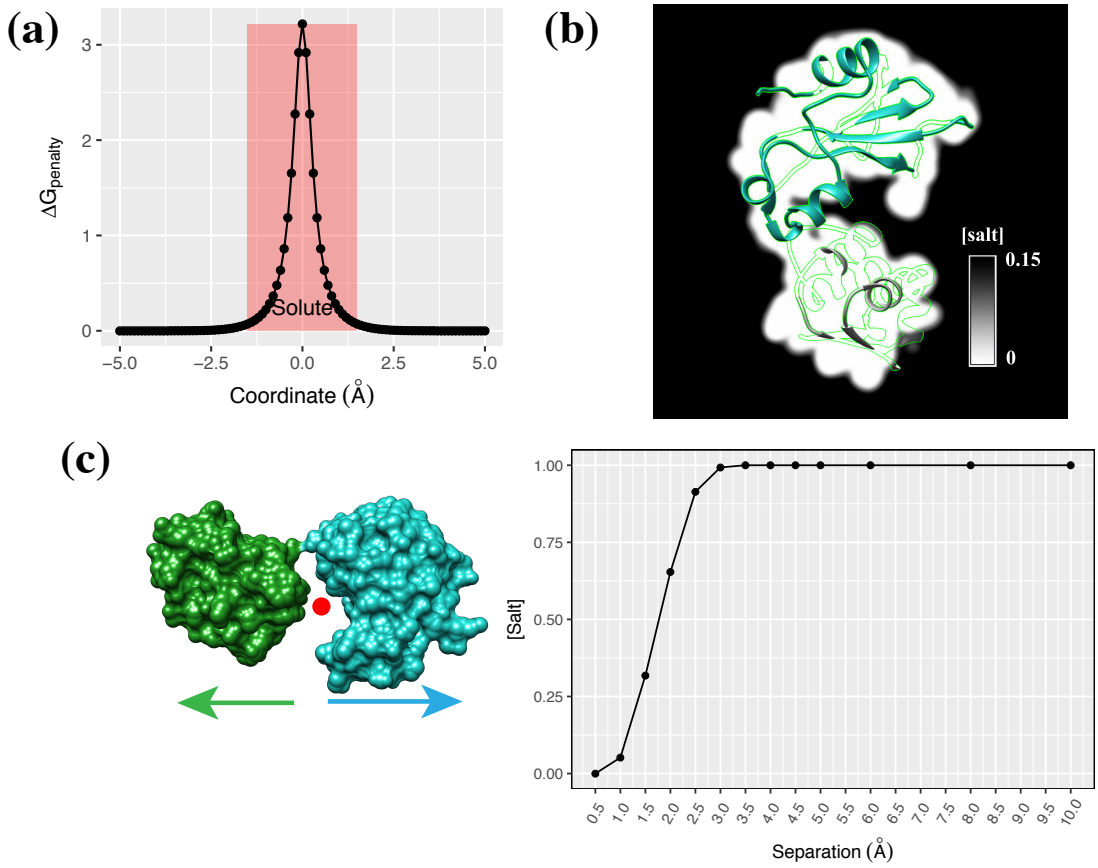


***Figure** 2.3.1: **Salt treatment using the Gaussian model**. (a) The penalty term added to a salt's electrostatic energy obtained after solving PBE is plotted as a function of position in space. The space contains of a solute, represented by a rectangular slab of width 4Å (filled with pink color). Everything outside is assumed to be filled by the solvent. (b) An illustration of the salt concentration distribution generated using Delphi around the Barnase-Barstar complex (PDB: 1X1X). (c) Salt concentration at the binding interface of the Barnase-Barstar complex (computed at the red point shown in the cartoon representation of the complex) is plotted as a function of the distance of separation of the monomers.*

In

Figure 2.3.1(c), the change in the salt concentration in the binding interface region of the complex is shown as its monomers are separated away in an arbitrary but consistent direction.

2.4 Water distribution across lipid bilayers using Gaussian-based dielectric model

Lipid bilayer membranes in animal cells are exposed to the extra-intracellular fluids, which are aqueous electrolyte solutions. These membranes sustain very high hydrostatic and osmotic pressures (as high as 18KPa [97]) to preserve the shape of the cell and contain the cytoplasmic contents. Therefore, interaction, diffusion and

permeation of water with and across lipid membranes are vital for osmoregulation and cell lysis. Subsequently, any lipid-water model should be appropriately represented for a computational study.

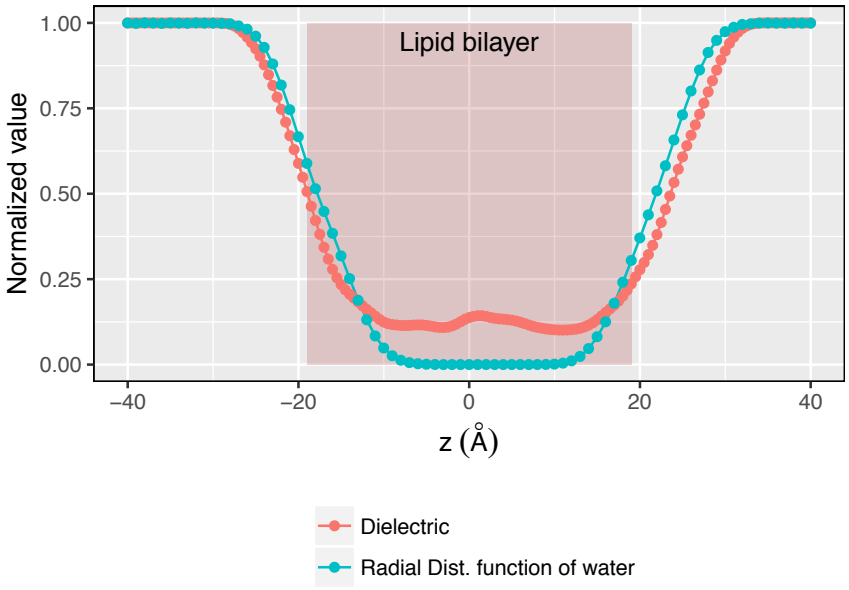


***Figure** 2.4.1: **Dielectric distribution and water's radial distribution function across a lipid bilayer membrane.** The **Figure** shows the normalized values of the radial distribution function of water's oxygen atom and the dielectric distribution obtained using the Gaussian model along the transverse direction perpendicular to a lipid membrane's plane. The membrane region is depicted by a rectangular slab of 38 Å width which is the typical value of the POPC head-to-head distance (bilayer thickness). The normalization is done with respect to the maximum value of the corresponding data. In the case of dielectric, the maximum value was 80 (solvent dielectric).*

Using the Gaussian-based dielectric model, it is shown that the dielectric distribution across a lipid membrane matches well with the averaged distribution of

water surrounding it which is computed from a 12ns explicit water NPT-MD simulation of a POPC-lipid bilayer patch. The results are illustrated in **Figure** 2.4.1**.** For better perspective, the values are normalized with respect to their respective maximum (e.g. 80 for dielectric constant). It can be seen that water molecules propagate inside the membrane resulting in a smooth profile from bulk water density to zero density in the core of the lipid bilayer. The dielectric constant profile replicates the trend by smoothly decreasing from 80 in the bulk phase to lower values inside the membrane. This finding provides additional support for our claim that the Gaussian-based dielectric function mimics the effect of water molecules near the macromolecular interfaces. The shaded region in **Figure** 2.4.1 is a crude representation of the membrane slab of thickness 38Å; the typical thickness of POPC membranes [98].

2.5 Summary

This chapter presents the mathematical formulation of the Gaussian-based description of atoms as opposed to the conventional hard-sphere model. From there, the derivation of the dielectric function using the Gaussian function as its basis is presented. The model's qualitative aspects are exhibited through three different

examples that show the ability of the Gaussian model to capture a realistic distribution of the solvent around biomolecules. In addition to the qualitative appeal, previous studies have also reported its success in predicting pKa's [72], optimum pH and proton transfer analysis[73], predicting change in binding free energy upon mutation[99, 100], etc. It has also been shown that this dielectric model along with salt contribution outperforms the tradition 2-dielectric model in predicting the pKa shifts of ionizable and polar residues incurred due to the protein's configuration[101].

As the recent advances in solvation models continue to provide a more realistic picture of macromolecular behavior in water, efforts are also needed in developing time-inexpensive models for solvation and binding that can deliver experimentally measurable quantities. This is of importance because relevant experimental techniques deliver quantities that are ensemble averaged and are not merely pertinent to measurements made on a single molecule. At present, ensemble averaged quantities can be obtained by protocols like MM/PBSA[102] and MM/GBSA[103], which are rather time-consuming. The Gaussian-based dielectric model, with its current abilities, can reproduce the ensemble average polar component of solvation energy from a single energy-minimized structure of a protein, which is discussed in detail in Chapter 4.

3 CONCEPTUAL VALIDITY OF THE GAUSSIAN MODEL: EVIDENCE FROM EXPLICIT WATER MD SIMULATIONS

In the previous chapters, the key concepts relevant to this dissertation have been discussed. A major highlight of Chapter 1 is the introduction of the continuum electrostatic in the framework of implicit solvent models using PBE formalism. It discusses how this model offers a viable alternative to the explicit solvent models of biomolecular systems by emphasizing the concreteness of its physical assumptions and its ability to be time-efficient. In those descriptions, the typical setup of a solvated biomolecular system is discussed and the concept of distinct but homogeneous dielectric media is presented. Chapter 2 presents an alternative way of representing such systems using a Gaussian-based model of atoms from which a Gaussian-based smooth dielectric model is derived. Its central idea of delivering a smooth, molecular-surface-free dielectric distribution is demonstrated through various examples that reflect the physical appeal and simplicity of this model. However, the discussions of the Gaussian model have been more perspective-oriented and qualitative. The aim of the current chapter is to present a Proof of Concept of

the Gaussian-based dielectric model using observations from explicit solvent MD simulations and eventually show the validity of the central assumptions of this model.

3.1 Revisiting the motivations behind the Gaussian-based dielectric model

The Gaussian-based model of dielectric distribution was first presented in Ref [104]. Later Li *et. al.*,[86] presented an alternative formalism of this model with the aim of generating a surface-free description of solvated biomolecular systems which would mimic the effect of conformational dynamics of the solute on solvation. As opposed to the traditional practice of representing the space as a disjoint union of two distinct dielectric media, solute and the solvent, the Gaussian model proposes that the presence of a strict dielectric surface is unphysical and that dielectric value must be smoothly and continuously distributed in space.

The origin of the idea behind the Gaussian-based dielectric model is rooted in reality. In physiological conditions, biomolecules are very dynamic and the effects of this flexibility must be considered in order to determine the correct solvation energies. These conformational dynamics entails both, small and large frequency

motions of the various units that build a biomolecule which emerge from the intra-molecular and inter-molecular forces of interaction and kinetic energies possessed by the constituent atoms. This flexibility continuously updates local interactions of solvent-exposed atoms with the solvent and other solute atoms. As a result, local structural and energy changes are constantly underway. Since dielectric distribution affects the structure-energy relations via screening of the electrostatic interactions within the solute and between the solute and solvent[105, 106], the effects of the conformational dynamics can be correctly captured using a dielectric model that respects the role of the varying local environments harbored in the system. In other words, an inhomogeneous dielectric distribution model can serve the purpose by retaining the simplicity and time-efficacy of the implicit solvent models, on one hand, and preserving the local structural phenomena, on the other. The Gaussian-based dielectric model was designed exactly for this reason.

The key assumption of the Gaussian-based dielectric model is that the regions with higher packing density (high $\rho(\vec{r})$) will experience restricted motion compared to those regions with looser packing (low $\rho(\vec{r})$). This translates to the ability of the atoms in either region to respond to an external electric field by virtue of its ability to rotate its dipole moment vector in order to minimize the effects of the

"perturbation". Due to higher packing, the dipole moment due to the group of atoms in those regions will feature very large rotational correlation times while those in the loosely packed regions will feature just the opposite. This implies that the highly packed regions will possess low polarizability (and therefore lower dielectric value) and the loosely packed regions will exhibit high polarizability (and therefore higher dielectric). In a sense, this assumption establishes an inverse relationship between the local structural packing density and the local dielectric value. MD simulation studies have been able to demonstrate such space-dependent dielectric properties of proteins[107, 108], specifically that the dielectric values are higher moving away from the center of a protein.

This relation also scales perfectly to the solvent medium. By representing the solvent region as a uniform structure-less dielectric medium, the traditional 2-dielectric setup automatically disregards the important solute-solvent interactions that are dominant at the interface. These interactions tend to anchor the solvent molecules at the interface and bereave them of their bulk properties. As a result, the residence time of water molecules is higher close to the solute or at any hydration site where interaction with the solute is viable[109, 110] and solvent mobility (translation and diffusion) is limited. Therefore, as one probes the solvent region, it

is safe to say that the dielectric value will gradually increase with the distance from the solute's surface. In addition, solvent molecules locked in the cavities and other interstitial accessible regions of the solute may exhibit dielectric properties which are very unlike that of the bulk and perhaps closer to that of the solute.

Energetically, the interaction of the polar groups of the solute with the solvent can have significant effect on its stability. There is an anti-correlation of the structural stability and the solvation energy, in that, favorable intramolecular Coulombic interactions imply a relatively unfavorable solvation energy and *vice versa (***Appendix A.4).** This establishes a balance between the two factors since both the energies are summed up when computing the polar binding free energy of a macromolecular complex in a solution.

The motivation of the Gaussian-based dielectric model, thus, is to be able to mimic the conformational flexibility in a system through a smooth dielectric distribution model and deliver solvation energy that matches well with experimental observations.

3.2 Molecular dynamics of a protein in explicit water

3.2.1 Selection of a protein with cavity waters

For this particular study, a protein with cavity waters in its crystal structure was sought. The protein of choice was the interleukin-1β (IL-1β) whose crystal structure is known to harbor 5 different internal cavities of which four are occupied by water (PDB ID: 2NVH) [111]. The same protein has also been used by Hazra *et. al.* [90] to present a Super-Gaussian model of dielectric distribution that emphasizes the importance of a distinct cavity dielectric value in addition to that of a solute and solvent. **Figure** 3.2.1 shows the crystal structure of this protein and separately highlights the cavities (and its waters wherever found).

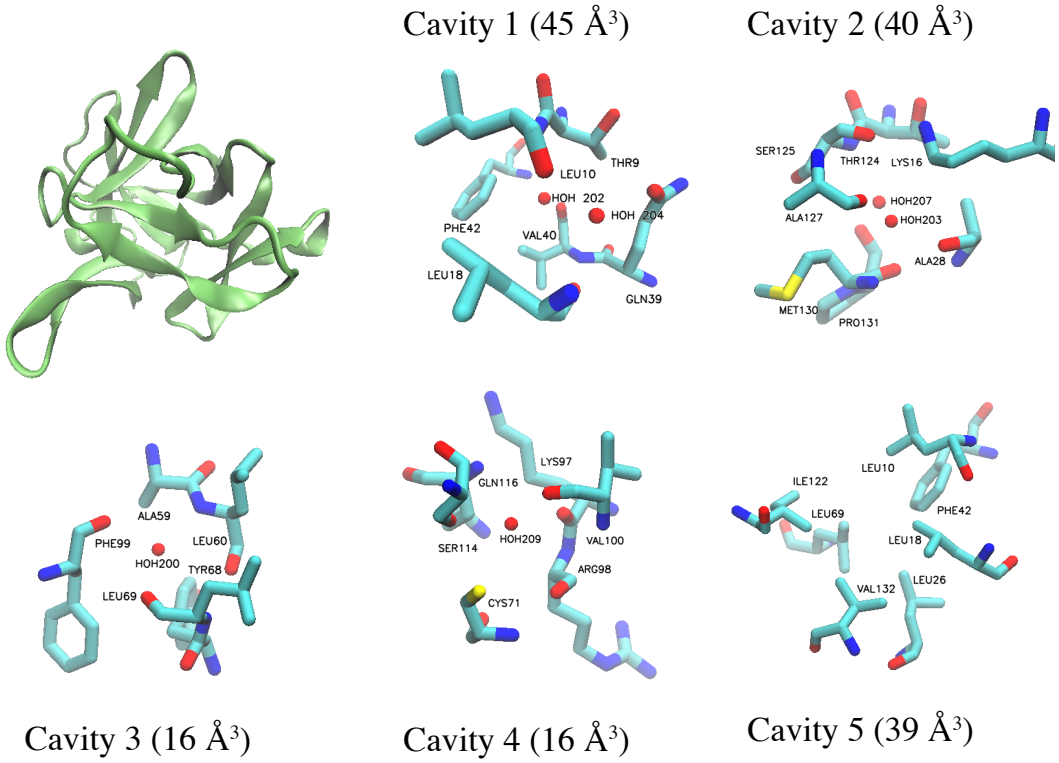


***Figure** 3.2.1: **A protein and its cavity with crystal waters.** The protein interleukin-1β (IL-1β) with PDB ID: 2NVH is shown on top-left corner. The protein is known to have five different cavities in its crystal structure of which 4 of them are occupied by crystal waters (Cavity 1-4). Cavity 1 and 2 contain two water molecules and cavity 3 and 4 contain one water molecule. Cavity 5 is the central non-polar cavity with no water present in it in the crystal structure. All the cavities are labelled and the water oxygen atoms are labelled and shown as red spheres. The volume of these cavities are also reported as mentioned in Ref[111].*

3.2.2 Molecular Dynamics

This protein was solvated in a bath of explicit solvents after removing all the crystal waters, including those that are present in the cavity. This was done to see if water molecules placed artificially find their way to and through these cavities during a MD run. Before running the MD simulations, the protein's crystal structure was protonated using the AMBER99SB force-field[112]. On protonation, all the Arginine (ARG), Histidine (HIS) and Lysine (LYS) residues had a charge of +1e and all the Glutamic acid (GLU) and Aspartic acid (ASP) residues had a charge of -1e. The total charge on the protein was -1. The protonated crystal structure was solvated in a bath of explicit water molecules with water molecules assigned the TIP3P form[113]. To make the system electrically neutral, Na and Cl ions were added while desiring that their concentration be 0.15M. First, 10000 steps of steepest descent minimization were performed while harmonically restraining the heavy atoms in the protein in order to guide the solvated system to the nearest local energy minimum. This was followed by 3 independent 200ps of isothermal-isochoric (NVT) equilibrations which subsequently branched off to 3 independent runs. Each NVT equilibration was followed by a 2ns long isothermal-isobaric (NPT) equilibration

period with temperature set at 300K and pressure set at 1 atm. During both the equilibration phases, the harmonic restraints on the heavy atoms were retained. After equilibrations, the system was allowed to evolve for 30ns under NPT conditions and no harmonic restraints. During the runs, the non-bonded forces were only applied within a cut-off of 12Å and the list was updated every 10 steps (20fs; each time-step was 2fs). The simulations were done using Periodic boundary conditions to minimize boundary effects and Particle-Mesh-Ewald (PME)[114] was employed for long range non-bonded electrostatic force-calculations.

Configurations of the solvated system were sampled every 10ps. For analysis, only the last 20ns of the total 30ns per run was used which yielded a total of 2000 frames per run. In total 6000 frames were, therefore, analyzed. All the simulations and most analyses were carried out using the GROMACS package (v 5.0.5)[115, 116].

3.2.3 Force-field and water model combinations

In the preceding paragraph, the use of AMBER99SB force-field with TIP3P water model is mentioned. All of the processes were also repeated using the OPLS force-field[117] in conjunction with TIP4P water models. This was done to gauge

qualitative differences in water behavior incurred due to differences in the force-field/water-model combinations.

3.3 Analysis: Tempo-spatial properties of cavity and bulk water

The trajectory worth 20ns per run was analyzed to answer three main questions.

a) Were the cavities in the protein able to attract water during the simulation (occupancy)?

b) What was the typical residence time of the water occupying the cavities and how that differed from the non-cavity waters?

c) What was the dipole rotational relaxation time of these category of waters?

The last two analyses were also done to probe any space-dependent features of the water in the bulk, i.e. the residence and the dipole rotational relaxation time of water as a function of its distance from the surface of the protein were evaluated. Each of these questions are answered respectively in the following paragraphs.

3.3.1 Occupancy of the cavities

In each of the sampled snapshot of the solvated system, the number of water present in each of the 5 cavities were examined. A water was deemed to be present in a cavity at a certain time if any of its atoms (H, O, H) lay within 3 Å from the center of mass of the protein residues lining that cavity at that time. If at least one water fit this definition, the cavity was occupied at that time. For runs made using AMBER99SB/TIP3P and OPLSAA/TIP4P combinations, the occupancy is presented in the form of a histogram in **Figure** 3.3.1. for each of the five cavities, the histogram shows the typical number of water molecules present in them through each run.

As is evident, cavities 1 and 2, showed the tendency to have 1-3 water molecules occupying it at most times, though odd cases of no water molecule being present in them are also noticeable. Cavities 3 and 4 were typically found to be occupied by one water molecule for a vast majority of the simulation when the AMBER99SB/TIP3P combination was used. For the OPLSAA/TIP4P combination, cavity 3 was mostly unoccupied. Only for run labelled '1', it was found to contain one water throughout. Cavity 4, on the other hand, maintained its occupancy at 1

though there are visible cases when it had no water molecules and very rarely two water molecules. Cavity 5 featured zero occupancy consistently across all runs through both combinations.

Overall these observations align well with the experimental occupancies of these cavities [111]. All the cavities except Cavity '5', were typically occupied by water molecules. These observations address the first question pointed out at the start of this section by indicating that indeed, cavity waters are detected during MD runs even though the crystal cavity waters were removed before executing these runs. These observations also lay the groundwork for the next set of analyses where differences in the tempo-spatial properties of cavity and non-cavity (or bulk) water molecules are examined.

Of a special note is the behavior of Cavity '5' which retains its non-polar nature in full glory. It is an interesting subject of examination but it is beyond the scope of the current study. In fact, cases like this highlight the limitations of the model like the Gaussian-based dielectric model which are derived fundamentally from geometric features and not from biochemical features.

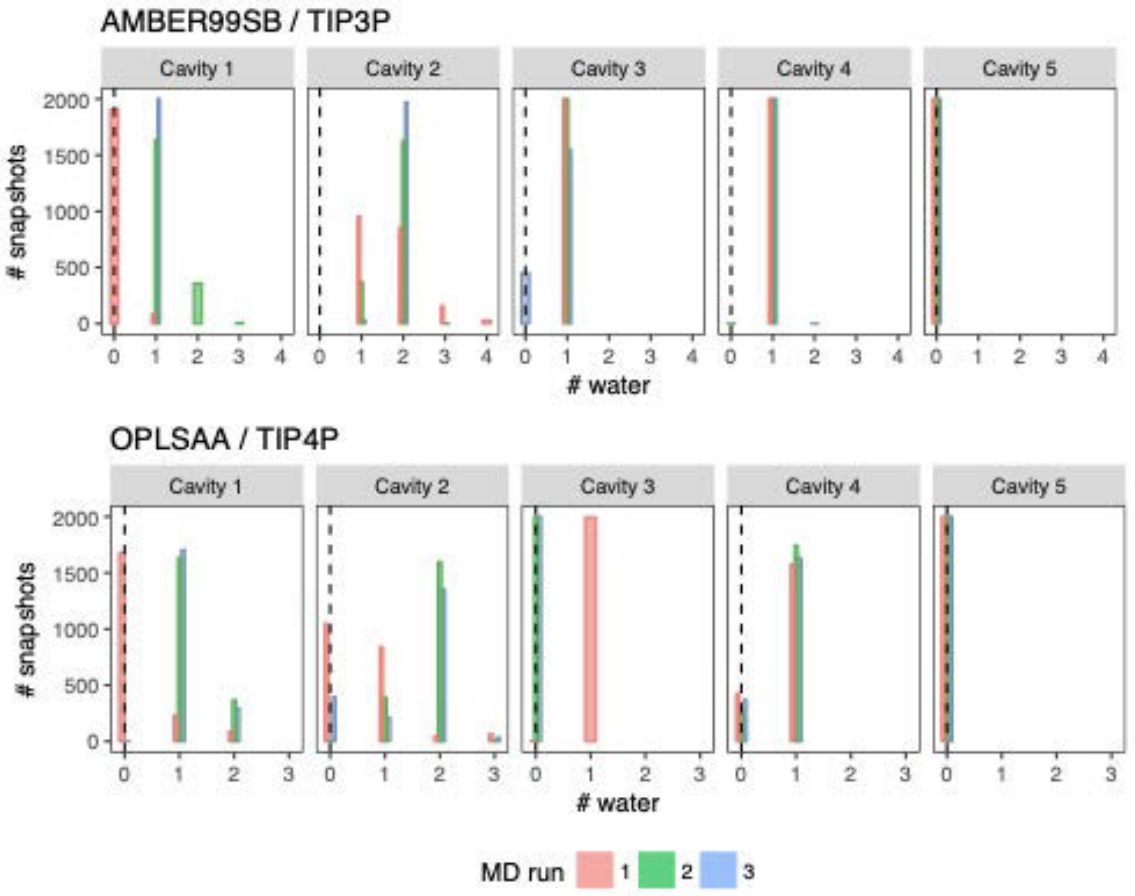


***Figure** 3.3.1: **Occupancy of cavities**. Histograms in this **Figure** show the typical occupancy of the 5 cavities in 2NVH. The data is collected across 2000 snapshots from all the MD runs performed using AMBER99SB/TIP3P (top) and OPLSAA/TIP4P(bottom) force-field/water-model combinations.*

3.3.2 Cavity vs bulk water: Mean residence time

Essentially, the mean residence time of water at a hydration site tells the "average" time a given water molecule might spend residing there. It can be computed in various ways[118] but for this study, the definition adopted by Makarov

et. al. [110] was used. Like theirs, the concept of time correlation function was used. The idea was to compute the probability of finding the same water molecule (through its index in a structure file) some time Δt later from a "starting point" and obtain an average of those probabilities for all the valid "starting" points in the trajectory. Mathematically, this is given by:

$$P(\Delta t) = \left\langle \frac{\#((water\ present\ at\ t_0 + \Delta t) \cap (water\ present\ at\ t_0))}{\#(water\ present\ at\ t_0)} \right\rangle_{t_0} \quad (20)$$

In the above equation, the correlation is computed for a set of time lags Δt and for a fixed value of it, the average probability is computed by averaging over all the "starting points" denoted using t_0. The probability, as such, is determined by calculating the number (#) of water indices common in snapshot at time t_0 and $(t_0 + \Delta t)$ and dividing that number by the total number of water molecules present at time t_0. For the rest of this chapter, the term $P(\Delta t)$ will be referred to as the *Index correlation function (ICF).*

Once the ICFs were computed, the profile was fit using a bi-exponential curve, whose form was inspired by the work of Makarov *et. al.*[110]. The idea behind using a bi-exponential curve instead of a mono-exponential curve was that hydration sites

can be visited by water which may tend to stay for shorter as well longer periods of time. The authors posit that both kinds of diffusion behavior are possible and must be accounted for. They also admit that this form doesn't capture the true behavior as it is more complicated than a binary perception they adopt. The curve has the following expression:

$$P(\Delta t) = a_0\left((1-w)e^{-\frac{\Delta t}{\tau_1}} + we^{-\frac{\Delta t}{\tau_2}}\right) \tag{21}$$

The annotations of the various symbols are as follows: Term a_0 denotes the occupancy of the site in question and ranges from 0 to 1 where 1 implies fully occupied by water. Weight factor w indicates the weightage of the term that is associated with water molecules with longer residence times and therefore $(1-w)$ is associated with those with shorter residence times. If w is closer to 1, the site is likely to be occupied with water with tendency to stay there for prolonged periods of time and *vice versa*. The corresponding residence times are given by τ_2 and τ_1. Thus, when assessing the values of these four parameters, they must not be interpreted independently of each other. Of specific interest is the value of τ_1 or the short residence time as the typical values lie within the range of the simulation time.

For cavity as well as the bulk waters, the ICFs were computed. For the former, the ICFs for each cavity was computed individually. For the bulk, an extended set of analyses were performed. The bulk solvent region was divided into 5 different concentric shells, each 3 Å thick, that were positioned 0, 3, 6, 9 and 12Å away from the solvent exposed residues of 2NVH (identified using NACCESS[119]). A schematic illustrating this division of the bulk volume into 5 concentric shells is presented in **Figure** 3.3.2. The ICFs were then fit using the bi-exponential curve and the resulting values of the parameters were compared. The ICFs for the four cavities (Cavity '5' is excluded because it had no water) are plotted in **Figure** 3.3.3 and that of the 5 hydration shells around the protein are plotted in **Figure** 3.3.4**.**

Upon comparing the two plots, a clear visual distinction can be made. Speaking qualitatively, the decay rate of the ICF of the cavity waters is much slower than those in the bulk hydration shells. In other words, the timescale of decay was larger for the former. With a bi-exponential fit, the value of τ_1 (the short mean residence time) was also obtained. The typical value of τ_1 for cavity waters was in the ballpark of 3000ps vs ~20ps or less for the hydration shells in the bulk.

This is a clear quantitative proof from an explicit water MD that the local environment of a water molecule has a profound effect on its mean residence time.

This qualitative finding is not novel. It is well resolved that water in the vicinity of proteins residues tends to interact with them and is therefore, anchored to that region. For example, the water at the interface of the *Barnase-Barstar* complex mediates their binding and therefore must feature properties very different from the bulk[53]. In the particular case of our study, the cavity water molecules interact with several residues, especially those that line the cavity and depending upon the hydrophobicity of these residues the mean residence time is influenced (see **Figure** 3.2.1). If the residues are polar or charged, it would interact favorably with water via H-bonds or salt-bridges or both. Energetically, it imparts more favorability to the solvation energy. If they were non-polar, residence time can be expected to be lower. As a matter of fact, Cavity 5, with its completely non-polar lining, does not harbor a single water molecule.

In the bulk region, comprised of disjoint hydration shells, water only in the first two shells appear to be affected by the protein's configuration. For shells farther away (more than 6Å away from the surface of the protein), the interactions with the protein are very weak. At such positions, the water molecules are only affected by the surrounding water molecules. By virtue of their smaller weights and hydrogen

binding abilities, the water molecules in those regions diffuse very freely and feature bulk-like properties.

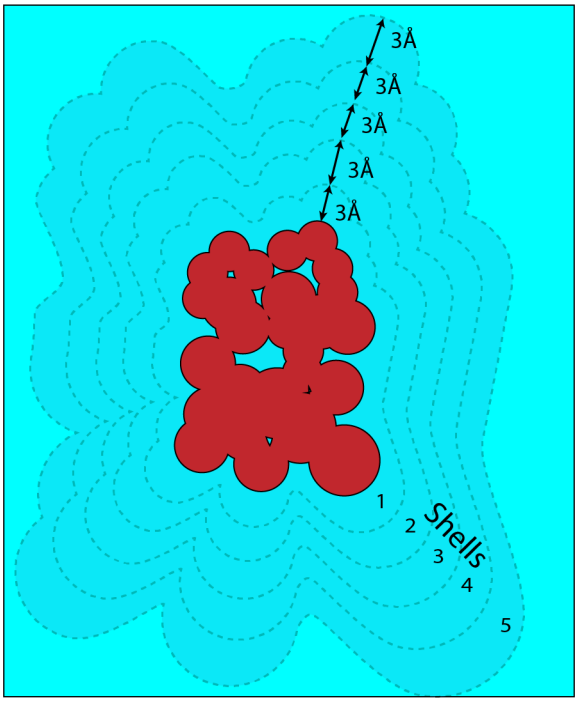


***Figure** 3.3.2: **Concentric hydration shells in the bulk volume.** An illustration of the division of the bulk of the solvent into 5 concentric shells is shown. Each shell is 3Å thick and placed in the manner shown. Each of these shells are treated as hydration sites in their own rights and their label (1 through 5) indicate their distance from the molecular surface of the protein (2NVH in this case).*

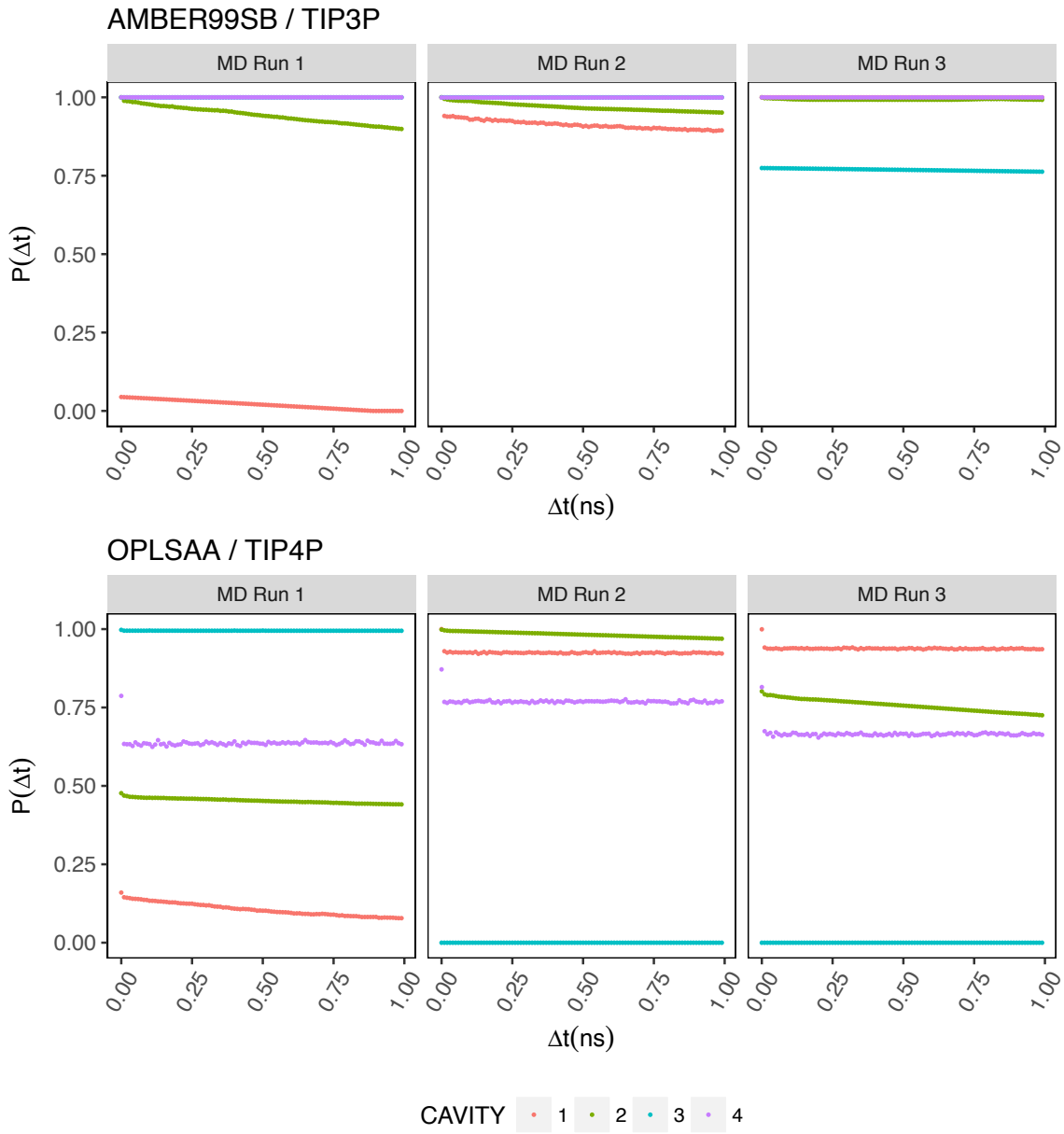


***Figure** 3.3.3: **Index correlation function (ICF) of the cavity waters**. ICF of cavity waters computed across the three MD runs using AMBER99SB/TIP3P (top) and OPLSAA/TIP4P (bottom) are shown for each of the 4 cavities. Cavity 5 is excluded because no water molecule was ever found to visit it.*

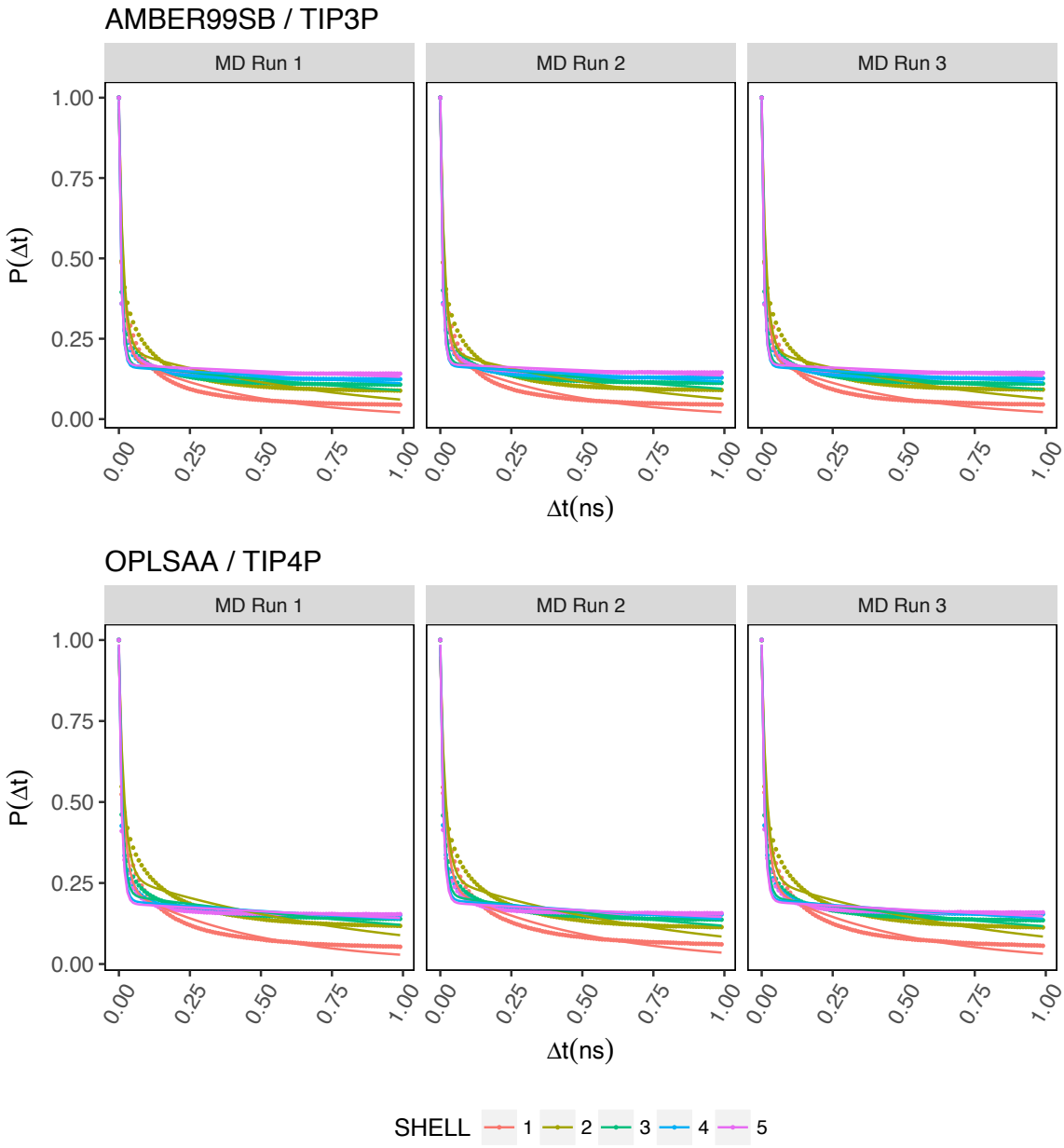

AMBER99SB / TIP3P
MD Run 1
MD Run 2
MD Run 3
P(Δt)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Δt(ns)
OPLSAA / TIP4P
MD Run 1
MD Run 2
MD Run 3
P(Δt)
Δt(ns)
SHELL
1
2
3
4
5

***Figure** 3.3.4: **Index correlation function (ICF) of the water in the bulk's hydration shell**. ICF of waters in the five different hydration shells in the bulk, computed across the three MD runs using AMBER99SB/TIP3P (top) and OPLSAA/TIP4P (bottom), are shown. The respective bi-exponential fits are also shown using lines of the same color as the respective data points.*

However, some interesting inferences can be made about the bulk water properties of the system. In addition to being remarkably different from the cavity waters, the bulk water was found to show distance-dependent attributes. After the bi-exponential fit to the ICF of the bulk hydration shells, computed using equation 20, a comparison of the fit parameters indicated that the residence times (τ_1, τ_2), occupancy (a_0) and the weight factor (w) vary as a function of distance from the surface of the protein. These variations provide insights into the effect of the local environment on bulk properties.

As is evident from **Figure** 3.3.5, the profile indicates that closer to the protein's surface, the short-term residence times (τ_1) are larger. The profile of the long-term residence time (τ_2) shows a contradictory behavior. However, the latter behavior must be interpreted carefully and in conjunction with the weight factor, which is indicative of the likelihood that a water might stay for prolonged durations. Though τ_2 tends to increase with distance from the surface of the protein, the weight factor decreases. Collectively it indicates that the likelihood of prolonged residence

times in the bulk are very low and with increasing distance from the surface, long term water molecules are highly unlikely.

Overall, a gradual transition of the water's temporal properties is evident from these analyses. The fact that the different force-field/water-model combinations do not introduce any noticeable difference in the profiles of these quantities suggests that these properties are intrinsic to the water as a chemical entity (**Figure** 3.3.3, **Figure** 3.3.4 and **Figure** 3.3.5). In essence, its local environment seems to be a critical factor in influencing its mean residence time.

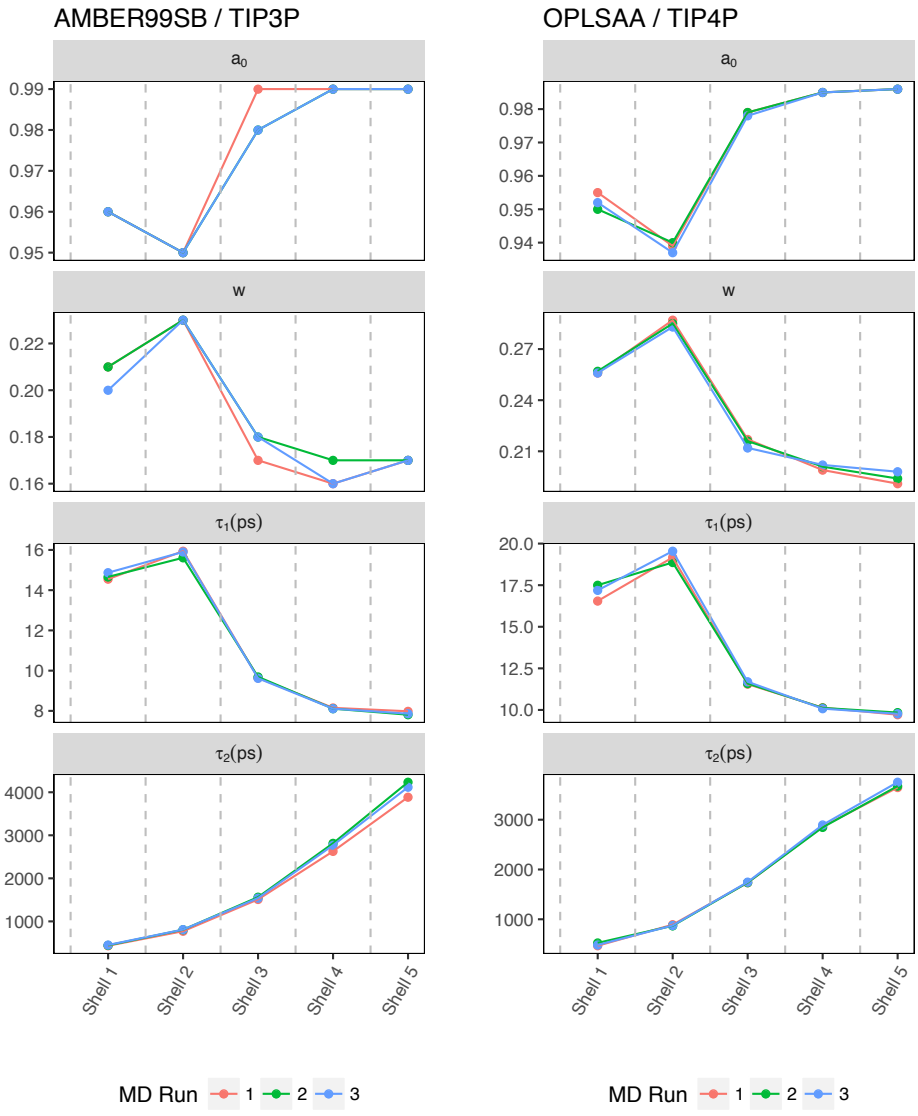


***Figure** 3.3.5: **Distance-dependence of the parameters of the bi-exponential fit to the ICF of the hydrations shells in the bulk**. The plots show the value of four parameters (labelled as: a_0, w, τ_1, τ_2) in each of the concentric hydration shells. Their values, obtained through simulations using AMBER99SB/TIP3P (left) and OPLSAA/TIP4P (right) combinations are shown.*

3.3.3 Cavity vs bulk water: Dipole rotational relaxation time

In this section , the differences in the spatial properties of the water in the bulk and in the cavities are analyzed. Like the above analyses of mean residence times, the gradient of bulk properties as a function of the distance from the surface of the protein is also examined.

The quantity of interest in this context is the orientational relaxation time of the water molecules and is derived using a 2nd order rotational autocorrelation function (RAF) of its dipole moment vector [120]. The functional form of the function is

$$C_2(\Delta t) = \langle P_2(\hat{u}(t_0 + \Delta t) \cdot \hat{u}(t_0)) \rangle_{t_0} \tag{22}$$

The function $P_2(x)$ is the 2nd order Legendre's polynomial given by:

$$P_2(x) = \frac{3x^2 - 1}{2} \tag{23}$$

Dipole moment's ability to "rotate" in a region, as a response to any external field, is a critical indicator of the local polarizability. In RAF, this ability to "rotate" in space is quantified by the average angle the total dipole moment vector sweeps in some time period Δt. Essentially, the argument to the $P_2(x)$ is the *cosine* of the angle

between the unit vector of the dipole moment at some time t_0 and a later time $t_0 + \Delta t$. For a fixed value of the time-gap or Δt, different starting or reference times, t_0, will feature different angles (and different cosines) whose average is of interest in this analysis.

To better understand its implications, it is important to understand how the average of the function looks like. Say there was a dipole moment vector that never underwent any rotation. In that case, the average angle for any time-gap is 0, which implies that the cosine is 1. Therefore, RAF will yield a value of 1 because

$$\langle P_2(x = 1)\rangle = 1 \tag{24}$$

On the other extreme is the case of a dipole moment which is freely rotating in space and therefore sampling all the angles in the range of 0 to $180°$ (or 0 to π radians, cosine +1 to -1). For that vector, the RAF will yield 0 since,

$$\langle P_2(x)\rangle = \frac{1}{2}\int_{-1}^{1} P_2(x)dx = 0 \tag{25}$$

All the intermediate behavior should lie within these limits.

The water molecules in the cavities and in the hydration shells were subjected to the above assessments. In addition to this, the data from equation 22 was fit using a mono-exponential curve to determine the time-scale of orientational relaxation[120]. For the lack of a better justification, bi-exponential fit wasn't used in this case. The fit function had the following form:

$$C_2(\Delta t) = a_0 e^{-\frac{\Delta t}{\tau_R}} \tag{26}$$

The coefficient a_0 is not occupancy in this definition but simply a dimensionless scaling factor present there to provide an optimal fit of the curve to the actual RAF profile. τ_R denotes the dipole orientational relaxation time. Essentially, it is indicative of the timescale required for a dipole moment vector to rotate in response to an external stimulus. A larger value will suggest that the dipole moment takes longer to respond and therefore, the medium in question has a lower polarizability. Respectively, a smaller value will indicate a more polarizable medium.

The results of the analyses are presented in the plots in **Figure** 3.3.6 and **Figure** 3.3.7.A clear visual distinction of the RAF profiles of the cavity and the bulk hydration shells is evident. As was observed with the ICF profiles in the section *Cavity vs bulk water: Mean residence time*, the cavity RAF have a slower decay rate

than the bulk hydration shells. The difference is in fact more stark than that of the ICF profiles. After fitting the observed data to the mono-exponential expression in equation 22, the values of two parameters - a_0 and τ_R were also determined. Of special interest is the latter, the rotational relaxation timescale. For the cavity waters, the value of τ_R was as low as 400ps and as high as 6000ps. It's also worth mentioning that the fit obtained with the mono-exponential expression showed poor fit quality and therefore the exact values are not very meaningful in an absolute sense. However, they provide a good relative estimate when the contrast with bulk water timescales are also considered. For the bulk hydration shells, the typical values of τ_R were below 5ps; nearly 1000 times smaller than that of the bulk. This is a vivid proof of the significant difference of the polarizabilities of the water molecules locked in protein cavities and those present in the bulk. This aligns well with the common understanding that water molecules occupying interstitial sites of the protein are involved in interactions with proteins and are stripped off their bulk-like properties.

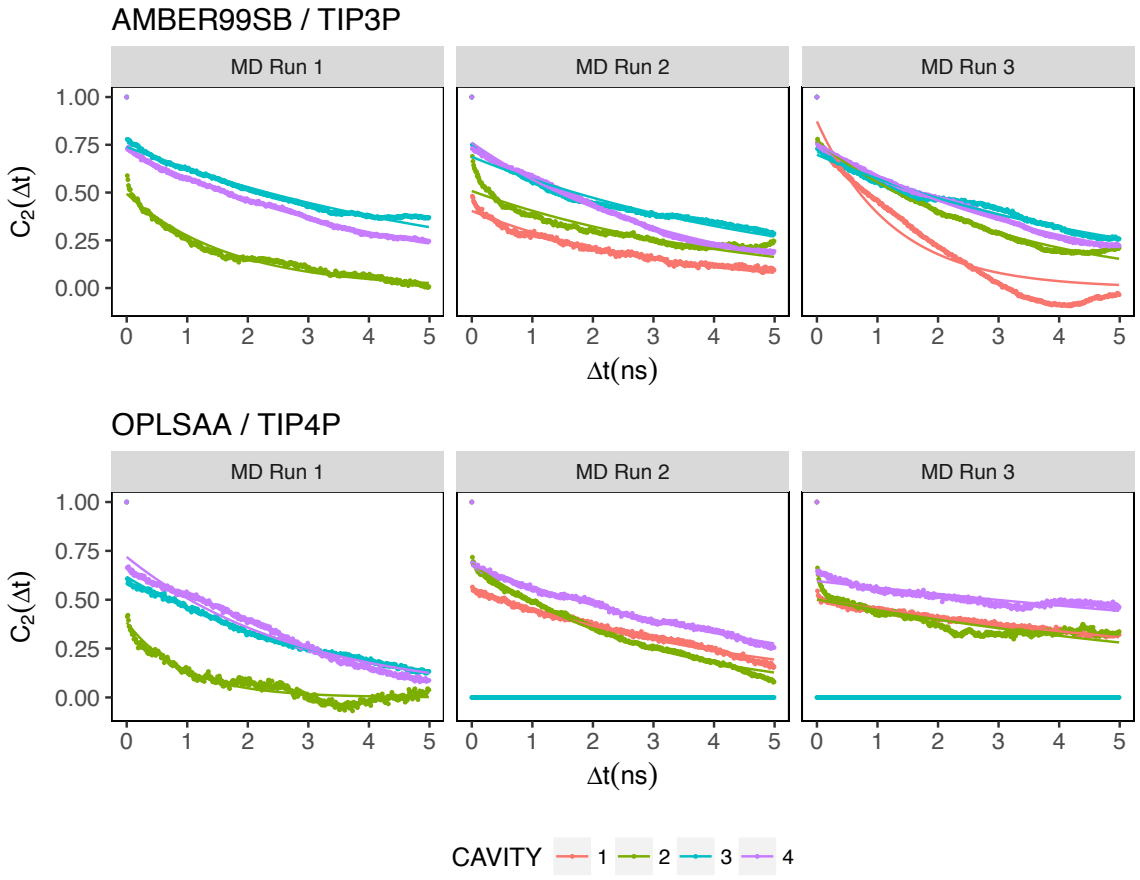


***Figure** 3.3.6: **Rotational auto-correlation function (RAF) of the cavity waters**. RAF of cavity waters computed across the three MD runs using AMBER99SB/TIP3P (top) and OPLSAA/TIP4P (bottom) are shown for each of the 4 cavities. Cavity 5 is excluded because no water molecule was ever found to visit it. For these plots, the solid lines denote the mono-exponential fit curve.*

In addition to the difference between the cavity and bulk waters, space-dependent profile of the rotational relaxation timescales was also observed across the bulk hydration shells. The variation of the fit parameters, a_0 and τ_R, as a function of the distance from the protein's surface is evident from the plots in **Figure** 3.3.8**.**

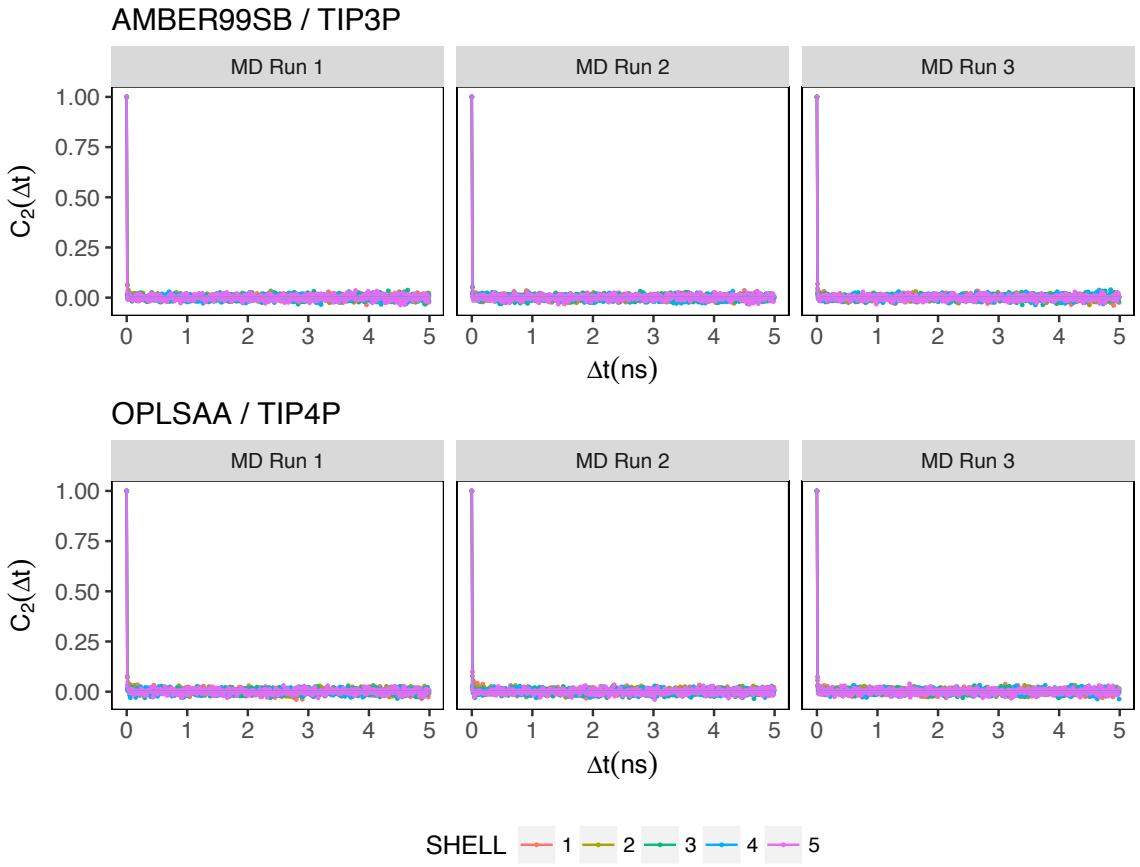


***Figure** 3.3.7: **Rotational auto-correlation function (RAF) of the water in the bulk's hydration shell**. RAF of waters in the five different hydration shells in the bulk, computed across the three MD runs using AMBER99SB/TIP3P (top) and OPLSAA/TIP4P (bottom), are shown. Though obscured, the respective mono-exponential fits are also shown using solid lines of the same color as the respective data points.*

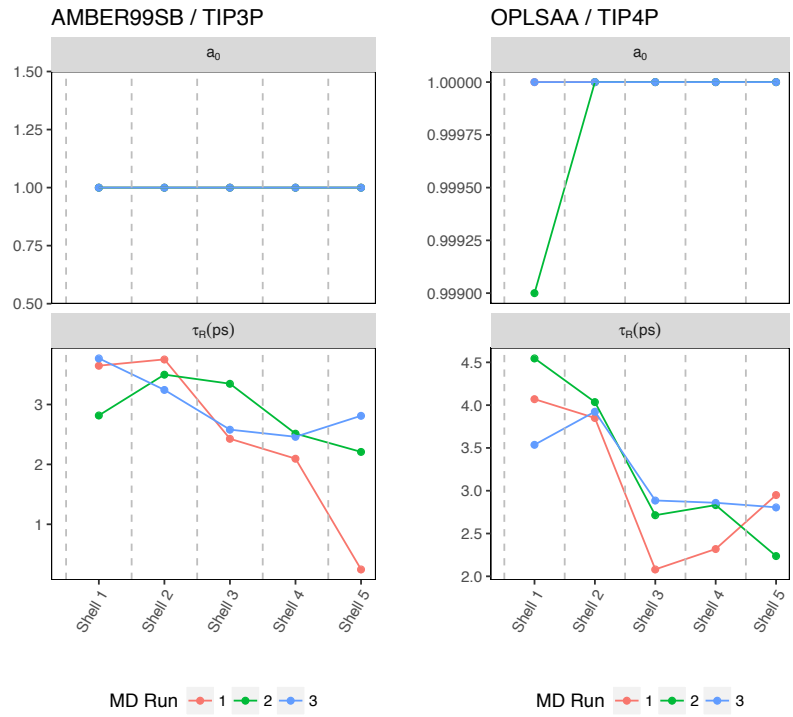


***Figure** 3.3.8: **Distance-dependence of the parameters of the mono-exponential fit to the RAF of the hydrations shells in the bulk**. The plots show the value of two parameters (labelled as: a_0, and τ_R) in each of the concentric hydration shells. Their values, obtained through simulations using AMBER99SB/TIP3P (left) and OPLSAA/TIP4P (right) combinations are shown. The typical values of thee quantities can be inferred from the scales of the y-axes of the plots.*

As one probes the water molecules farther away from the protein's surface, the value of τ_R tends to decrease. This means that rotational motions are restricted closer to the surface and there is more leeway farther away from it. Once again, this is not unexpected because water molecules engage in solute-solvent interactions, via H-bonds and salt-bridges when they are in the vicinity of the protein. Thereupon,

the rotations are only favorable to certain configurations. Farther out in the bulk, the degeneracy is higher and therefore orientational changes aren't very costly. These general inferences appear to be valid for either combinations of the force-fields and the water models and therefore reflect a property intrinsic of the water as a chemical entity.

Overall, there is a spatial variance of the dipole rotational relaxation time and in a general sense an inhomogeneity in its distribution as a property of the solvent. This inhomogeneity emanates from the differences in the local environments of the water molecules. Of major significance is the relevance of these inferences to the established relationship of the dipole orientational ability and the dielectric of a medium. Our inferences from explicit solvent MD simulations establish that local effects are critical in influencing the dielectric properties of a region.

3.4 Solvent exposure and dipole orientational relaxation timescales of protein residues

A similar set of calculations were also performed for the residues of the protein under investigation. The 2nd order rotational auto-correlation function (equation 22) was computed using the dipole moment information of each of the 153 residues and

using the mono-exponential fit (equation 26), their rotational orientational relaxation timescales were estimated. The objective of this exercise was to seek if solvent exposure of a residue influences its rotation timescale. This is expected to give an overview of the effect of the solute-solvent interactions from the perspective of a protein. The results of this analyses are shown in **Figure** 3.4.1 where the timescale τ_R is plotted against the solvent accessibility of the residues which were computed using NACCESS[119]. An inverse relationship is visible in these plots. For a guide to the eye, the data is fit using a curve which has an overall negative slope.

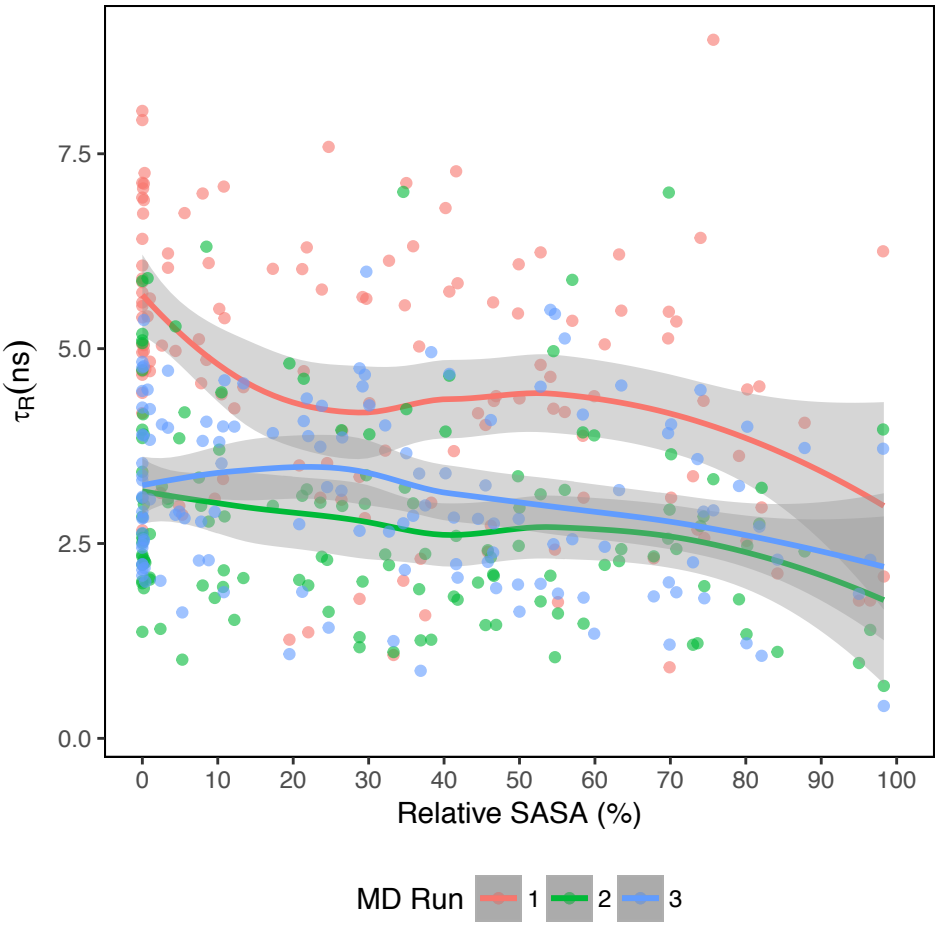


***Figure** 3.4.1: **Solvent exposure and dipole orientational relaxation timescales of protein residues.** For all the three MD runs, the dipole orientational relaxation timescales (τ_R) are plotted versus the relative solvent accessibility surface area (SASA) of the residues of the protein with PDB ID 2NVH.*

Upon a careful observation and comparison of the typical values of τ_R of the protein residues and bulk waters, one can notice the difference in scales or the order of magnitudes. Whereas the bulk water molecules exhibited timescales of in the ballpark of 1-5ps, with the ones close to the surface featuring a timescale of 5ps, the

protein residues exhibited typical values in the nanosecond regime. Thus, their relaxation is more restricted and this justifies lower dielectric values for the proteins. The fundamental reason is the structural constraints present in the protein emanating from its 3D structure. Protein residues are restrained by the backbone and other side-chain/side-chain and side-chain/backbone intramolecular interactions which do not typically permit low frequency, high amplitude motions in the nanosecond regime. Nevertheless, inferences made using MD trajectories are very helpful and can be corroborative to the design of the dielectric models.

By merging these inferences with the inferences regarding the environment-dependent polarizability of water, it can be reasoned that it is best if a dielectric distribution model of a solvated biomolecular system assigns lower dielectric values to the protein regions and higher dielectric values to the bulk by simultaneously maintaining a smooth and gradual gradient of the dielectric values across them. Farther away from the protein's atoms, the local dielectric must increase till it reaches the maximum value, which is that of the bulk solvent. The solute-solvent interface should be assigned an intermediate dielectric value which would reflect the combined effect of increased mobility of the protein atoms/groups at the interface and relatively restricted mobility of the solvent there. That the Gaussian-based

dielectric model exactly captures this gradient through its formalism (see section 2.1) has already been shown by Li *et. al.[86]* and Chakravorty *et. al.*[121]

3.5 Summary

In this chapter, several observations are presented and discussed to make a case that the local environments are critical to determining the properties of the solvent. By choosing and analyzing the trajectory of a protein with internal cavities, the differences in the tempo-spatial properties of the cavity and bulk water molecules are analyzed. The observations are connected to the dielectric properties that they may acquire and how their local environments can be influential in that regard. The restricted ability of the cavity waters to rotate as freely as the bulk waters is demonstrated. The variation of the rotational abilities of the bulk water molecules as a function of the distance from the protein's surface is also presented extensively. Simultaneously, the effect of solvent exposure on a residue's dielectric response (through rotational timescales) is also shown. All of the inferences are shown to indicate that inhomogeneous dielectric distribution best captures the conformation dynamics which are intrinsic to the solvated biomolecular systems. Through these

efforts, the validity of the conceptual basis of the Gaussian-based smooth dielectric distribution is strengthened.

4 USING THE GAUSSIAN-BASED DIELECTRIC MODEL TO REPRODUCE ENSEMBLE AVERAGE POLAR SOLVATION ENERGY OF A PROTEIN FROM A SINGLE CONFORMATION

Typically, the ensemble average polar component of solvation energy (ΔG_{solv}^{polar}) of a macromolecule is computed using molecular dynamics (MD) or Monte Carlo (MC) simulations to generate conformational ensemble and then single/rigid conformation solvation energy calculations are performed on each of those snapshots. The primary objective of this chapter is to demonstrate that the PB model using a Gaussian-based smooth dielectric function can reproduce the ensemble average (ΔG_{solv}^{polar}) of a protein from a single structure. It is shown that the Gaussian dielectric model can reproduce the ensemble average $\Delta G_{solv}^{polar}\left(\langle \Delta G_{solv}^{polar} \rangle\right)$ from an energy minimized structure of a protein. The best case, however, is when it is paired with an *in vacuo* minimized structure. In other minimization environments (implicit or explicit waters or crystal structure) the traditional two-dielectric model can still be selected with which the model produces correct solvation energies. The observations reported in this chapter reflect the model's ability to appropriately mimic the motion

of residues, especially those forming salt-bridges and how that is a key factor in deciding a dielectric model's ability to reproduce the ensemble average value of polar solvation free energy from a single structure. The contents of this chapter have been published previously[121] (copyright permission[2]).

4.1 Motivation

A routine protocol for computing the average solvation energy is to obtain a representative ensemble of structures (snapshots) by MD/MC simulations and then perform TI, FEP or Bennett Acceptance Ratio (BAR) calculations on each of the snapshots (while keeping each of them rigid)[122-124]. But these methods are extremely demanding of computational time and resources, since a typical ensemble may consist hundreds or thousands of snapshots. One can, alternatively, also subject these snapshots to PB modeling and obtain the corresponding polar solvation energy. The calculated polar solvation energies together with non-polar solvation energies

---

[2] Reprinted (adapted) with permission from (*J. Chem. Theory Comput.* 2018, 14, 2, 1020-1032). Copyright (2018) American Chemical Society.

delivered from surface area/volume of individual snapshots are expected to represent an experimentally measured solvation energy. Such an approach is an essential component of energy calculations employing molecular mechanics Poisson-Boltzmann surface area (MM/PBSA)[102] and molecular mechanics Generalized Born surface area (MM/GBSA)[103] methods. However, the bottleneck of MM/PBSA and MM/GBSA approaches is the generation of representative ensemble of structures, which is very expensive computationally, especially if applied for large scale modeling.

As an alternative to explicit modeling of conformational changes, one can also mimic the effect of these changes on the solvation energy via appropriate dielectric constant of the macromolecule. Dielectric distributions are known to affect the structure-energy relations via screening of the electrostatic interactions within the solute and between the solute and solvent[105, 106]. However, biological macromolecules are not rigid bodies and experimentally observable quantities are ensemble averaged. The traditional two-dielectric PB calculations cannot mimic these conformational changes within an ensemble because it uses two distinct but respectively uniform dielectric constants for the solute and water phase. This drawback has motivated the idea and usage of heterogeneous dielectric

distributions[104, 125-129]. They have been shown to yield better predictions for protein folding[78] and binding free energies[130] when benchmarked against experimental data while at the same time, they can also reveal the effects of mutations on these processes[131].

The objective, therefore, is to examine if the Gaussian-dielectric model, by virtue of its physically justified heterogeneity, can successfully capture the local effects of conformational changes on the solvation properties. The primary goal was to be able to render this average computationally, by incorporating the effects of the aforementioned dynamics while still preserving the time efficacy of the implicit solvent models, and thus to serve as a starting point for developing a fast and efficient single structure MM/PBSA method. **Figure** 4.1.1 provides an illustration on a cartoon plot to better represent the motivations behind the works reported in this chapter.

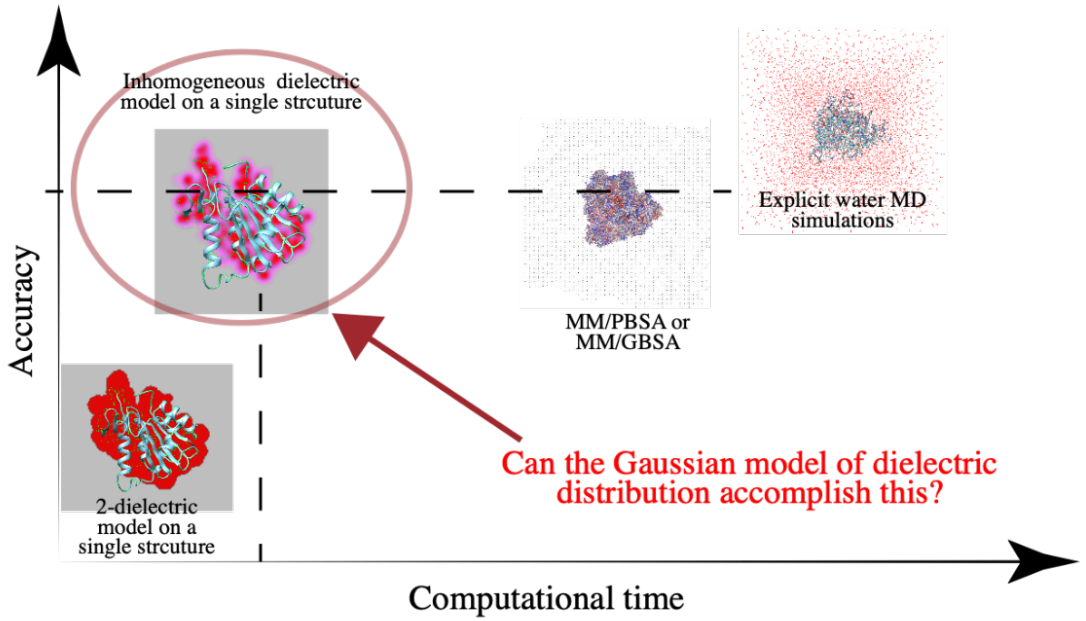


***Figure** 4.1.1: **Can the Gaussian-based dielectric model reproduce ensemble average properties from a single structure**? This illustration provides a visual description of the question being asked in this chapter. Essentially, it highlights the "gap" that, if filled, will offer a promising and faster alternative to the conventionally used methods of computing polar components of solvation free energy while still retaining the physical meaningfulness.*

4.2 Methods

4.2.1 Set of representative proteins

Protein structures for this work were obtained from the Protein Data Bank (PDB) [132]. To obtain a dataset of reasonable size that can be managed in parallel with extensive MD simulations data, the resolution of the structures was limited

between 0.8 and 0.99 Å with at most 200 amino acids. Besides we required the structures to be monomeric. The proteins retrieved were required to have at most 30% sequence similarity. In addition, it was ensured that [3]these structures did not contain a ligand or modified residue. This search yielded 74 globular proteins from the PDB as of April 29, 2017[3].

4.2.2 Structure preparation

The protein structures were prepared for MD simulations using GROMACS v5.0.5[116] with atomic parameters of the AMBER99SSB[133] force field. All the titratable residues were kept in their charged states. To build the explicit water solvated systems, these structures were solvated using TIP3P water molecules[113] and ions were added wherever neutralization was needed.

---

[3] PDB ID of 74 proteins used: 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

4.2.3 Energy minimization

The explicit water solvated systems were subjected to 10,000 steps of steepest descent (SD) energy minimization using GROMACS v5.0.5[116]. The heavy atoms were harmonically restrained to their original positions with a force of 1000 kJ/mol/nm while everything else in the system was set free to move.

Two other minimizations, involving only the protein structures, were also carried out. These were performed *in vacuo* and Generalized Born Implicit Solvent (GBIS)[134] environments. Only 5000 SD steps were used since the system size was drastically smaller than the explicit water systems. For both of these cases, cutoffs for the non-bonded interactions were lifted but all the heavy atom harmonic restraints were retained. For GBIS minimization, the external dielectric was set at 80.0 (emulating water environment) and that for *in vacuo* was 1.0. The parameters for minimizations are provided in **Appendix A.1.**

4.2.4 MD simulations

Post energy minimization, only the explicit water solvated systems were subjected to 3 independent MD simulations for 20ns each (with different initial

atomic velocities) to allow versatility in the resulting ensemble of structures. Prior to production phase of the MD, they were equilibrated under constant volume-temperature (NVT) conditions for 500ps (with heavy atoms harmonically restrained) followed by 2000ps (=2ns) of constant pressure-temperature (NPT) equilibration at 300K temperature and 1 atm pressure (with the same restraints). In the 20ns MD that followed, the restraints were lifted. The initial 10ns of was discarded as they were considered to not have equilibrated yet. The structures for analysis were sampled from the last 10ns at every 10ps. This yielded 1000 snapshots per MD run, all of which were subjected to PB based solvation free energy calculation after stripping off the explicit water molecules (and ions for neutralization wherever present). For all the equilibrations and MD, particle mesh Ewald (PME)[114] based electrostatic calculations were invoked in conjunction with periodic boundary conditions. The parameter configurations for equilibration and MD are provided in the **Appendix A.2.**

4.2.5 Ensemble average polar solvation energy from PB vs alchemical MD methods

Three independent MD simulations in explicit water rendered 3000 thermodynamically weighted configurations per protein[135], which we shall refer to

as its ensemble. For each member of the ensemble, the polar component of the solvation free energy (ΔG_{polar}^{solv}) was computed to obtain ensemble average[103, 136, 137]. The method used for this purpose requires an explanation, which is provided below.

Ideally to compute the ΔG_{polar}^{solv} for a molecule, alchemical free energy calculation methods in explicit solvent setups are preferred. Thermodynamic integration based molecular dynamics (TI-MD) is an example of such a method. Authors of Ref.[138, 139] have calculated the polar component of the solvation free energy of 19 proteins using TI-MD. These values were computed by fixing the protein's structure in space and coupling the partial atomic charges to a coupling parameter 'λ' which was varied from 0 to 1. As the protein's electrostatic properties traverse a set of alchemical intermediate states due to λ, the energy cost associated with it in the presence of explicit water molecules is calculated and eventually summed up to render the total solvation energy (polar + non-polar component). Should the protein structure be allowed to move, the resultant energy cost would include effects of protein molecular mechanical energies which cannot be resolved to get the exclusive polar solvation energy. This procedure can be iteratively applied

on each "snapshot" of an ensemble (from MD/ Monte Carlo) to determine the ensemble solvation energy.

Delphi[70, 71] calculates and outputs the polar component of solvation energy of a "snapshot", which is termed "corrected reaction field energy". To examine if the PB-based calculations provide similar or identical polar solvation energies as TI-MD, the ΔG_{polar}^{solv} for the 19 proteins used by the authors of the aforementioned work[138, 139] were computed while preserving the structural coordinates, charges and radii. A *scale* of 2.0 grids/Å, a '*perfil*' of 70 and the traditional 2-dielectric method was used while setting the protein internal dielectric to 1.0 and solvent dielectric to 80.0. The ΔG_{polar}^{solv} values from both methods are compared **(Figure** 4.2.1**)**. It is evident that *Delphi* delivers ΔG_{solv}^{polar} almost precisely identical to that obtained by TI-MD in explicit water (correlation = 0.99 and RMSD = 17.93 kcal/mol). This reinforces the claim that, provided the structures are rigid, PB calculations with *Delphi* (with protein internal dielectric=1 and solvent dielectric = 80) can deliver ΔG_{polar}^{solv} that would otherwise require a much longer TI-MD runs. Therefore, by using *Delphi* with the above protocols to calculate ΔG_{polar}^{solv} for each "snapshot", the ensemble polar solvation energy was calculated in a manageable time.

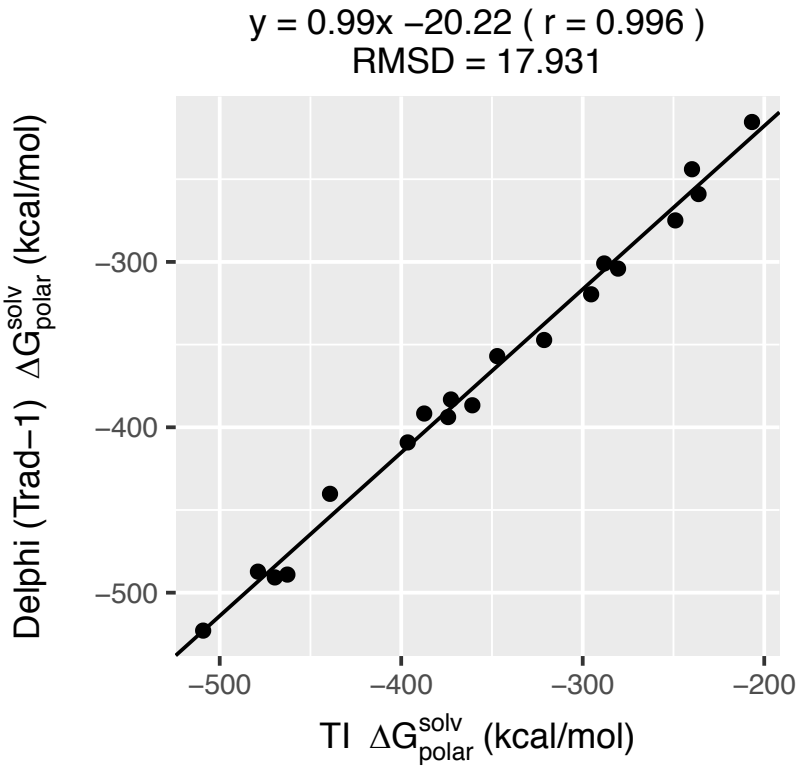

y = 0.99x −20.22 (r = 0.996)
RMSD = 17.931
Delphi (Trad−1) ΔG solv polar (kcal/mol)
−300
−400
−500
−500
−400
−300
−200
TI ΔG solv polar (kcal/mol)

***Figure** 4.2.1: **Explicit solvent thermodynamic integration vs Implicit solvent PBE**: The comparison of the polar solvation energies of 19 net-neutral proteins obtained from explicit solvent thermodynamic integration (TI) simulations and implicit solvent Poisson-Boltzmann (PB) calculations using the traditional 2-dielectric model with Delphi is shown. For both the cases, the protein structures were kept rigid. The TI simulations were performed by the authors of Ref[138, 139]. The Pearson correlation (r) and RMSD (in kcal/mol) of the comparison are also mentioned.*

4.2.6 Polar Solvation energy of energy minimized structures

For each of the protein energy minimized (EM) structures (minimized in 3 different environments), ΔG_{polar}^{solv} was computed using the traditional 2-dielectric model as well the Gaussian-based smooth dielectric model[63]. With a *scale* of 2.0 grids/Å, a '*perfil*' of 70 was used for the former and that of 50 was used for the latter. For the Gaussian-model, a '*sigma*' = 0.93 was applied. For all the calculations, the probe radius was set at 1.4 Å with zero electrolyte concentration and the external dielectric constant was set to 80 (emulating water environment). The boundary potentials were determined using the dipole method.

In the rest of the chapter, all the PB calculations performed using the traditional 2-dielectric method will carry a label 'TRAD-x' and that for the Gaussian-based smooth dielectric method will carry a label 'GAUSS-x'. 'x' in these labels

indicate the protein internal dielectric constant. For instance, 'TRAD-1' and 'GAUSS-1' will identify as the corresponding methods with protein internal dielectric set at 1.

4.2.7 Modified Gaussian-based smooth dielectric model in Delphi

A modification was incorporated in the algorithm that computes the polar solvation energy using the Gaussian-based smooth dielectric function in *Delphi*. The key idea of the Gaussian-based approach is that a strict surface doesn't separate the solute interior from the external medium, as is assumed in the traditional 2-dielectric models. The mathematical details are presented in equations 14-16. This delivers a position-dependent dielectric distribution ($\epsilon(\vec{r})$) when the solute is present in a medium of dielectric constant 'ϵ_{out}'.

In the original implementation of the Gaussian-based model in *Delphi*, the polar component of solvation energy is calculated by taking the difference of the grid energies obtained from modeling a solute in (i) external solvent medium (medium-1) and (ii) medium with dielectric constant same as the internal dielectric constant (medium-2). However, this requires building a surface between solute and medium-2, a surface that does not conceptually exist in a surface-free approach of the

Gaussian-based model and therefore is artificially drawn when calculating grid energies in medium-2. For that a user-specified dielectric value is used to delineate the iso-dielectric surface.

In this work, two-fold modifications were made. First, the "surface" was drawn based on a user-defined probability (ρ_{SF}) instead of a dielectric value. This is done to fix the solute "volume" regardless of the ϵ_{ref} value of the internal reference dielectric constant since different ϵ_{out} can influence the position of the iso-dielectric surface but not that of a iso-probability surface. In this work, we used the atomic density value of 0.759, which corresponds to a dielectric of 20 for $\epsilon_{ref} = 1$. Second, a smoother transition from this "surface" to the external region for medium-2 was made using an exponential function. By fixing the medium-2 as vacuum ($\epsilon_{out} = 1$), the smoothing term secures the surface-less approach of the Gaussian-based model to a great extent. This results in the following dielectric distribution when the external medium is vacuum:

$$\begin{aligned} &if\ \rho_{in}(\vec{r}) \geq\ \rho_{SF}\colon \varepsilon'(\vec{r}) =\ \varepsilon(\vec{r}) \\ &\quad if\ \rho_{in}(\vec{r}) <\ \rho_{SF}\colon \varepsilon'(\vec{r}) \\ &\qquad\qquad =\ 1 + (\varepsilon(\vec{r}) -\ 1)e^{-(\rho_{in}(\vec{r}) - \rho_{SF})(\varepsilon_{ref} - \varepsilon_{out})} \end{aligned} \tag{27}$$

Here, $\epsilon(\vec{r})$ is the dielectric value of a 3D point when the solute is present in vacuum and $\epsilon(\vec{r})$ is the dielectric value assigned to that point when the protein's presence in solvent was modelled. 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 ρ_{SF}. The above schematic is shown in **Appendix A.3** for an arbitrary placement of atoms along a single dimension.

4.3 Results and Discussion

4.3.1 Ensemble average from Energy minimized structures

The ΔG_{polar}^{solv} of the protein structures obtained after minimization in three different environments - *in vacuo*, Generalized Born Implicit Solvent (GBIS) and explicit water (TIP3P), were compared with the ensemble average polar solvation energy, $\langle \Delta G_{polar}^{solv} \rangle$ (procedure outlined in the Methods). For the sake of completeness, the crystal structure of the proteins was also subjected to this comparison. The crystal structures were, first protonated and then energy minimized while its heavy atoms were heavily restrained (force constant of 1e6 KJ mol-1nm-2) to keep the

backbone atoms positions unchanged. We shall refer to these structures as optimized crystal structures hereafter.

The results of these comparisons are shown in terms of probability distribution of the differences of ensemble average $\langle \Delta G_{polar}^{solv} \rangle$ and the ΔG_{polar}^{solv} of the EM and optimized crystal structures (**Figure** 4.3.1). In the figure, the difference $\langle \Delta G_{polar}^{solv} \rangle - \Delta G_{polar}^{solv}(EM)$, extends to both negative and positive values. Since both $\Delta G_{polar}^{solv}(EM)$ and $\langle \Delta G_{polar}^{solv} \rangle$ are negative, it is vital to understand how these differences should be interpreted. A negative difference implies $\langle \Delta G_{polar}^{solv} \rangle < \Delta G_{polar}^{solv}(EM)$, depicting that the ensemble average is more negative than the polar solvation energy of the EM structure. In terms of magnitudes, the EM structure ΔG_{polar}^{solv} is smaller than the $\langle \Delta G_{polar}^{solv} \rangle$ (underestimation). On the other hand, a positive difference implies $\langle \Delta G_{polar}^{solv} \rangle > \Delta G_{polar}^{solv}(EM)$, i.e. the ensemble average is less negative than the corresponding polar solvation energy form the EM structure. Magnitude wise, the EM structure ΔG_{polar}^{solv} is larger than the $\langle \Delta G_{polar}^{solv} \rangle$ (overestimation). Therefore, if $\langle \Delta G_{polar}^{solv} \rangle - \Delta G_{polar}^{solv}(EM) \approx 0$, such a case successfully reproduces the ensemble average using a EM structure alone. With this, we now turn to describing the trends observed in Figure 4.3.1.

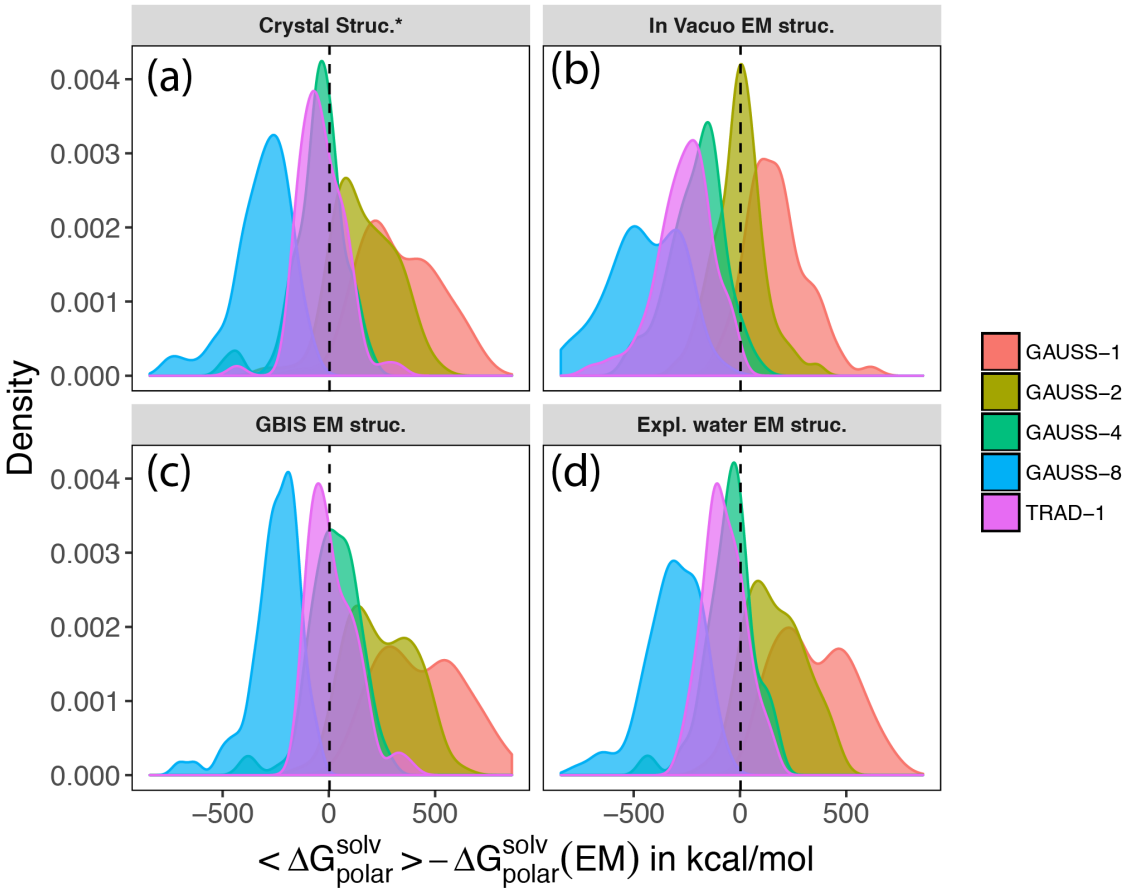


Figure 4.3.1: ***Performance of the Gaussian and the traditional 2-dielectric model in predicting the ensemble average polar solvation energy.*** *The Figure shows the density distribution of the difference,* $\langle \Delta G^{solv}_{polar} \rangle - \Delta G^{solv}_{polar}(EM)$, *obtained when Gaussian or traditional dielectric models are used on (a) crystal (aka. Xtal) structure (* added protons are optimized) and structures minimized (b) In Vacuo (c) in GBIS and (d) in explicit solvent (TIP3P). The labels 'TRAD-x' and 'GAUSS-x' indicate the traditional 2-dielectric and Gaussian-based smooth dielectric distributions, respectively. 'x' is the protein's internal dielectric value. The dashed vertical line is at the zero mark in each plot.*

Both, Gaussian-based (GAUSS) and traditional (TRAD) dielectric models were used with the optimized crystal and EM structures to compare with $\langle \Delta G_{polar}^{solv} \rangle$. For the former, values of 1, 2, 4 and 8 were used as internal reference dielectric constant. For the latter, only a single value (=1) was used because values larger than 1 resulted in highly underestimated ΔG_{polar}^{solv} with respect to the ensemble averaged $\langle \Delta G_{polar}^{solv} \rangle$. In Figure 4.3.1, a visual inspection reveals that the traditional dielectric model (TRAD-1) has a very similar degree of agreement with the $\langle \Delta G_{polar}^{solv} \rangle$ when paired with the optimized crystal structure Figure 4.3.1(a) and structures optimized in solvent (Figure 4.3.1(c, d)). With the *in vacuo* optimized structure, the trend is conspicuously different (Figure 4.3.1(b)). Quantitatively, when expressed in terms of the mean relative unsigned error, the best agreement is attained by the GBIS minimized structures, followed by the optimized crystal structure and structure minimized in explicit solvent (see Table 4.1).

Table 4.1: Average relative error and average absolute error from the ensemble average polar solvation energy of that from the optimized crystal and energy minimized structures.

Minimization Environment	Dielectric distribution model				
	TRAD-1	GAUSS-1	GAUSS-2	GAUSS-4	GAUSS-8
Crystal Structure*	5.59% [a] (90.90) [b]	19.34% (338.14)	10.35% (180.00)	5.31% (94.06)	18.09% (312.99)
In Vacuo	15.26% (262.13)	10.55% (179.48)	5.13% (85.01)	11.52% (206.69)	25.76% (449.86)
GBIS	5.14 % (85.71)	24.82 % (432.90)	14.41% (248.53)	5.16% (92.97)	14.50% (250.07)
Explicit Water (TIP3P)	6.21% (101.12)	20.07% (348.26)	10.16% (174.61)	5.30% (92.72)	18.18% (315.74)

*After optimizing the added hydrogens while restraining the heavy atoms in the crystal structure with a force constant of 1e6 KJ $mol^{-1}nm^{-2}$.

[a] Mean relative unsigned error

[b] In the parentheses, average absolute error (in kcal mol^{-1}).

It can, therefore, be tempting to use the GBIS minimized structure of a protein with the traditional model (internal $\epsilon = 1$) to obtain its ensemble $\langle \Delta G^{solv}_{polar} \rangle$. However, it must be done with caution. This is because the low relative mean error value is a statistical result and when it comes to individual proteins, as the plots suggest, some of them feature an underestimation of $\langle \Delta G^{solv}_{polar} \rangle$ ($\langle \Delta G^{solv}_{polar} \rangle - \Delta G^{solv}_{polar}(EM) < 0$; left of the black lines in Figure 4.3.1. To attain a better agreement, these cases demand that the internal ϵ be less than 1. Such a modification is

physically unreasonable. In fact, such an underestimation is true for the majority of proteins regardless of the environment of optimization (Table 4.2).

At the same time, the Gaussian-based dielectric model reveals a better agreement with the ensemble $\langle \Delta G_{polar}^{solv} \rangle$. This is also inferable visually from the figure, owing to the close placement of the peaks of the error distribution plots to the zero line. Unlike the traditional method, the Gaussian-based model offers a good match regardless of the minimization protocol. Quantitatively, the mean relative unsigned error varies depending on the ϵ_{ref} value used for a particular Gaussian-based model but there is always a case for all of the optimization environments where the mean relative unsigned error $\approx 5\%$ (Table 4.1). For instance, GAUSS-4 has an error comparable to and better than what the TRAD-1 incurs for the optimized crystal structure and structures optimized in solvent. GAUSS-2 with *in vacuo* minimized structures, moreover, not only offers a better agreement with $\langle \Delta G_{polar}^{solv} \rangle$ than the TRAD-1 model but it offers the best agreement amongst all the cases (lowest mean relative unsigned error; 5.13%). Furthermore, with minor adjustments of the input parameters for the Gaussian-based model in *Delphi* (see *Delphi* Manual[1]), one can tune the degree of agreement. This circumvents the problem of having to use unreasonable dielectric values for proteins, unlike the traditional method. Therefore,

the *in vacuo* minimized structure, when paired with the Gaussian-based dielectric model, can offer the best approximation to the ensemble average polar solvation energy.

Table 4.2: The percentage of cases (out of the 74 proteins) where the difference in the ensemble average polar solvation energy and polar solvation energy of optimized crystal and EM structure obtained using TRAD-1 dielectric method is negative. These cases would require decreasing the protein internal dielectric below 1 to correct for the error incurred by the TRAD-1 model, which is physically invalid.

Minimization Environment	% cases where $\left\|\langle \Delta G^{solv}_{polar} \rangle\right\| > \left\|\Delta G^{solv}_{polar}(EM)\right\|$
Crystal Structure	66.22 %
In Vacuo	100.00 %
GBIS	54.05%
Explicit Water (TIP3P)	78.34 %

Before delving into extensive analyses of the factors that influence the performance of the dielectric models, it is important that we address some of these trends observed for the traditional and Gaussian-based dielectric models in detail.

Observations in Figure 4.3.1 indicates differences in the behavior of the traditional dielectric model ($\epsilon_{in} = 1; \epsilon_{out} = 80$) when used with differently optimized structures. From the traditional model, the ΔG^{solv}_{polar} of the optimized crystal or

solvent-minimized structures is in good agreement with the $\langle \Delta G_{polar}^{solv} \rangle$ but that from the *in vacuo* EM structure is significantly underestimated. This can be understood as follows.

Upon a pairwise comparison of ΔG_{polar}^{solv} from TRAD-1 model, the following trend was observed:

$$\begin{aligned} &\Delta G_{polar}^{solv}(In\ Vacuo) \\ &\quad > \left(\Delta G_{polar}^{solv}(Xtal) \approx\ \Delta G_{polar}^{solv}(TIP3P)\right. \\ &\quad >\ \left.\Delta G_{polar}^{solv}(GBIS)\right) \approx \langle \Delta G_{polar}^{solv} \rangle \end{aligned} \tag{28}$$

When these comparisons were extended to protein coulombic energies, the following was seen (protein dielectric = 1.0).

$$\begin{aligned} &U_{coul}(In\ Vacuo) \\ &\quad < \left(U_{coul}(Xtal) \approx\ U_{coul}(TIP3P)\right. \\ &\quad <\ \left.U_{coul}(GBIS)\right) \approx \langle U_{coul} \rangle \end{aligned} \tag{29}$$

A clear reversal of the trend in Equation 28 is seen in Equation 29. Furthermore, **Figure** A. 4 illustrates this comparison qualitatively as well as quantitatively. One can notice the differences of U_{coul} and ΔG_{polar}^{solv} of solvent-

minimized (and optimized crystal) structures calculated w.r.t the *in vacuo* minimized structure. The opposite trends of comparison for these energy terms are evident. This is because a molecular structure with a high negative value of coulombic energy almost certainly contains oppositely charged particles placed more closely than in a structure with a less negative U_{coul}. The former stabilizes the packing in the gas phase, but reduces interactions with water. This loss in favorability towards solvation comes from the closely placed charges of opposite polarities forming a very small dipole. The smaller dipole consequently "annihilates" the atomic charges, thus compromising the favorable electrostatic interaction that could have existed with the polar solvent. As a result, the solvation is unfavorable relative to a configuration with a higher (less negative) U_{coul}. Since the *in vacuo* EM structures are likely to have oppositely charged atoms (or residues) placed more closely due primarily to the absence of any de-solvation, they feature more negative U_{coul} and therefore, relatively less favorable ΔG_{polar}^{solv}. The other configurations incur the effects of the solvent and consequently have a less negative U_{coul} but a more favorable ΔG_{polar}^{solv}. This validates the good agreement that structures minimized in solvent have with the ensemble average, in terms of ΔG_{polar}^{solv} computed using traditional dielectric

model. This is because the ensemble comprises configurations generated in an explicit solvent environment (see section **4.2**).

The inherent heterogeneity of the dielectric distribution underlying the Gaussian-based model complicates the above analysis provided for the traditional method. Not only do the formulations for the coulombic energy become non-trivial, it is practically difficult to exactly pinpoint a surface that segregates the solute region from the solvent due to its surface-free nature[63]. This precludes a simple interpretation of the trends of ΔG_{polar}^{solv} obtained from the Gaussian model. Nonetheless, the Gaussian model preserves the general trend of the effects of dielectric constant on the ΔG_{polar}^{solv}, i.e. increasing the solute dielectric decreases the latter's absolute value. This is apparent from the density plots in **Figure** 4.3.1 where increasing the ϵ_{ref} of the Gaussian model shifts the peak to the left of the zero-mark, indicating that the deviation from the ensemble $\langle \Delta G_{polar}^{solv} \rangle$ increases, or that the ΔG_{polar}^{solv} becomes less negative. These trends are separately depicted in the **Figure** A. 5, that compare how the solute internal dielectric affects ΔG_{polar}^{solv} for the two dielectric models. This inverse relation of the solute dielectric and ΔG_{polar}^{solv} prevails in systems as simple as a dipole embedded in a spherical cavity where the increase

of the cavity dielectric decreases ΔG_{polar}^{solv} value. Underlying this relationship is the fact that an increased solute dielectric increases the screening of the interaction of solute dipoles with water phase, thus making solvation less favorable.

Figure 4.3.1 also indicates that using the same value for GAUSS's ϵ_{ref} and TRAD's ϵ_{in} yields a more negative value for the former. This is because the ΔG_{polar}^{solv} calculated with Gaussian-based smooth dielectric model (as implemented in *Delphi*) depends not only on the reference value of internal dielectric constant (ϵ_{ref}), but also on the "surface" that separates the solute from the external medium. The numerical demarcation of this "surface", drawn based on a cut-off of effective dielectric value (iso-dielectric surface)[63] or effective atomic density (iso-density surface, ρ_{SF}) results in solute-solvent (solute-vacuum) interface being placed, in some regions, slightly inside the traditional molecule surface. This decreases the effective size of the solute thus making the ΔG_{polar}^{solv} more negative than what the traditional 2-dielectric model would deliver.

4.3.2 Role of salt-bridges (SBs)in the energy minimized structures

So far, the results have indicated that a Gaussian-based smooth dielectric distribution (GAUSS-2) in conjunction with *in vacuo* minimized structure reproduces

the ensemble average, $\langle \Delta G^{solv}_{polar} \rangle$, with smallest mean relative unsigned error (**Table** 4.1). They have also indicated that ΔG^{solv}_{polar} for optimized crystal or EM structures obtained with different dielectric models exhibit different but reasonably good agreements when compared with ensemble $\langle \Delta G^{solv}_{polar} \rangle$. To determine the causes for these differences, several structural properties of the minimized structures were tested and compared.

It is well known that constant breaking and forming of salt bridges is a salient feature of protein dynamics[140-142] and their dynamics affects the dielectric distribution of the protein interior[127, 143]. For our purposes, the fluctuation between the closed/open forms of a SB is quantified in terms of occupancy. Occupancy is defined as the percent of the 3000 configurations (of the host protein) wherein the SBs were closed (O-N distance < 3.4 Å). Therefore, a SB with a 100% occupancy is never found to be broken in the ensemble while one with 0% occupancy is only identified in the minimized structure but never in the ensemble. Anything in between should be interpreted likewise. That the SB pairs identified across all the 74 proteins featured fluctuations, is evident from the histogram of occupancies in **Figure** A. 4 in **Appendix A.5,** where occupancies pervade all the values from 0-100%**.**

The results suggest that the population of the charged/titratable residues forming SBs is a clear cause that differentiates the abilities of minimized structures to reproduce the ensemble averages. The comparison of the population of salt bridges (in closed conformation) after minimization shows that the *in vacuo* protein EM structures have a higher number of these than the EM structures from the other two environments (**Figure** 4.3.2(a)). This has been further demonstrated by computing relative number of SBs using the number in the corresponding EM structure from explicit water environment for normalization (**Figure** 4.3.2(b)). From that, the *in vacuo* structures clearly exhibit a high population of salt bridges while the number of salt bridges in GBIS based minimized structures are slightly larger than the explicit water ones. In fact, more than 90% of the *in vacuo* structures have more SBs than the corresponding explicit water based EM structures. This indicates that incorporating solvent effects in any form (implicit or explicit) can have a similar influence on the salt bridge formation. Such an influence can be ascribed to the screening of coulombic forces due to higher solvent dielectric and the desolvation energy due to partial burial of the SB forming titratable groups when they form a salt bridge. At the same time, the absence of these effects in vacuum allows the residue pairs to orient their side chains in a manner that would allow them to stay

bonded via a stable SB (a closed salt-bridge). This argument also seconds the trend of coulombic energy of various configurations of a protein described in the preceding section.

The significance of the number of charged residues and their ability to form salt bridges becomes more prominent upon examining how other structural features such as structural backbone RMSD and intra-protein hydrogen bond network vary after minimization in different environments. Energy minimization protocol is not expected to cause a significant change of a protein's conformation, especially the backbone conformation, but it may result in different hydrogen positions. To verify, the structural RMSD of the backbone and number of intra-protein hydrogen bonds (hydrogen bonds within the atoms of a protein) after minimization, with respect to the corresponding crystal structure, were calculated. The comparison for structural RMSD is shown in **Figure** 4.3.2(c). It is noticeable that large backbone changes did not occur post minimization regardless of the environment. Essentially, the backbone atomic positions were preserved. The RMSDs in all the cases were less than 0.5 Å. In the same way, **Figure** 4.3.2(d) illustrates the comparison for the number of intra-protein hydrogen bonds. It is evident that all of the three environments yielded similar numbers after minimization. The above analysis indicates that EM in

different environments results is very similar backbone structures and intra-protein hydrogen bonds and therefore, cannot be the reason for the differences in the polar solvation energies. However, the number of closed SBs in the EM structures bear a qualitative correlation to the differences in their polar solvation energies.

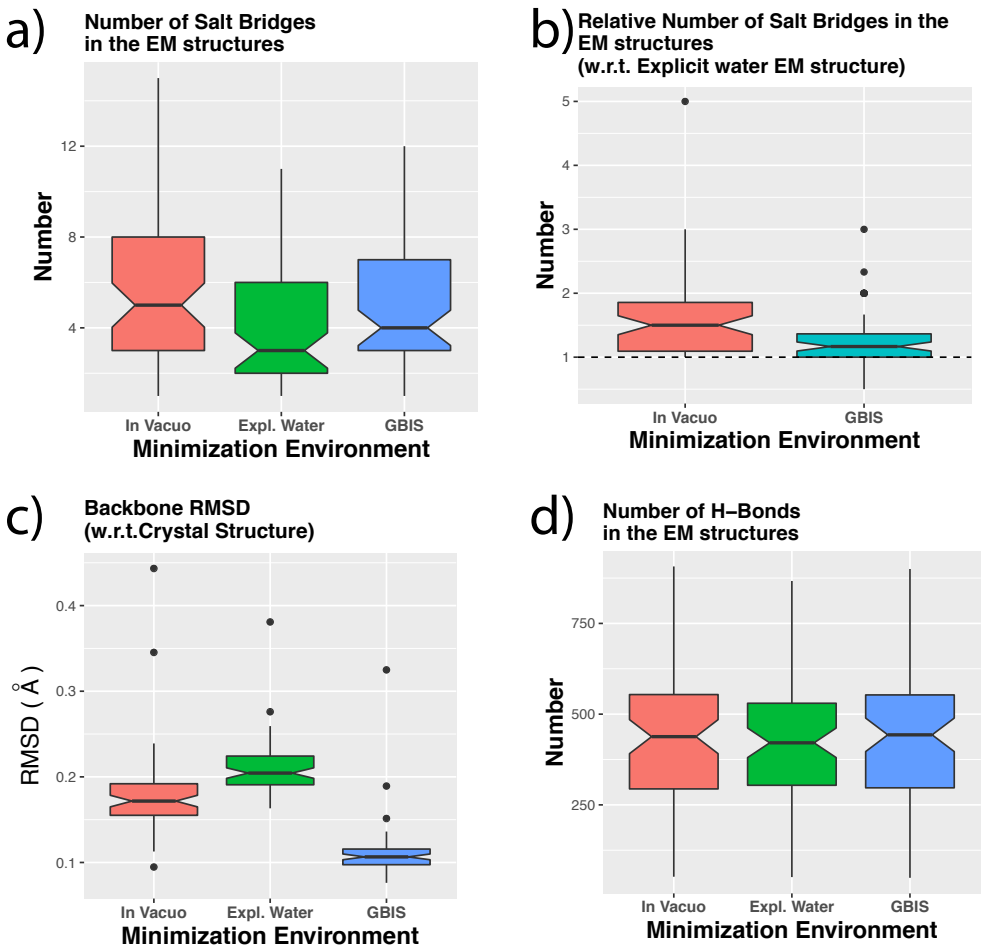


***Figure** 4.3.2: **Differences in the structural properties of the energy minimized configurations**. Boxplots showing (a) the distribution of the number of SBs in the energy minimized (EM) structures from the three environments, (b) the number of SBs for in vacuo and GBIS EM structures relative to that from explicit water environment, (c) the backbone structural RMSD of the structures relative to the crystal structure after minimization in the corresponding environment, (d) the number of intra-protein hydrogen bonds in the EM structures after minimization in different environments. The dotted horizontal line in (b) indicates the unity mark.*

Seeking to find a quantitative association of the change in polar solvation energy ΔG_{polar}^{solv} and the number of SBs formed or lost upon solvation, the difference of ΔG_{polar}^{solv} (computed using the traditional method and expressed as $\Delta\Delta G_{polar}^{solv}$) of the *in vacuo* minimized structure and the GBIS/Explicit solvent minimized structure against the difference in the number of SBs ($\Delta Number_{SB}$) in the two structures were plotted (**Figure** 4.3.3). As one can infer from the reasonably high r^2 values (0.525 and 0.884 for GBIS and Explicit Solvent, respectively) that a linear relation is evident. This is a clear indicative of how the solvent can affect the number of SBs and subsequently alter the polar solvation free energy. Moreover, since the ordinate in the plots is the true difference $\left(\Delta\Delta G_{polar}^{solv} = \Delta G_{polar}^{solv}(In\ Vacuo) - \Delta G_{polar}^{solv}(in\ solvent)\right)$, a greater loss of the SBs yields a more favorable solvation

(ΔG^{solv}_{polar} is more negative). This is a direct consequence of the antagonistic relation between ΔG^{solv}_{polar} and the coulombic energy U_{coul}.

The above quantitative association of SBs and the polar solvation energy have further implications when the dynamics of the proteins are considered. In the next section, we draw more weight onto these inferences and demonstrate how the breaking and forming of SBs in MD simulations is well mimicked by the Gaussian-model but not the traditional one. This, we show, influences the success or failure of a dielectric distribution model to reproduce ensemble average.

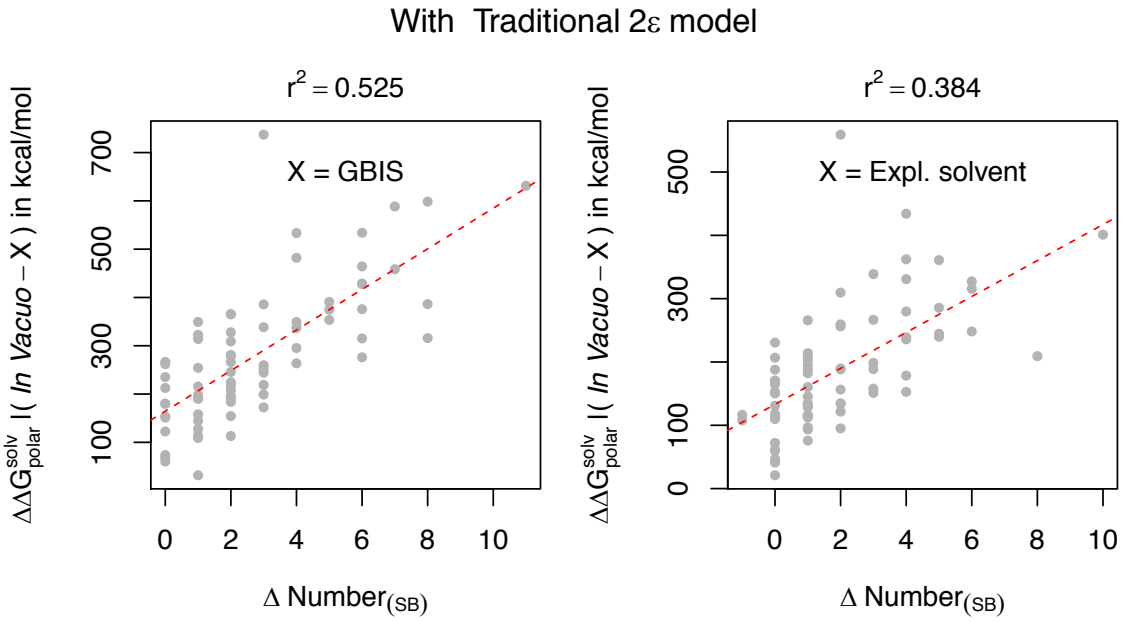


Figure** 4.3.3:* ***Polar solvation free energy and the number of salt bridges. *The difference of the ΔG_{polar}^{solv} , computed using the traditional 2ε dielectric model $\left(\Delta\Delta G_{polar}^{solv}\right)$ of the in vacuo and solvent minimized structures is plotted as a function of the difference of the number of salt-bridges in those structures. Left plot corresponds to GBIS and the right plot corresponds to explicit solvent (TIP3P). The quality of the linear fit (dotted red line) is quantified by the square of Pearson coefficient (r^2).*

4.3.3 Gaussian-based smooth dielectric model to mimic the fluctuations of the SBs.

To assess the implication of the fluctuation of the SBs on the ability of the either dielectric model to reproduce ensemble $\langle \Delta G_{\text{polar}}^{\text{solv}} \rangle$, the relation of the error of $\Delta G_{\text{polar}}^{\text{solv}}$ from these models (for *in vacuo* EM structure) with occupancies of the SBs was sought. We plotted the error $(\langle \Delta G_{\text{polar}}^{\text{solv}} \rangle - \Delta G_{\text{polar}}^{\text{solv}}(In\ Vacuo))$ against the number of SBs with occupancy $< 50\%$ (see **Figure** 4.3.4). The plot indicates if the error incurred by a dielectric distribution model deteriorates as more of the SBs present in the EM structure break during the MD. One can notice, from the linear trend in **Figure** 4.3.4**(a)**, that it is indeed the case with the traditional model. At the same time from **Figure** 4.3.4**(b)**, the error of the GAUSS-2 method is not only smaller than that of the TRAD-1 method but is independent of the occupancy of the salt-bridges.

This indicates that as more of the SBs, extant in the EM structure, have a tendency to break and stay 'broken' during the MD, the traditional 2-dielectric

method fails to reproduce the $\langle \Delta G^{solv}_{polar} \rangle$ ($r^2 = 0.545$). This can also imply that the ability of a dielectric model to capture ensemble $\langle \Delta G^{solv}_{polar} \rangle$ from a single structure significantly depends on its ability to mimic or capture the effect of fluctuations of the salt-bridges. This demonstrates that the Gaussian-based dielectric model (GAUSS-2) is able to capture the SB fluctuation's effect resulting in smaller error (than the TRAD method) and calculated polar solvation energy has no dependence on the occupancy of the SBs ($r^2 = 0.011$).

This can be attributed to the very basis of the Gaussian-based model. As is described in Ref.[63] and elaborated in the Methods section, the dielectric assigned to a region depends on the local atomic density, i.e., a region with lower atomic density is assigned a higher dielectric value and vice-versa. Consequently, the less dense regions will also have more room for motion owing to lesser likelihood for steric clashes with other solute atoms.

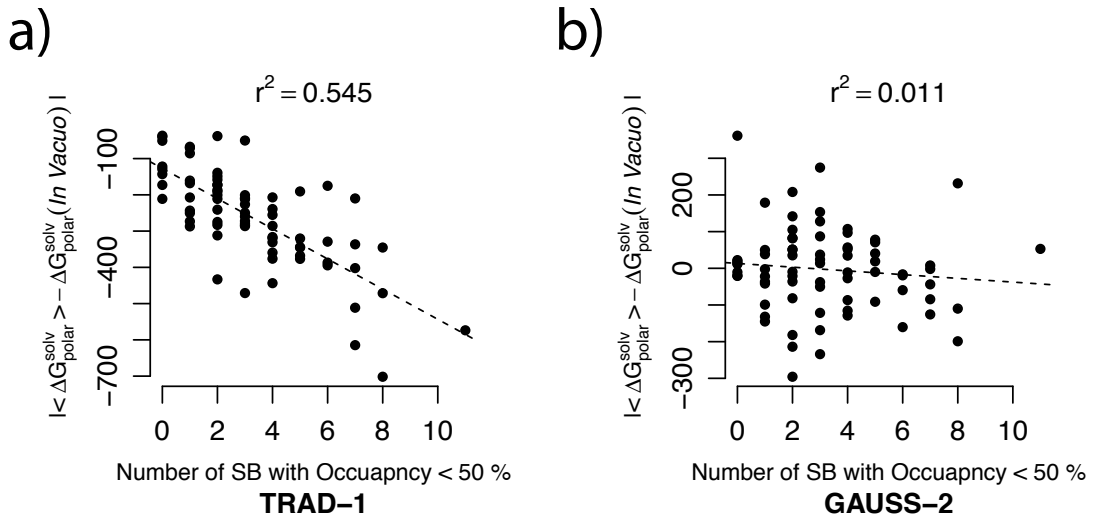


***Figure** 4.3.4: **Effectiveness of a dielectric model revealed by its ability to capture the dynamics of salt bridges.** The error in ΔG_{polar}^{solv} from using (a) traditional 2-dielectric method and (b) the Gaussian-based smooth dielectric model with in vacuo minimized structures with respect to the ensemble average (expressed as $\left|\langle \Delta G_{polar}^{solv} \rangle - \Delta G_{polar}^{solv}(In\,Vacuo)\right|$) are plotted as a function of the population of the salt bridges which were present for more than 50% of the frames in its MD generated ensemble (occupancy > 50%). The solid black lines depict the linear model fits to these comparisons and the r^2 value is mentioned for each of these linear fits. All energy units are kcal/mol.*

As a result, the Gaussian-method would assign regions of potentially high mobility a higher dielectric constant. Therefore, if the Gaussian-method yields a good agreement with the ensemble $\langle \Delta G_{polar}^{solv} \rangle$, it must be able to capture the SB fluctuations appropriately. Thus, it is expected that it should assign relatively higher dielectric constant in the vicinity of those SBs which a lower occupancy (more room for

fluctuation) than around those which have a higher occupancy (due to spatial restrictions arising from higher atomic density). To assess if that is indeed true, the local dielectric around the O-N atom pairs of the SBs identified in the *in vacuo* EM structure are computed to determine its relation with SB occupancy. The results are depicted in **Figure** 4.3.5.

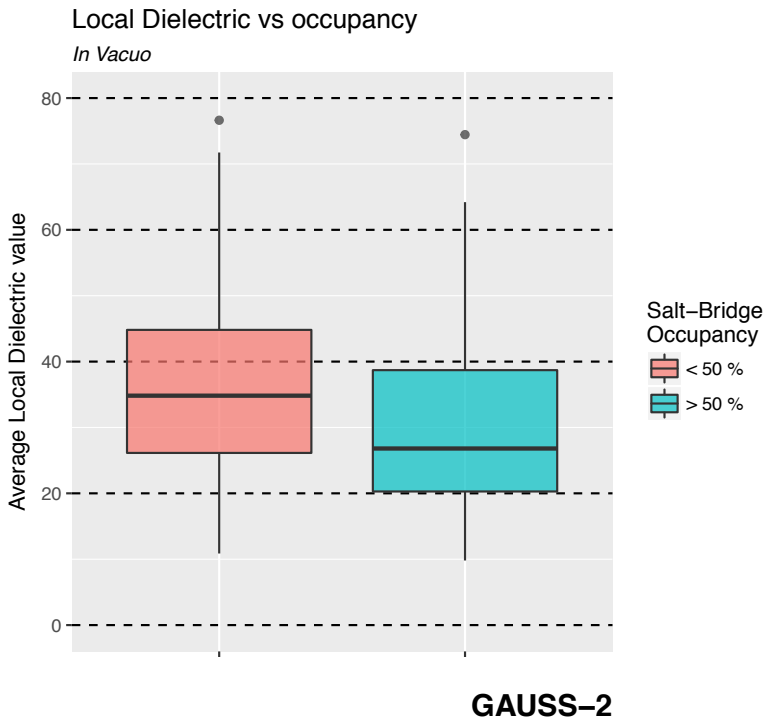


***Figure** 4.3.5: **Dielectric values assigned by models around salt-bridges**. Boxplots showing the distribution of the average dielectric constant assigned by the Gaussian-based smooth dielectric model in the locality of the salt-bridges (SBs) which have an occupancy < 50%(red) and > 50% (blue).*

In fact, the SBs with lower occupancies (< 50%) have a higher local dielectric constant on an average compared to the SBs that have an occupancy of more than

50%. This ratio is significant provided that regions populated with salt-bridges have, in general, a higher average dielectric constant than the buried regions rich in non-polar and polar residues[63] (see **Figure** A. 6).

Therefore, being able to capture the effects of fluctuation of the salt-bridge residues plays pivotal role in the success of a dielectric distribution in reproducing the ensemble average solvation energy.

4.4 Summary

The primary objective was to ascertain if the Gaussian-based smooth dielectric distribution (as implemented in *Delphi*) for modeling the dielectric distribution can mimic the natural dynamics of a protein and therefore, yield its ensemble average polar solvation energy using a single structure alone. The Gaussian-based model, in parallel with the traditional 2-dielectric model, was paired with structures minimized in different environments (*in vacuo*, GBIS and explicit water) and crystal structure of 74 proteins to study its ability to approximate the ensemble $\langle \Delta G_{polar}^{solv} \rangle$. Our study shows that the traditional dielectric model is able to reproduce a protein's $\langle \Delta G_{polar}^{solv} \rangle$ only with its crystal structure or a structure minimized in solvent. However, for most of the proteins, one would have to decrease

the dielectric internal dielectric (ε_{in}) to below 1, in order to achieve better approximations. This unreasonable modification can be circumvented by the use of Gaussian-based dielectric model. Not only does it yield a better agreement with the ensemble $\langle \Delta G^{solv}_{polar} \rangle$ for physically valid internal dielectric values (known as ε_{ref}), its performance is appreciable regardless of the minimization environment. In fact, for most of the cases, Gaussian-based dielectric model performs better than the traditional model, even if subtly. Upon comparing the overall results, we show and therefore, suggest that the use of Gaussian-based dielectric model with $\epsilon_{ref} = 2$, paired with a protein's *in vacuo* minimized structure, is best suited for reproducing its ensemble average polar solvation energy.

A detailed analysis revealed the reasons for the aforementioned differences in performance and other solvation energy trends. We found that the conformational states of SBs (open/closed) in a protein's minimized structure play an important role in offering one dielectric model an advantage over the other in terms of reproducing its ensemble average polar solvation energy. This means that a dielectric model, that best mimics the flexibility of the SB forming residues from their configuration in the EM structure, is better at reproducing the ensemble average polar solvation free energy. The Gaussian-based dielectric model is shown to accomplish this and

therefore is capable of generating ensemble average polar solvation energy of a protein from its *in vacuo* energy minimized structure. Our findings can henceforth, serve as a starting point for developing a time-inexpensive single structure MM/PBSA method.

5 DESCRIBING MOLECULAR GEOMETRY BY GAUSSIAN BASED MODEL OF ATOMS: A NOVEL GRID BASED ALGORITHM FOR DETERMINING MOLECULAR VOLUME AND SURFACE AREA

This chapter presents a novel method of implementing the Gaussian based atom model to compute solute/molecular volume and surface area (SA) in *Delphi*. The novelty lies in the exploiting *Delphi*'s finite difference setup designed to solve PBE to determine the pairs of atoms that overlap in space. The work in this chapter has also resulted into a publication[144] (copyright permission[4]).

5.1 Variations in the Non-polar solvation free energy models

Molecular geometry, best described by using the molecular volume (MV) or surface area (MSA) or both of a molecule, has served as a fundamental factor in

[4] See Appendix A.14 for copyright permissions.

modeling the non-polar properties of biological macromolecules. The prominent role of non-polar interactions in the formation of protein aggregates in solvent, protein-drug binding and membrane formation is well documented[145-147]. Furthermore, binding free energy changes occurring due to mutations in proteins also conceive the important role played by the change in the surface area of the mutant sites in predicting the pathogenicity of the mutation[76, 99]. Besides binding, studies on folding and unfolding of proteins have signified the role of MV and MSA[78]. From a geometrical perspective, these quantities help differentiate the level of packing of native from non-native and unfolded states of proteins. From a thermodynamic perspective, changes in volume and surface area signify the effect of pressure of protein folding in isothermal conditions, which is essentially the condition inside a biological cell. In addition, pressure-induced unfolding or denaturation and the associated volume changes of a protein has also been shown to have a significant dependence on the volumes of the internal cavities in its native structure[148, 149].

The non-polar component, which can be thought of as the energy required to create a cavity in the bulk of the solvent large enough to accommodate the solute in question, can be computed using various models. At the core of these models lies the assumption that the non-polar energy is related to the solute's volume or surface

area (SA) or both. The use of different models and their inclusion in the protocol for computing free energy can largely depend on one's understanding of the underlying physics and one's expectations from these computations. Some empirical models assume a linear relationship between the non-polar energy and the molecular surface area (MSA)[150-156] while others also include the molecular volume (MV)[68, 157-162]. By identifying some limitations of these linear models in describing the physical reality[163-165], recent models have suggested that in addition to the linear cavity term, an attractive van der Waals (vdW) term [67, 68, 158, 159, 161, 166-168] is also required to determine the total non-polar contribution to the free energy. Key variations amongst these different models originate from their definition of protein volume and SA. Most models use the solvent-accessible SA (SASA)[69, 153, 163, 169-171] of the proteins to quantify the size of the cavity while some also justify the use of van der Waals surface area (vdWSA)[48, 166-168]. As for the volume, some use van der Waals volume (vdWV)[166, 168], others the solvent accessible volume (SAV) [161, 162]. In addition, these models may also differ in the method they use to represent individual solute atoms, for example, as classical hard-spheres that draws a strict boundary between the solvent and solute regions or as regions occupied by a smooth volume density (expressed as Gaussians) which promotes a strict-

surface-free approach[85, 86] of describing solvated systems (which has been extensively discussed in the preceding chapters).

5.1.1 Need for efficient algorithms

This variety in non-polar energy models has called for different computational methods of computing volumes and SA. Besides differing on the model of atoms, these computational methods can also be distinguished in terms of the algorithm they use for identifying atomic overlaps[64, 172, 173] or that for delineating surfaces of contact of the probe and the solute atoms[33, 35, 174-176]. The use of one model over the other is certainly influenced by the time-efficacy and robustness besides the all-important physical meaningfulness. But as the number of structures in the Protein Data Bank (PDB) grows and genomic expansion studies are being undertaken widely, researchers are using a large number of structures in their studies and are sampling larger configurational spaces for a better and holistic understanding of biomolecular processes. As a result, the time-efficacy of a computational method has become a significant factor in influencing its choice over others.

The idea of the algorithm presented in this chapter combines a novel grid-based approach of identifying overlapping atoms and the analytical approach of

computing MVs and surface area using a Gaussian-based description of atoms[64]. The primary motivation is to integrate a method of computing MV and MSA, and therefore, a method of computing non-polar energy terms, into *Delphi*[70]. The use of a smooth Gaussian-based model will make this merger consistent with its smooth Gaussian-based approach of representing the dielectric distribution of solvated biomolecular systems[86, 104] (discussed in the preceding chapters). This integration is expected to provide a comprehensive platform for computing the free energy using a single package and thereby, offer a wide range of users a convenient way of analyzing and evaluating the energy of system configurations sampled from large-scale simulations using the MM/PBSA protocol.

The novel grid-based algorithm is designed to identify pairs of solute atoms that overlap in space by simultaneously using the robust grid-based finite-difference method that *Delphi* uses to solve the Poisson-Boltzmann Equation (PBE). By doing so, it is shown that little to no additional time is spent in identifying overlapping atom pairs. After the pairs have been identified, a depth-first tree based algorithm, used by the popular AGBNP[166] package, is used to compute the volumes and surface areas.

5.2 The Gaussian model of computing molecular volume and surface area

The mathematical basis of the Gaussian model of atoms has been extensively discussed in Chapter 2. This formalism is adapted from the seminal work of Grant and Pickup[64]. Of relevance to this chapter is the formalism that derives the volume and surface area of a molecule whose atoms are represented using the Gaussian model. Therefore, the preliminary concepts of Gaussian density are skipped here (see section 2.1 for details). The following shows how the volume and surface area are computed using the Gaussian models.

5.2.1 Gaussian product theorem for computing volumes and SA of overlapping regions

The product, $g_{ij} = g_i g_j$, of the Gaussian density functions of two atoms i and j describes the volume of their overlap. The product is itself a Gaussian density function centered at

$$\overrightarrow{r_{ij}} = \frac{\alpha_i \vec{r_i} + \alpha_j \vec{r_j}}{\alpha_{ij}} \tag{30}$$

where $\vec{r}_i \; and \; \vec{r}_j$ are their respective positions and $\alpha_i \; and \; \alpha_j$ are the exponent scaling factors. In the product Gaussian term, the resultant Gaussian exponent acquires the form

$$\alpha_{ij} = \alpha_i + \alpha_j \tag{31}$$

Correspondingly, in the Gaussian formalism the overlap volume, V_{ij}, of two atoms is given by the volume integral of their product density

$$V_{ij} = \int_V dV g_{ij} = p_{ij} e^{-\left(\frac{\Lambda_{ij}}{\alpha_{ij}}\right)} \left(\frac{\pi}{\alpha_{ij}}\right)^{\frac{3}{2}} \tag{32}$$

where the resultant height factor p_{ij} is given as $p_{ij} = p_i p_j$ and the factor $\Lambda_{ij} = \alpha_i \alpha_j \left|\vec{r}_i - \vec{r}_j\right|^2$. This strategy can be extended recursively to obtain analytic expressions of the Gaussian overlap volumes of any order. For instance, the third order overlap volume of atoms *i, j* and *t* is expressed as

$$V_{ijt} = \int_V dV \; g_{ijt} = p_{ijt} e^{-\left(\frac{\sum_{m,n=\{i,j,t\}|m\neq n} \Lambda_{mn}}{\alpha_{ijt}}\right)} \left(\frac{\pi}{\alpha_{ijt}}\right)^{\frac{3}{2}} \tag{33}$$

Where

$$p_{ijt} = p_{ij} p_t \tag{34}$$

$$\alpha_{ijt} = \alpha_{ij} + \alpha_t$$

And

$$\overrightarrow{r_{ijt}} = \frac{\alpha_{ij}\overrightarrow{r_{ij}} + \alpha_t\overrightarrow{r_t}}{\alpha_{ijt}} \tag{35}$$

To compute the MV, overlap volumes are added to or subtracted from the arithmetic sum of the hard-sphere volumes of all the atoms, based on their order (inclusion-exclusion formula). The alternative inclusion and exclusion ensure that there is no redundancy in the contribution by a certain overlap region to the total volume.

$$V_{molecule} = \sum_i \frac{4}{3}\pi R_i^3 - \left(\sum_{i<j} V_{ij}^{overlap} - \sum_{i<j<t} V_{ijt}^{overlap} + \sum_{i<j<t<s} V_{ijts}^{overlap} + \cdots \right) \tag{36}$$

The terms in the parenthesis of equation 36 in the right-hand side comprise the total overlap volume. Note that they occur with alternating signs of the form $(-1)^n$ where n is the order of the overlap.

The surface area SA_i of atom 'i', is defined as the derivative of the MV with respect to the radius of that atom. The total SA of the molecule is obtained from equation 37 as the sum of the individual atomic surface areas as

$$SA_{molecule} = \sum_i SA_i$$

$$= \sum_i \left(\frac{\partial V_i}{\partial R_i} - \sum_j \frac{\partial V_{ij}}{\partial R_i} + \sum_{j,t} \frac{\partial V_{ijt}}{\partial R_i} - \sum_{j,t,s} \frac{\partial V_{ijts}}{\partial R_i} + \cdots \right) \tag{37}$$

In the context of the Gaussian model, overlap volumes and their derivatives are available in analytic form. For a generic overlap term of order n, the derivative with respect to the radius of atom i is given by

$$\frac{\partial V_{ij\cdots n}}{\partial R_i} = \frac{\partial V_{ij\cdots n}}{\partial \alpha_i} \left(\frac{\partial \alpha_i}{\partial R_i} \right)$$

$$= \frac{2\kappa_i}{R_i^3} \left[\frac{3}{2\alpha_{ij\cdots n}} + \left| \vec{r_i} - \overrightarrow{r_{ij\cdots n}} \right|^2 \right] V_{ij\cdots n} \tag{38}$$

5.3 Identifying overlapping atom pairs and computation of volume and SA

The above mathematical description of the model emphasizes on the importance of the overlapping volume and SA terms to these calculations. These terms are contributed by atoms that share a region, which means that each atom has its own set of neighboring atoms that affect its volume/SA. Typically, such pairs of atoms are found using a distance criterion, wherein two atoms, i and j, are said to be overlapping if:

$$\left|\vec{r_i} - \vec{r_j}\right| \leq R_i + R_j + \epsilon \tag{39}$$

where R represents their respective radius and $\vec{r}$ designates their center coordinates, such as those provided in a PDB file. ϵ, typically, has a small value that provides allowance for those pairs of atoms which wouldn't overlap were they to be described as classical hard-spheres. Finding out this pair-list, also known as neighbors list, therefore, requires $O(N^2)$ operations, in theory. Algorithms like cell-linked-list[177], domain-decomposition method[178], Verlet list[179] and others[180-182] were contrived solely to cut down on the computation time and are mainly incorporated with MD simulation packages.

The grid-based approach makes use of the 3D setup of grids constructed by *Delphi* in order to solve the Poisson-Boltzmann equation (PBE) using finite difference method[41, 70]. The neighbor list of the atoms is computed in this symmetrical 3D mesh of grids (also called box) on which the molecule in question is projected into. The box is large enough to accommodate the molecule fully and have an additional space around it to account for the solvent phase. Based on the number of grids per Å (*a.k.a* resolution or "scale"), the fineness of the 3D mesh can be manipulated. The details of the grid construction can be found in (section 1.5.1).

As is described in the section *Overview of Delphi's workflow* in Chapter 1, the first step of *Delphi*'s algorithm is to determine the dielectric distribution of the system contained in the box (the *Space* module). With the information of the coordinate of the atoms and their radii, grid points are surveyed and based on its distance from the center, a dielectric value is assigned. Since evaluating all the grid points can be extremely expensive, only a cubic region around the atom in question, large enough to accommodate its spherical volume is scanned[70]. Consecutive atoms are projected onto to the grid points and a 3D dielectric distribution map is constructed. It is at this step the neighbor list of atoms is generated. As consecutive atoms are projected onto the grid, computation of neighbor list runs in parallel. It

uses the following criteria to identify neighbors: *two atoms are considered as neighbors if the local cubic box around them share at least one grid point.* If the boxes are larger, more neighbors will be identified and vice versa. However, overestimation of the number of neighbors will not necessarily overestimate the volume. It will simply increase the computation time.

5.3.1 Grid-based algorithm for finding overlapping atoms

To provide the exact schematic of this approach's algorithm, an example molecule of 5 atoms is used for demonstration. Without any loss of generality, the grids are portrayed in 2D and the atoms are be described as circles of radius equal to their van der Waal radius. This is illustrated in **Figure** 5.3.1. In the figure, the flow of steps is represented by a number on each of the panel, going from '1' through '6'.

Step (1). A mesh, large enough to encompass all the atoms of the input molecule, is defined. A labelling system is used wherein each grid point is labelled by an integer. To initialize our grid, we assign '0' to each grid point.

Step (2). A separate $(N+1) \times (N+1)$ square matrix, depicting atom pairs that overlap in space is defined. We will refer to the matrix as the ***atom-overlap***

matrix or ***AOM***. All the atoms in the molecule with indices $1, 2, \cdots, N$ are considered along with a dummy atom of index 0. An element of this matrix is defined as $AOM_{m,n} \in [True, False]\ \forall\ m, n \in [0,1,2, \cdots, N]$ such that if atoms m and n overlap in space, $AOM_{m,n} = True$ otherwise *False*.

Step (3). The first atom (with index '1') in the list is placed onto the grid. As the grid-points in the vicinity of atom 1, contained in its local cubic box (shown as squares in the **Figure**), are surveyed by *Delphi*, grid points that lie within a distance of kR_1 from the center of atom 1 ($\vec{r_1}$), i.e. those that satisfy the distance criterion $\left|\vec{r}_{grid} - \vec{r_1}\right| \leq kR_1; k \in \mathbb{Z}^+$, are made to undergo a change in their integer label. *'k'* here is a factor that affects the volume of the box that is searched for grid points that fit the criteria. From the initial '0' label, they are assigned a label of '1' since they lie in the vicinity of atom 1. This change in label of the grids is accompanied by updating $AOM_{0,1} = True$. Essentially, the matrix element with row-index equal to the old label and column-index equal to the new label is updated to *True*.

Step (4). The second atom (with index 2) is placed onto the grid. Grid points that satisfy the above distance criterion with respect to atom 2, are surveyed and their labels are updated accordingly. Those with '0' are now labelled as

'2', causing $AOM_{0,2} = True$ and those with '1' are now labelled as '2', causing $AOM_{1,2} = True$.

That $AOM_{1,2} = True$ exists implies that atoms 1 and 2 potentially overlap.

Step (5). The third atom (index 3) is placed onto the grid. At this point, grid points are labelled as either '0' or '1' or '2'. Grid points satisfying the distance criteria with respect to atom-3 result into updating $AOM_{0,3}, AOM_{1,3}$ and $AOM_{2,3}$ to *True*.

Step (6). Similarly, atom 4 and 5 are treated and the corresponding elements in the AOM are updated.

It must be noted here that since atoms are used in an increasing order of their index, the above procedure will only update the upper triangular block of the AOM. This doesn't result in losing any information because if atoms 'm' and 'n' overlap (where $m < n$) due to $AOM_{m,n} = True$, then it directly implies that $AOM_{n,m} = True$.

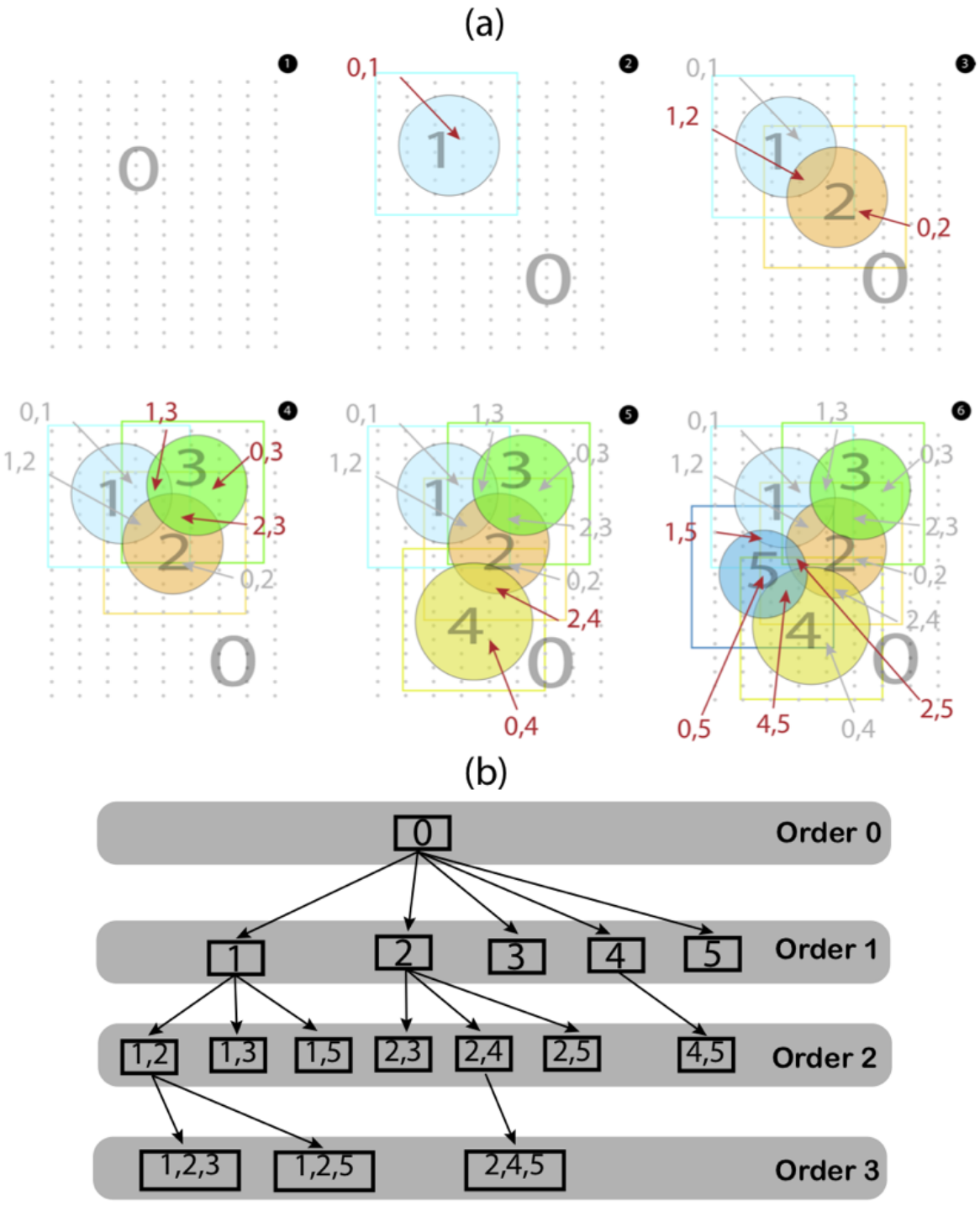

(a)
0
0,1
1
0
0,1
1,2
1
2
0,2
0
0,1
1,3
0,3
1
3
2,3
2
0,2
0
0,1
1,3
0,3
1,2
1
3
2,3
2
0,2
4
2,4
0
0,4
0,1
1,3
0,3
1,2
1
3
1,5
2,3
5
2
0,2
4
2,4
0
0,5
4,5
0,4
2,5
(b)
0
Order 0
1
2
3
4
5
Order 1
1,2
1,3
1,5
2,3
2,4
2,5
4,5
Order 2
1,2,3
1,2,5
2,4,5
Order 3

Figure** 5.3.1:* ***An illustration of the grid-based algorithm designed for identifying atom pairs that overlap in space. *(a) The algorithm of identifying overlapping atom pairs is visually illustrated. Each atom is shown as a colored circle surrounded by a square of the same color depicting the local box that is searched for grid-points in its vicinity. The systematic flow of the steps is indicated by the label on the top-right corner of each panel in the* ***Figure****. Two atoms 'i' and' 'j' that overlap update the atom overlap matrix (AOM) element* $AOM_{i,j}$ *to True. At each step, new indices of AOM that get updated to True are shown in red. The numeric labels placed at different regions are meant to indicate the integer label on the grid-points present in a region. (b) A rooted tree constructed using the neighbor list of all the atoms in the molecule and an additional dummy atom with index '0'. Each level or order is marked using grey horizontal bars. From top to bottom, levels of increasing orders are shown.*

The final *AOM* used to prepare the neighbor list. The following steps are performed.

Step (1). For each atom, an empty neighbor list (to store integers or atom-indices) is defined.

Step (2). An iterator navigates through the upper triangular part of the symmetric *AOM* and checks for all the elements $AOM_{m,n} | m \leq n$ which are *True*.

Step (3). For any $AOM_{m,n} = True$, index n is appended to the neighbor list of atom m.

For our example molecule with 5 atoms, Table 5.1 shows the AOM (in the upper-triangular form) and the list of neighbors for each atom. The outcome can be confirmed by the arrangement of atoms in **Figure** 5.3.1**(a)**. Also note that, atom with index '0' is a dummy atom and it automatically has all the "real" atoms of the molecule in its neighbors list. This helps in the construction of a rooted tree that is used for computing overlap volumes and surface area.

Table 5.1: The Atom Overlap Matrix or AOM (top panel) and the neighbor list of atoms inferred from it (bottom panel) for the 5-atom example molecule obtained using the grid-based neighbor search algorithm. For clarity only the upper triangular part of the symmetric matrix is shown.

Atom Overlap Matrix (AOM)						
	0	1	2	3	4	5
0	-	T	T	T	T	T
1		-	T	T	-	T
2			-	T	T	T
3				-	-	-
4					-	T
5						-
'True' is represented as 'T' and 'False' is represented as '-'						

Neighbor List of the atoms	
Atom Index	List of neighboring atoms
0 (Dummy Atom)	[1, 2, 3, 4, 5]
1	[2, 3, 5]
2	[3, 4, 5]

3	[]
4	[5]
5	[]

5.3.2 Depth-first traversal method for computing total volume and surface area of overlap information

The neighbor lists of the atoms are used to construct a rooted tree of overlaps, the hierarchy of which follows the order n of the overlaps. Each node of the tree holds the value of the overlap volume which are arithmetically added, according to equation 36, to yield the total MV. For computational efficiency, overlap volume terms with values less than 0.001 A^3 are neglected. A volume cutoff of this kind is necessary since the Gaussian overlap volume of two distant atoms, albeit infinitesimally small, will never be zero. In parallel to volume calculations, the SA term for each node is also computed and the total molecular SA is obtained.

The basic premise of constructing a tree by a "depth-first" algorithm and using it for volume/SA computation is identical to the one used in reference [168]. Each atom is assigned an integer index (starting from 1) and a dummy atom with index '0' is used to build a rooted tree with it being the designated root. Each

subsequent level in the tree is assigned an order based on its distance from the root; the root is assigned an order 0 at the first step of the process. All the atoms are then defined as the children of the root, hence, forming the next level down the hierarchy with order 1. Each of these atoms then initiate a separate branch of the tree. The tree grows more levels by incorporating new nodes of the next order that contains the information of all the common neighbors of its ancestors. Eventually, a node of any order is designed to contain the information of all the neighbor common between itself and its ancestors. Computationally, a node of order 'k' is represented by an ordered list of 'k' atom indices such that the atom with the k^{th} index is a common neighbor t of all the 'k-1' atoms preceding it. Geometrically, that implies that all the k-atoms overlap in space. For e.g. if a node (1, 2, 3, 7) exists for an arbitrary molecule, it would imply that atom-7 is a common neighbor of atoms-1, 2 and 3. It would also mean that the four atoms overlap in space. A branch of the tree is terminated when a new common neighbor isn't found or when the volume of that particular node is smaller than the cutoff value (0.001 $Å^3$, see above). For computational efficiency, we limited the order of nodes to 6. As the branch reaches a "dead-end", the next branch from the top of the tree is worked upon in the same recursive manner till all the branches growing out of the root have been covered. For

our example case of the 5-atom molecule, **Figure** 5.3.1**(b)** is an illustration of its depth-first tree.

It must be noted that, though the Gaussian-model projects a physically meaningful picture of a protein-solvent system, the mathematical formulation can harbor some unphysical issues. Therefore, it is necessary that they are eliminated correctly. An example is negative surface areas for deeply buried atoms surrounded by many neighboring atoms[168]. For such atoms, it is likely that certain orders of the overlap volume, which have a negative contribution to its total surface area, add up to be larger than its individual volume (e.g. order 2). To correct for this, we devised a physically appealing way of filtering the contribution of these atoms to the total SA. This filter uses a smooth sigmoid function of the form

$$SA_{filtered},i = SA_i\left(\frac{1}{1+\ e^{g(-SA_i+SA_{cutoff,i})}}\right) \tag{40}$$

Here '*i*' depicts an atom and SA_i is the surface area computed by the Gaussian model**.** '*g*' is a dimensionless constant with a value 5, assigned after optimization. $SA_{cutoff,i}$ is a threshold value of an atom's SA that decides its contribution to the

total SASA of the molecule. Only the atoms with values larger than the cutoff contribute. The cutoff is computed using a hard-sphere approximation and hence depends on the radius of a solvent-probe (R_{probe}; 1.4 Å for water) and the radius of that atom (R_i). An atom is considered solvent accessible if it can allow at least one solvent molecule (in its hard sphere form) to share a tangential plane with it. The cutoff, therefore, acquires the following form.

$$SA_{cutoff,i} = SA_i \left[\frac{1 - \cos(\theta)}{2}\right] \tag{41}$$

where the angle 'θ' is the solid angle subtended by a cone of height $R_{probe} + R_i$ and base radius R_{probe}. It can be expressed as

$$\theta = 2\tan^{-1}\left(\frac{R_{probe}}{R_{probe} + R_i}\right) \tag{42}$$

Figure 5.3.2 provides a visual reference which exemplifies the case of a solvent of probe radius 1.4 Å and an atom of radius 2 Å.

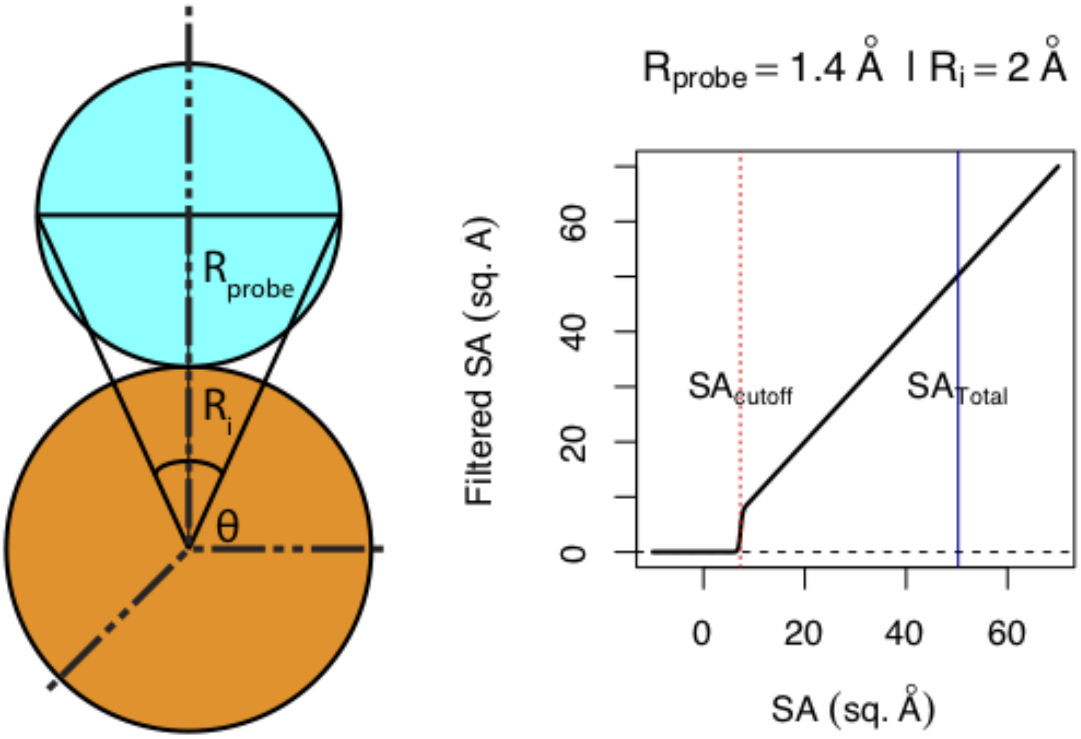


***Figure** 5.3.2: **Solvent accessibility based filter for determining atomic and molecular surface areas.** (Left) Illustration of the physical basis of the function used to compute cutoff atom-specific surface area and filter out the contribution of atoms with negative surface area terms. (Right) The output yielded by the filtering function.*

5.4 Validation of the algorithm

The grid-based approach was validated at three different levels. First, the volumes and SAs for a library of 74 proteins of sizes ranging from 50 to 200 residues (used previously [121]) were calculated using implementation as well as AGBNP[166, 168] and then were compared to determine the numerical differences. Second, the effect of grid-resolution on the output of volume and SA was examined. Third, the

accuracy in identifying "correct" neighbors using the grid-based approach was evaluated by comparing the neighbors identified using a standard $O(N^2)$ analytical approach (see Equation 39).

5.4.1 Validation of the volume/SA output

Figure 5.4.1**(a, b)** shows the comparison of the volumes and SA of 74 proteins computed using our implementation of the Gaussian model and that of AGBNP. The quality can be adjudicated by the slope and intercept of a linear regression fit as well as the correlation (R^2) accompanying the figures. Slopes approximately equal to 1.00 (with relatively infinitesimal intercepts) and correlations equal to 1.00 indicate that our implementation is precise. In addition, it is also acknowledged that the resolution of grids *("scale" in Delphi*) can have an effect on the volume/SA value by having an effect on the neighbor search process. Therefore, different values of scale were also used and the resulting volume/SA outputs were compared with AGBNP. The results are shown in **Figure** 5.4.1**(c)** in terms of the root mean square relative difference (RMSRD; see **Appendix A.11**) incurred as a function of the grid resolution, which indicates that the differences are small, i.e., ~0.40% at a low resolution of 1 grid/Å and ~0.15% at 2 grids/Å and become

infinitesimal (<0.1%) at 3 and 4 grids/Å. But since increased resolutions mean non-linear increase in computational times (cubic power), one should consider a balance between accuracy and computational time.

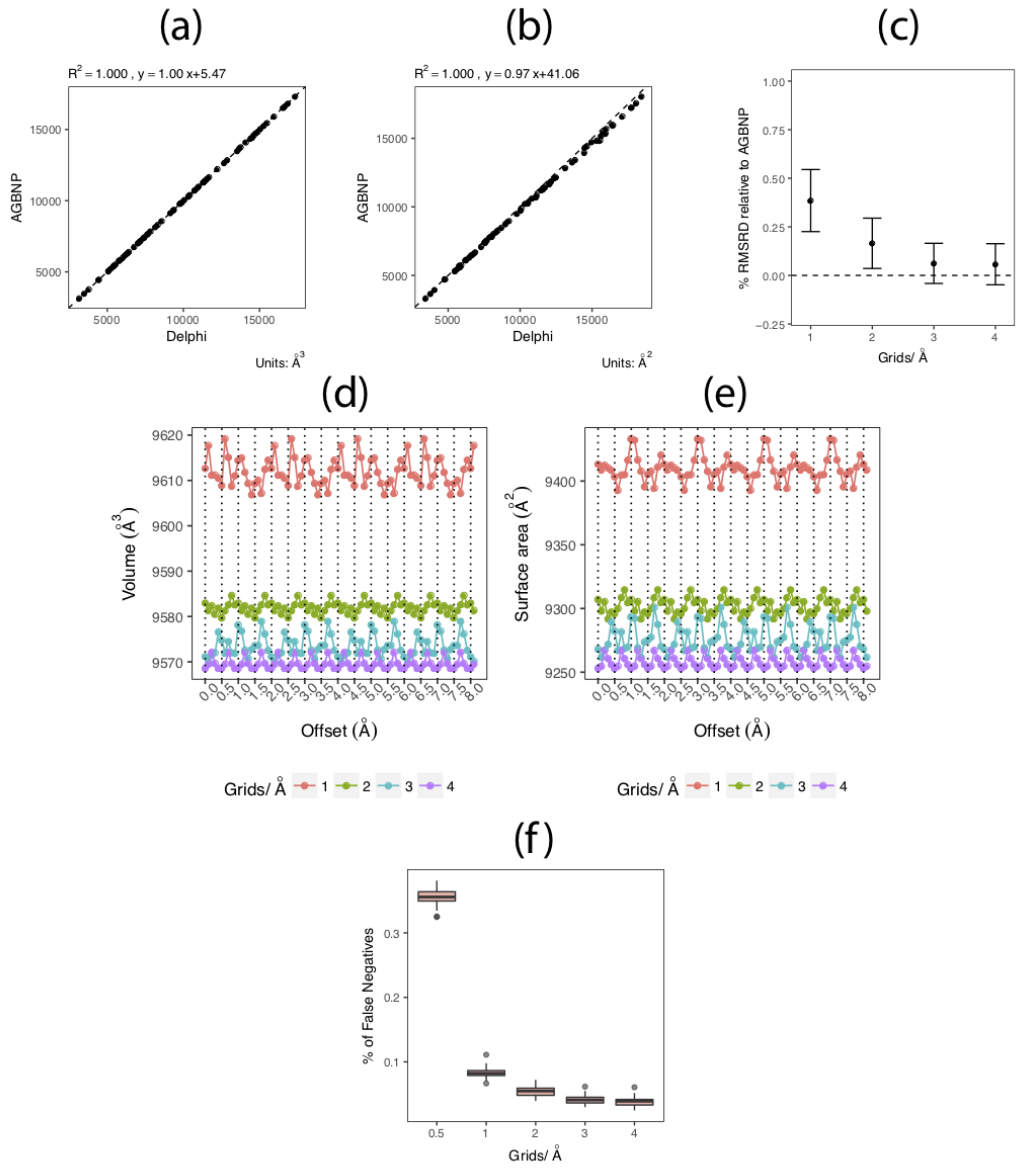


***Figure** 5.4.1: **Validation of the grid-based algorithm for identifying overlapping atom pairs**. Comparisons of (a) the molecular volumes and (b) the molecular surface areas of 74 proteins obtained using the grid-based algorithm in conjunction with the Gaussian-model and obtained using AGBNP. (c) Percent relative difference (RMSRD) of the molecular volumes of the 74 proteins with respect to the values output by AGBNP as a function of the scale or grid-resolution. (d) Volume and (e) surface area of Barstar (PDB: 1X1X, chain D) plotted as a function of the offset in its position from the center of the grid box. (f) Percentage of falsely missed atom pairs overlapping in space (False Negatives) by the grid-based algorithm plotted as a function of the grid-resolution (grids/Å).*

5.4.2 Effect of positioning in the grid box

For the second level of validation, the effect of differently positioning the solute inside the box was examined without changing the position of the grids. This was important because in the initial phase of a *Delphi* run, the coordinates of a 3D structure (from PDB for instance) are projected onto these grids using a distance-dependent interpolation technique (see Chapter 1: *Overview of Delphi's workflow*). For this test, Barstar (PDB ID: 1X1X, chain D) was chosen and its position was changed continually along an arbitrarily chosen direction (without loss of generality), by offsetting it's coordinates from the center of the box in small incremental steps and computing the volume and SA using the Gaussian model. **Figure** 5.4.1(d, e) show the outcomes as a function of the offset distance for different values of scale.

A tendency to vary periodically w.r.t to the offset is seen in these plots. The periodicity also varies with grid resolution, in that, the period is inversely proportional to the number of grids/Å. This is because with different offsets, the projection of the atom coordinates on the grids changes and this affects the neighbors identified in the process and congruent grid placements occur in multiples of the grid resolution. Eventually, that results into variations in the volume and SA outputs. But these variations are minor in comparison to the average values ($< 0.05\%$). This leads to a conclusion that the grid-based approach is appreciable precise and is only minutely sensitive to the arrangement of the grid points in the box.

5.4.3 Accuracy in predicting overlapping atoms

At the third level of validation, we evaluated our method's accuracy in determining the "correct" neighbors. The above tests showed that our approach is minutely sensitive to the grid resolution and positioning of the solute in the box. This is because of the differences in the placement of the grids with respect to the atomic coordinates of the solute. In a standard approach, neighbor list for the atoms can be computed using a distance-based criterion where two atoms with coordinates separated by a distance lower than the sum of their radii are considered as neighbors

(Equation 39). But in our grid-based approach, two atoms are considered as neighbors if their box of grids surrounding the respective spherical volumes share common grid-points (see section *Grid-based algorithm for finding overlapping atoms*). Therefore, we compared the neighbor list yielded by the grid-based approach, at different grid resolutions, with that obtained by using the standard $O(N^2)$ approach. This test was expected to report neighbors that are common to both the approaches (*True Positives*) or are neighbors based on one approach and not the other (*False Negatives* or *False Positives*). The grid-based approach would ideally "pass" the test if it can identify at the very least all the neighbors that the standard approach would. Any additional neighbors detected (*False Positives*) would later be filtered out based on the volume of their shared region (i.e. < 0.001 Å^3). However, if a vast percentage of neighbors is only found by the standard approach and not by the grid-based approach (*False Negatives*), it would question the method's credibility. Our focus is to detect the percentage of such cases. **Figure** 5.4.1(f) shows the outcomes. As a function of the grid resolution, the percentage of *False Negative* cases are plotted. Each boxplot depicts the range of percentage of *False Negatives* found across a library of 74 proteins and the solid black line close to the center of these boxplots is the median value of the distribution (See **Appendix A.12**). There are two major

observations: 1) The percentage of the False Negative cases are infinitesimally small (<0.4%) if not exactly 0.0 at a very coarse resolution of 0.5 grids/Å. With finer resolutions, the percentage drops to ~0.05%. This means that the grid-based approach could likely miss 1 in every 2000 neighbors identified using the standard approach. This imparts an added confidence in the accuracy of this approach.

5.5 Performance of the algorithm

We also assessed the time efficiency and complexity of the grid-based approach. Theoretically, it is an $O(8NR^3G^3)$ complex algorithm, where 'N' is the number of atoms and 'R' is the average atomic radius and 'G' is the number of grids/Å. This is because for each atom out of N, a local cubical volume around its center is surveyed for the grid points which is later used by *Delphi* for assigning dielectric values and distributing charges. This local cube is of length proportional to $2R$ (average atomic diameter), making its volume $8R^3$ and the total number of grid points to be scanned equal to $8R^3G^3$.

But the integration of the grid-based algorithm in parallel with other grid-based operations performed by *Delphi* makes it difficult to evaluate the exact time. Therefore, we measured the total time taken by the grid-based neighbor search

algorithm and the volume/SA computation using our implementation of the Gaussian model and subtracted it from the time taken by *Delphi* when these calculations are turned off. This gives an estimate of the average time efficiency as a function of grid resolution and size of the solute.

Figure 5.5.1 plots the average time over 10 runs vs the number of atoms for different grid resolutions. It is clear that time taken for volume/SA computation along with the neighbor search part is typically < 3 sec for proteins with 1000-3000 atoms. Also increasing the resolution appears to drastically increases the time. The effect is prominent when the number of atoms is more than 1000. This is because with increased resolution, the number of neighbors identified by our approach is much larger than that by the standard distance-based approach. In other words, the percentage of *False Positives* increase (see **Figure** A. 7).

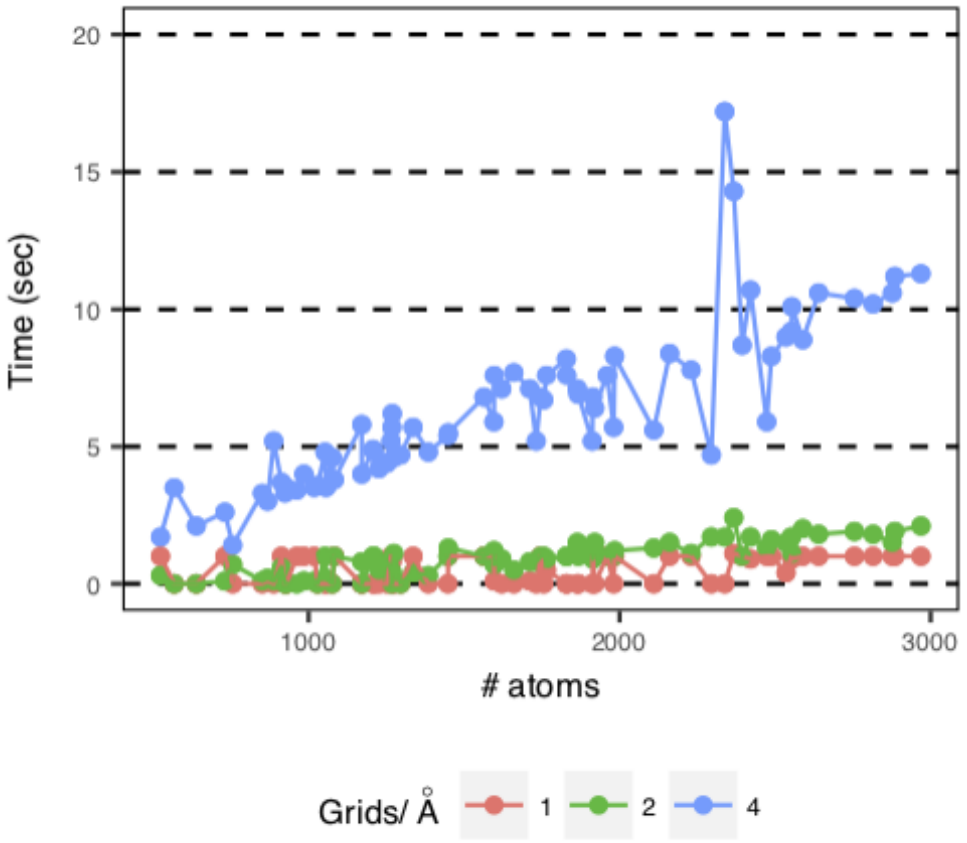


***Figure** 5.5.1: **Performance.** The average run time as a function of the number of atoms in the solute and grid resolution (grids/Å). 74 proteins were used for the test and the average time was computed by averaging over 10 runs on each protein. Since the standard deviations of the runtimes were infinitesimally small, error bars depicting them are deliberately not shown.*

5.6 Results and Discussion

The contents of this section review basic aspects of the Gaussian model and in addition also points out other aspects that have not been addressed meticulously

in the literature before. The idea is to emphasize the numerical accuracy, physical appeal and an inevitable limitation of the model.

5.6.1 Volume and surface area computed using the Gaussian model:

The Gaussian model, as proposed by Grant and Pickup in their seminal work, delivers the van der Waals volume (vdwV) and surface area (vdWSA) of molecules[64]. Using our implementation with the grid-based neighbor search algorithm and the library of 74 proteins, it was found that the volumes delivered by it differ from the hard-sphere vdWV by ~7-8% (the latter was computed using the package ProteinVolume[176]). Another package called 3V[183] was also used for a thorough benchmarking and it was found that the difference, in this case, was smaller (difference <1%). In terms of the surface area, the output from the Gaussian model differed by ~6-6.5% from the vdWSA computed using the hard-sphere model by FREESASA[184]. Once again for a thorough benchmark, another package called NACCESS[119] was also used and the difference was found to be ~4.5%. The exact values of these differences, expressed using RMSRD and the outcomes of a linear regression fit to the model comparisons are presented in Table 5.2. For all the

Gaussian model based calculations, a κ value of 2.227 was used[166, 168] (see Equation **2**) and a resolution of 2 grids/Å was used for grid-based neighbor search.

There are two major inferences we draw from these comparisons. First, our implementation of the Gaussian model precisely delivers the molecular vdWV and vdWSA indicating that the implementation correctly reproduces the expected behavior of the Gaussian model. Second, volumes/SA computed using the hard-sphere model do not offer a strict reference for benchmark and validation. This is evident from comparing the results computed different software. Indeed, different packages implementing the hard-sphere model yield different values (Table 5.2).

Table 5.2: Comparison between van der Waals volumes and surface area of proteins and surface area of individual atoms obtained using the Gaussian model and the hard-sphere model. The comparison is quantified by the slope, intercept of the linear regression fit, correlation (R2) and the root mean square relative difference (RMSRD)

	RMSRD	Slope	Intercept	Correlation (R^2)
van der Waals Volume of Proteins				
ProteinVolume	7.70%	1.08	7.40 Å^3	0.999
3V	0.24%	1.00	7.20 Å^3	0.999
van der Waals Surface Areas of Proteins				
FREESASA	5.9%	0.94	32.09 Å^2	0.999
NACCESS	4.3%	0.95	64.51 Å^2	0.999
van der Waals Surface area of individual atoms				
FREESASA	15.9%	0.89	0.70 Å^2	0.953

The comparison between the two models was extended further to the level of surface areas of individual atoms. Using the two models, surface area of individual atoms across the 74 proteins were computed and compared. The result, also shown in Table 5.2**,** clearly indicates that the Gaussian model can deliver precise surface areas of individual atoms with a difference of only 15.9% with respect to the values computed using the hard-sphere model. This good quality of the agreement is also evident from the slope and intercept of the linear regression and a correlation of 0.953. This ability to deliver proper surface areas of individual atoms provides the Gaussian model with an added advantage. Several packages like AGBNP[166, 168] and ACE[185], that run molecular dynamics using the Gaussian model, make use of this ability to correctly compute the energy and forces on individual atoms alongside the continuous and differentiable analytical expressions for this terms. In addition to this, atom-specific surface-tension coefficients used in conjunction with individual atomic surface areas have been shown to deliver non-polar part of the free energy in good agreement with that from explicit solvent simulations[69].

5.7 Physical appeal of the Gaussian model

In addition to numerical precision, one of the key features of a smooth Gaussian-based model is that the transition area between solute and the solvent phases does not have to be sharp. To demonstrate this, a profile of the change in the vdWV/SA of a protein complex as a function of the distance between the monomers is presented as they are separated in space. A test of the same nature was performed by Grant and Pickup in the process of parametrical optimization of the model[64]. For this study, chains A and D of the *Barnase-Barstar* complex (PDB ID: 1X1X) were separated in steps of 0.1Å starting from the bound state to 15Å away. At 15Å separation, the monomers are practically free (completely unbound).

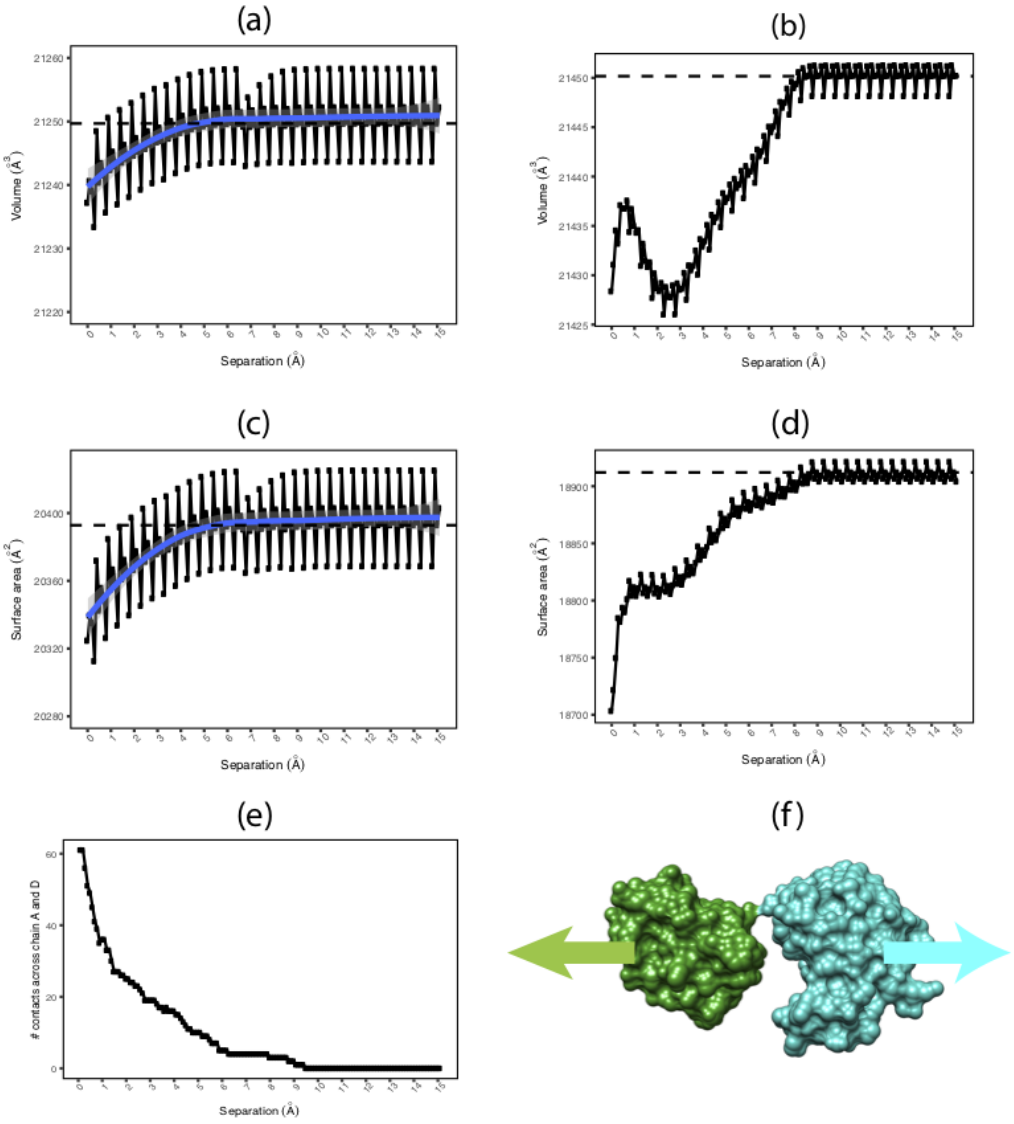

(a)
Volume (Å^3)
Separation (Å)
(b)
Volume (Å^3)
Separation (Å)
(c)
Surface area (Å^2)
Separation (Å)
(d)
Surface area (Å^2)
Separation (Å)
(e)
contacts across chain A and D
Separation (Å)
(f)

***Figure** 5.7.1: **Profile of the change in the van der Waals (vdW) volume and surface area.** Profile of the change in vdW volume of the Barsnase-Barstar complex as a function of the distance of separation of the monomers obtained (a) using the Gaussian model and (b) using the hard-sphere model. Profile of the change in vdW surface area obtained (c) using the Gaussian model trend and (d) using the hard-sphere model. The solid blue lines in (a) and (c) depict a non-linear fit to the profiles obtained using the Gaussian model in order to emphasize the overall smoothness of the trend. The vdW volume and surface area using the hard-sphere models were computed using 3V[183] with a probe of radius 0.0Å. (e) Change in the number of contacts, i.e. atom pairs from either monomer found to be within 4Å distance, as a function of the distance of separation of the monomers. (f) A cartoon representation of the setup in which the monomers of the Barnase-Barstar complex were separated for obtaining the above profiles of volume and surface area changes.*

The profiles are shown in **Figure** 5.7.1. Clearly, the change in vdwV/SA obtained using the Gaussian model (**Figure** 5.7.1 **(a, c)**) has an overall smooth trend (if one momentarily overlooks the periodic effects arising from use of the grid-based approach). In the completely unbound state, the volume and SA of the dimer is simply equal to the sum of these quantities for the individual monomers. On the contrary, the profile obtained using the hard-sphere model features some prominent bumps, discontinuities and noticeable regions of transitions. Between 1-2 Å of separation, the hard-sphere model yields an increase in the total volume of the complex and then it drops at around 2-3Å (**Figure** 5.7.1 **(b)**). Following this drop, it again increases monotonically till the total volume saturates at a value that equals

the sum of the volumes of the monomers. In terms of the surface area, there is a drastic initial increase till the monomers are separated by approximately 1Å after which there is a small discontinuity leading to a plateau in the profile (**Figure** 5.7.1 **(d)**). Once separated by approximately 3Å, the profile acquires a monotonically increasing trend till it saturates to a value that equals the sum of the surface areas of individual monomers (occurs at ~8Å separation).

Overall, the major difference in the profiles from the Gaussian and the hard-sphere model occur at small separations. This can be attributed to the method used by these approaches to treat the intersection or overlap volumes. To elaborate, the number of contacts between the two monomers (chains A and D) is plotted as a function of the distance of separation in **Figure** 5.7.1 **(e)**. A *contact* here is defined as a pair comprised of an atom from *Barnase* and an atom from *Barstar* whose centers are separated by no more than 4Å. As the monomers move farther apart, the number of contacts drops drastically at small separation distances and becomes zero at distances greater than 9Å. The region in between exhibits several abrupt transitions (2-9Å). The drastic and discontinuous drop in the region from 0-2Å explains the abrupt changes in the collective volumes of overlapping atoms at the interface. This is the cause for the bumps in the profile obtained using the hard-

sphere model at that region. This is also the region that emphasizes the ability of the Gaussian model to deliver smoothly-changing volumes. In the case of the Gaussian model, the volume of the overlapping regions (regardless of the order) between the atoms at the interface has a continuous expression (see Equation 33) which eventually renders a smooth change of the two quantities.

5.8 Gaussian model to compute solvent excluded volumes

The solvent excluded volume (SEV) and the corresponding surface area (SESA) are considered more faithful representations of the geometry of the solute-solvent interface than the van der Waals counterparts. These representations appropriately characterize those voids present in the solute structure, which are too small to fit a solvent molecule, as part of the solute phase. By virtue of this definition, SEVs are larger than the vdW volumes because the latter is a part of it. With our library of 74 proteins, SEVs were found to be 25% to 50% larger than their vdW volumes (average difference was ~38%).

In an attempt to enable the Gaussian model to deliver SEVs, Gallicchio *et. al.*[168] incorporated a modification in the Gaussian model. The modification involves augmenting the radius of all the solute atoms by an offset term R_{offset} so as to

account for the volume of crevices in the total volume. To enhance the physical appeal of this modification, an additional correction/modification was later incorporated [166]. The central idea of the modification was motivated by the fact that as R_{offset} increases the atomic radii in order to account for the interstitial crevices in the structure, it also causes the solvent exposed atoms to expand further into the solvent region. Therefore, the excess volume of the solvent exposed atoms can be discarded by computing the volume of only the solvent-exposed region of the atoms and subtracting it from their volume obtained with modified atomic radii (V_i^{offset}). This is illustrated in Figure 5.8.1**(a)** and Equation 43 expresses this correction term.

$$V_i = V_i^{offset} - V_i^{solvent-exposed\ region}$$

$$V_i = \frac{SA_i^{offset}\left(R_i + R_{offset}\right)}{3}\left(1 - \left(\frac{R_i}{R_i + R_{offset}}\right)^3\right) \tag{43}$$

The expression in Equation 43 ensures that the correction is only applied to the solvent exposed atoms with $SA_i^{offset} \neq 0$ computed using augmented atomic radii.

We added this modification in our implementation of the Gaussian model and evaluated a series of values of R_{offset} to find the value that yields the best agreement with the SEV computed using hard-sphere model with a solvent probe of radius 1.4Å. R_{offset} was systematically varied from 0.1-1.2Å and we found that 0.9Å gives the best agreement with a RMSRD of 2.3% (Figure 5.8.1 **(b, c)**). The goodness of the agreement is also confirmed by the quality of the linear regression which has a slope of 0.95, y-intercept of 326.22 and correlation of 1.00. In Table A. 1 we list the slope and intercept of the linear regression, correlation and the RMSRD for all the R_{offset} values in this range.

Although this empirical approach delivers a good agreement, the analysis so far as only emphasized on the numerical aspect. It was, therefore, important to examine if this modification is realistic in nature and if it retains the physical appeal of the Gaussian model. Two different approaches were used to address this - 1) if this modification truly accounts for the volume of the crevices in the structure and 2) if offers a physically meaningful description of the SEV.

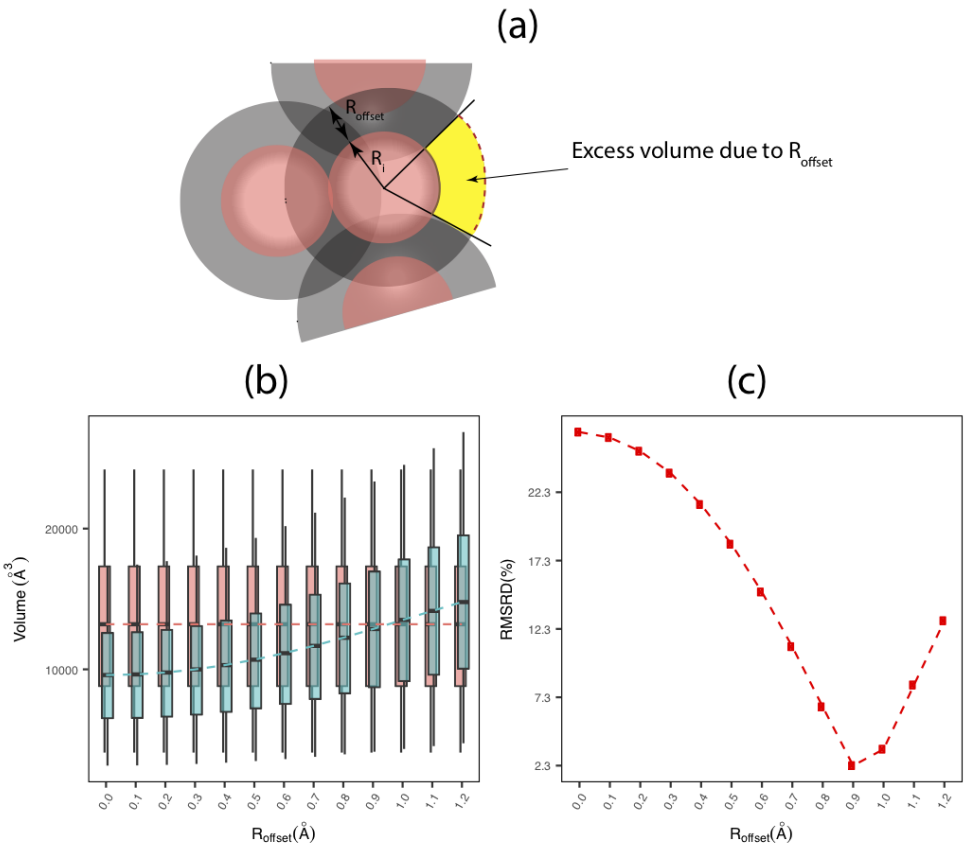

(a)
R_{offset}
R_i
Excess volume due to R_{offset}
(b)
Volume ($Å^3$)
20000
10000
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2
R_{offset}(Å)
(c)
RMSRD(%)
22.3
17.3
12.3
7.3
2.3
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2
R_{offset}(Å)

Figure 5.8.1: Optimization of R_{offset} input to the modified R_{offset} -based Gaussian model wrt the solvent excluded volume obtained using a hard sphere model. Distributions and relative percent deviations (RMSRD) are computed for the protonated and minimized crystal structures of 74 proteins. (a) Schematic showing the basis of the modified R_{offset} -based Gaussian model in which the excess volume of a solvent exposed atom (shown in yellow), obtained by augmenting its van der Waals radius by some R_{offset}, is subtracted out when the correction is applied. (b) Distribution of volume output by the modified R_{offset} -based Gaussian model (blue) computed with various R_{offset} values ranging from 0.0 through 1.2 Å, compared with the distribution of the hard-sphere solvent excluded volumes (pink) for the same set of proteins. Each distribution is represented by a boxplot (see Appendix A.12). The dashed lines connecting the medians of the boxes highlight the overall trend. (c) %RMSRD of the volume from the Gaussian model with respect to the solvent excluded volume as a function of R_{offset}.

5.8.1 Volume of Interstitial Regions:

A simple formula was used to derive the volume of the interstitial regions of solutes. By subtracting out the volume obtained with no R_{offset} (original model) from the volume obtained with a non-zero R_{offset}, the interstitial volume or $V_{interstitial}$ were obtained (Equation 44).

$$V_{interstitial}(R_{offset}) = Volume\Big|_{R_{offset}\neq 0} - Volume\Big|_{R_{offset}=0} \qquad (44)$$

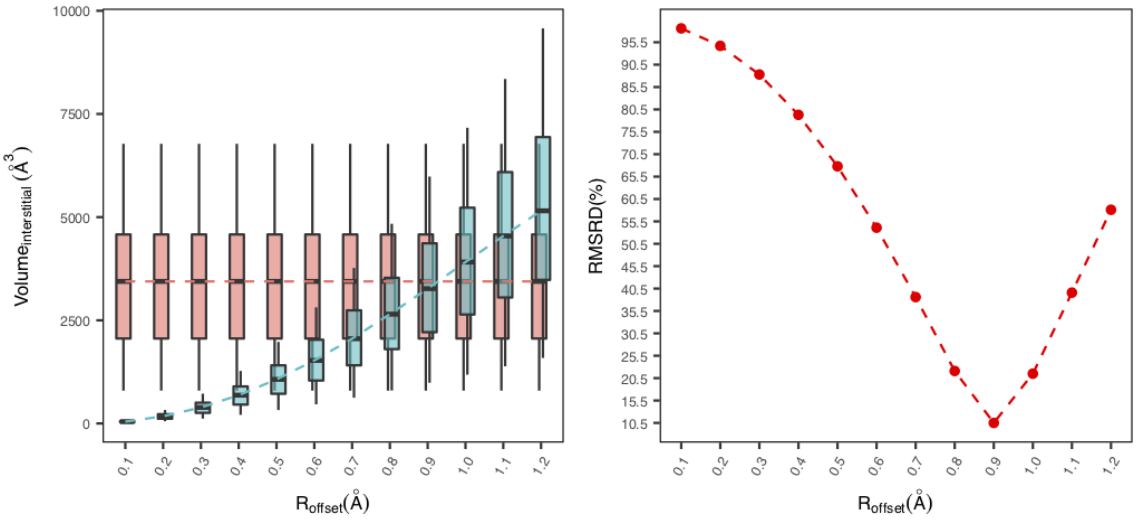


Figure 5.8.2: Comparison of the volume of the interstitial regions in the structure obtained using the modified R_{offset} -based Gaussian model and the hard-sphere model. Distributions and relative percent deviations (RMSRD) computed for the protonated and minimized crystal structures of 74 proteins. (b) Distribution of $\boldsymbol{Volume_{interstitial}}$ *computed using the modified R_{offset} -based Gaussian model (blue) computed with various R_{offset} values ranging from 0.0 through 1.2 Å, compared with the distribution of* $\boldsymbol{Volume_{interstitial}}$ *computed using the hard-sphere model by ProteinVolume[176] (pink). Each distribution is represented by a boxplot (see Appendix A.12). The dashed lines connecting the medians of the boxes highlight the overall trend. (b) %RMSRD of the* $\boldsymbol{Volume_{interstitial}}$ *from the Gaussian model with respect to the* $\boldsymbol{Volume_{interstitial}}$ *from hard-sphere model as a function of R_{offset}.*

By using the interstitial volumes obtained with the hard-sphere model using *ProteinVolume*[176] as a reference, the interstitial volumes computed using $\boldsymbol{R_{offset}}$ of 0.9Å were found to lie within 10.5%. In addition, a linear regression fit yielded a slope of ~0.81 and a correlation of ~0.99. That the other values of $\boldsymbol{R_{offset}}$ did not

perform as well as 0.9Å, as is evident from Figure 5.8.2, confirms that the numerical match obtained with this R_{offset} has a realistic foundation as well. In Table A. 2, the slope and intercept of the linear regression, the correlation and the RMSRD for all the R_{offset} values that were used are listed. The above calculations were done on the library of 74 proteins.

5.8.2 Physical Appeal of the R_{offset}-based Gaussian model.

In terms of physical appeal, the same case of *Barnase-Barstar* complex was used; separating the monomers from their bound state to a completely unbound state, to profile the change in volume. Except in this particular study, the vdW volume was replaced with the SEV from hard-sphere model and the volume from the modified R_{offset}-based Gaussian model. For the latter, we used R_{offset} of 0.9Å. The profiles are shown in **Figure** 5.8.3. Clear differences are visible in the trends obtained from the two models. That the profile of the volume from the modified $\mathrm{R_{offset}}$-based Gaussian model has a better physical foundation than the hard-sphere model is justified in the following two paragraphs.

The volume from the modified R_{offset}-based Gaussian model features a smooth monotonic decrease from an initial value to the value that equals the sum of the volumes of the individual monomers (**Figure** 5.8.3**(a)**). The inset in the plot shows the volume derived from equation 44 and shows that the excess volume computed by the $\mathsf{R}_{\mathsf{offset}}$-based Gaussian model (after the correction of the excess solvent exposed volume) monotonically and smoothly decreases as the separation increases. This smooth decrease can be better understood in terms of the change of average dielectric properties of the region between the monomers. As the monomers move apart, they gradually allow solvent molecules to occupy this region. But as the solvent molecules begin to enter the space between the interfaces, the interfacial residues from either monomer are expected to favorably interact with it to compensate for the loss of favorable interactions in the bound state. Consequently, the solvent molecules are not as mobile as their counterparts in the bulk and, therefore, tend to have a lower dielectric response, as has also been observed experimentally[53]. This is the foundation reflected in the Gaussian-based smooth dielectric model proposed by us[86].

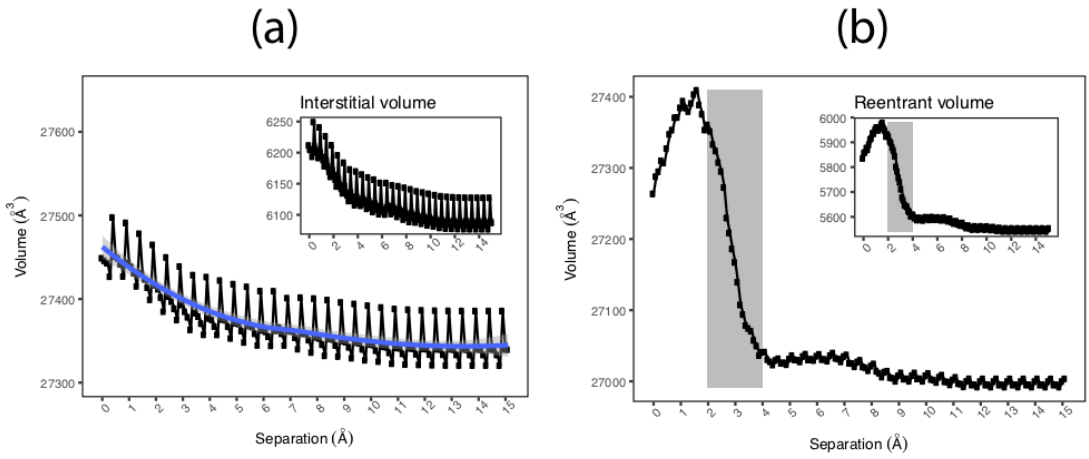


***Figure** 5.8.3: P**rofile of the change in the R_{offset} based Gaussian volume.** (a) Change in the volume of the Barsnase-Barstar complex output by the modified R_{offset}-based Gaussian model. The solid blue line depicts a smooth fit to emphasize the smooth trend. Inset: Difference of the volume output by the modified R_{offset}-based and the unmodified Gaussian model, that is supposed to depict the volume of solvent inaccessible crevices in the complex's structure, as a function of separation distance. (b) Change in the solvent excluded volume (SEV) of the complex computed using 3V[183] with a probe of radius 1.4Å as a function of the separation distance of the monomers. Inset: Volume of reentrant regions and solvent inaccessible crevices obtained by subtracting the van der Waals volume of the dimer from its SEV. The shaded region (gray) emphasizes the length scale of separation that is comparable to the diameter of the solvent probe (2.8Å).*

For the case of hard-sphere model, on the other hand, the volume increases from the initial value in the bound state till the separation is approximately 2Å. Subsequently, there is a drastic drop in the volume in the region from 2-4Å (shaded region in **Figure** 5.8.3**(b)**). The size of this region is typical of the solvent probe's

diameter and drop occurs because the concave reentrant surfaces, previously bounding the solvent inaccessible crevices at the interface between the monomers, disappear at this degree of separation. The inset plot in **Figure** 5.8.3**(b)** shows this loss of the solvent inaccessible volume bound by the reentrant surfaces. This volume is simply derived by subtracting the SEV of the dimer system from its vdW volume. The sudden loss of reentrant volume at the interfacial region implies that solvent molecules can enter this region and retain the dielectric response in bulk. Although, this model of dielectric distribution is conventional in PB modeling of solvated dimer systems, it fails to capture the possibility of interaction of the newly exposed interfacial residues with the solvent.

5.9 Limitations of the Gaussian model with large R_{offset}

There is a practical risk associated with using radius offsets comparable in magnitude to the radius of atoms (e.g. $\mathrm{R_{offset}}$ of 0.9 Å). The Gaussian model of Grant and Pickup was designed and optimized to deliver vdW volume and SA but only in the limit of weakly overlapping atoms[186, 187]. Thus, augmenting the atomic radius also increases the degree of overlap of atoms and this brings the Gaussian model very close or likely beyond its limit of applicability. With large overlaps and by virtue of

the Gaussian product theorem (Equation **4**), the volume of the overlapping regions is overestimated with respect to what a hard-sphere model would deliver with the same set of augmented radii. Mathematically, as the overlapping region of any two atoms grows in volume, the volume of the atom pair grows proportionally to the product of the volumes of the individual atoms (V^2, where V is the volume of one atom). Geometrically, however, if two atoms of similar volumes overlap significantly in space, the volume of the atom pair is proportional to the volume of the larger of the two atoms (or V). This fundamental problem can lead to errors in volume and SA estimates.

To test the effect of offsets, the van der Waals volume using the Gaussian model and the hard-sphere model were computed when both the models were provided with augmented atomic radii. This deliberately increased the degree of overlap of atoms due to their increased radii. By systematically varying $\mathsf{R_{offset}}$ from 0.0-1.2Å, their distribution was compared and the relative differences were measured (**Figure** 5.9.1). The overall trend indicates that as the value of $\mathsf{R_{offset}}$ is increased, the Gaussian and hard-sphere volumes start to deviate appreciably. Volumes obtained from the Gaussian model increase exponentially while volumes obtained using the hard-sphere model saturate after a certain point. With no offset, the

volumes from the two models deviate only by ~7% (the difference in the vdW volumes) but this increases to ~41% when the radii are augmented by an offset of 1.0Å and to ~55% when augmented by an offset of 1.2Å. This exponential deviation reflects the overestimation of the overlap volumes that is geometrically incorrect.

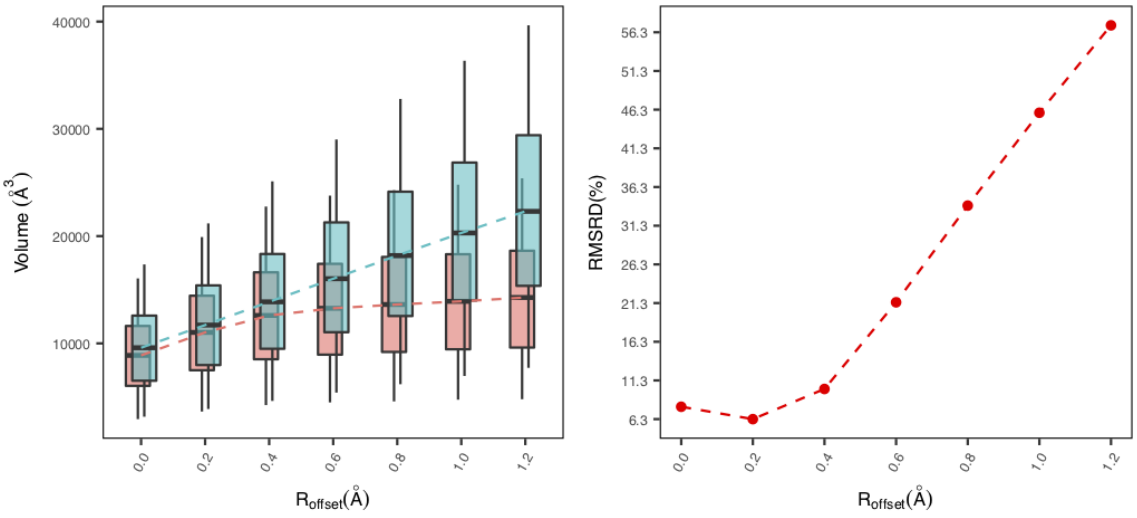


***Figure** 5.9.1: **Breakdown of the Gaussian model of molecular volume and surface area**. Comparison of the van der Waals volume from the R_{offset}-based Gaussian model (without correction of the excess solvent-exposed volume) and hard-sphere models when augmented radii for atoms are used. Distributions and percent deviations (RMSRD) are computed for the protonated and minimized crystal structures of 74 proteins. (a) Distribution of volume output by the R_{offset}-based Gaussian model (blue) computed with various R_{offset} values ranging from 0.0 through 1.2 Å, compared with the distribution of hard-sphere volumes (pink) computed using the same set of augmented radii. Each distribution is represented by a boxplot (see Appendix A.12). The dashed lines connecting the medians of the boxes highlight the overall trend. (b) %RMSRD of the volume obtained using the R_{offset}-based Gaussian model with respect to the volume output by the hard-sphere model as a function of R_{offset}.*

This aspect of the Gaussian model will pose methodological issues if it used to compute the solvent accessible surface area (SASA) of solutes using the definition used by the hard-sphere model. By that definition, SASA is essentially the van der Waal surface area obtained when the radius of each atom is augmented by the radius of the solvent probe (typically 1.4Å). But if an $\mathsf{R_{offset}}$ value of 1.4Å is used in order to obtain SASA using the Gaussian model, it will be asked to operate beyond its range of applicability. Though this idea was titillated by Grant and Pickup in their original work[64], Weiser *et. al.*[186] emphasized the issue with such an approach. Weiser *et. al.*[186] also discussed other parametrical modifications to obtain SASA but carefully described their limitations too.

5.10 Summary

This chapter presents a novel grid-based algorithm of identifying overlapping pairs of atoms in conjunction with the analytical approach of a Gaussian-based model[64] for computing MVs and surface areas(SAs). The primary motivation for this design was to integrate into *Delphi*[70], a new feature for determining non-polar parts of the free energy. This grid-based algorithm makes a simultaneous use of a cubic 3D grid-map constructed for *Delphi*'s finite-difference based operations and by

doing so, it incurs very little to no time in identifying the pairs of atoms that overlap in space. The validation of the grid-based algorithm in terms of the final volume/SA output, accuracy in identifying overlapping atom pairs and time-efficacy shows that the method is robust and credible for an integrated use with future versions of *Delphi* for MM/PBSA analyses. The integration of the Gaussian-based model of volume/SA with the Gaussian-based model of dielectric distribution of *Delphi*[86] also promotes a description of solvated biomolecular systems devoid of unphysical and discontinuous dielectric separation. This work brings us one step closer to having an integrated platform for MM/PBSA calculations using a physically appealing, surface-free approach to evaluate the thermodynamics of solvation, binding and folding/unfolding of proteins in the framework of implicit solvent models.

6 COMPENDIUM

Electrostatic and geometric properties, respectively, deliver the polar and the non-polar components of the solvation and binding free energy of a molecular system. The polar component signifies the stability of the molecule in a solvent bath (typically water) resulting from the balance between the intramolecular electrostatic interactions and molecule-solvent interactions. The non-polar component, on the other hand, signifies the work required to place a low-dielectric value cavity in the bulk of a polar, higher-dielectric solvent. By virtue of the state-function nature of the free energy, these components, add together to provide the energy of transfer of the molecule in question from gas-phase to the solvent.

This dissertation presents a detailed account of a Gaussian-based model of atoms designed to model electrostatic and geometric properties of biological molecules. The particular design of this model is constructed to work in conjunction with the Poisson-Boltzmann (PB) formalism of continuum electrostatic model. The design is motivated by the understanding that biomolecules and solvent constantly interact with each other and update their configurations and that these interactions are critical to its function and stability.

In the conventional setup, the solute region is depicted using a low dielectric value while the solvent region is assigned a higher dielectric value. The solute region is characterized by its constituent atoms represented by explicit positions, partial charges and radii (obtained from a force-field) while the solvent region is represented as a structureless dielectric continuum. A dielectric boundary is conventionally used to mark the separation of the two piecewise homogeneous dielectric regions. This is also referred to as the 2-dielectric model. But this arrangement overlooks the important solute-solvent interactions which are largely influenced by the density of atoms in the local environments. It also neglects the effect of the solvent molecules that find way into the interstitial regions in the solute's structure by treating them at the same footage as the bulk solvent. These elements can discount the critical solvation effects that can lead to a misinterpretation of the energies output by solving PBE.

In the Gaussian-based setup, the atoms of the solute are represented by smooth functions which symbolize the volume of a space that they occupy. By doing so, a smoother transition of dielectric values is established and a strict dielectric boundary, which imposes theoretical as well as numerical challenges, is discarded. With a matured and justified mathematical basis, the Gaussian-based model can be

used for estimating the polar component of the solvation free energy and determine geometrical attributes of the molecule which subsequently can be used to obtain the non-polar component as well.

In Chapter 1, the aforementioned concepts are extensively described. Since a major fraction of this work involved working with computational package called *Delphi*, its functionality, algorithm and applications are also introduced.

In Chapter 2, the mathematical details of the Gaussian-based atomic model and the derivation of a smooth Gaussian-based dielectric model from it are presented. The chapter provides a detailed understanding of the motivations of this model and illustrates its ability to resolve some key challenges retained in the conventional 2-dielectric setup. It also emphasizes the approach used to account for the contribution of the electrolyte/salt concentration in the Gaussian model. Through qualitative means mostly, the chapter offers a visual clarity of the Gaussian-based atomic and dielectric model and its ability to represent molecular systems and phenomena in a physically appealing way.

In Chapter 3, observations that support the idea of Gaussian-based model and its likes are presented and are carefully interpreted. Since the idea of the continuum electrostatics is to "average-out" the effect of the numerous solvent

degrees of freedom present in the explicit solvent representation, an attempt to assess observations made from explicit solvent simulations is discussed. It is shown that a simple averaging can ignore some very interesting features and structural attributes of the solvent and overlook the effect of local packing of atoms. The observations offer concrete evidence for an inhomogeneous dielectric model that provides differential treatment to the solvent as well as the solute depending largely on their local environments. Overall it presents a proof of concept of the Gaussian-based atomic model and its appropriate use for modeling a corresponding dielectric distribution.

Chapter 4 highlights the ability of the Gaussian-based model to faithfully reproduce the ensemble averaged solvation energy of a library of protein molecules using a single configuration respectively. The observations reflect the capacity of the Gaussian model to capture the effects of configurational flexibility on the ensemble averaged energies from energy minimized configurations only. In the discussions that follow, the underlying reasons for such are investigated and the conclusive importance of the dynamics of salt-bridges in a protein's structure is presented at good length. The outcome of the study presented in that chapter paves way for a

faster and a physically meaningful way of estimating ensemble average energy, which are more reasonable of a quantity to compare against experimental data.

Chapter 5 shifts the focus to the use of the Gaussian-based atomic model for determining the volume and surface area of a molecule. This transition is meant to highlight the versatility of this model by showing that it can similarly be used for estimating the non-polar component of solvation and binding free energies. Though not conceptually novel, the contents of the chapter are arranged around a novel grid-based algorithm of identifying overlapping atoms, which as an information is key to the computation of molecular volume and surface area. The motivation of this development was to expand the use of *Delphi* beyond the calculation of the electrostatic or polar components of solvation free energy. The ability introduced thereafter makes *Delphi*'s energy calculations more versatile and impactful as it can now estimate the total solvation free energy in a consistent manner.

Through the contents – the observations and their interpretations, presented here, the meaningfulness of the Gaussian-based atomic model for modeling electrostatic and geometric properties of biological molecule is demonstrated. These works also pave ways for future developments that are bound to be promising. Some of these works are currently underway.

APPENDICES

A.1 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.

A.2 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.3 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 (ρ_{SF}). This is done to fix the solute "volume" regardless of the 'ε_{ref}' value of the internal reference dielectric constant since the 'ε_{ref}' can influence the position of the dielectric-based surface but not that of a density-based surface.

(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 (equation 27).

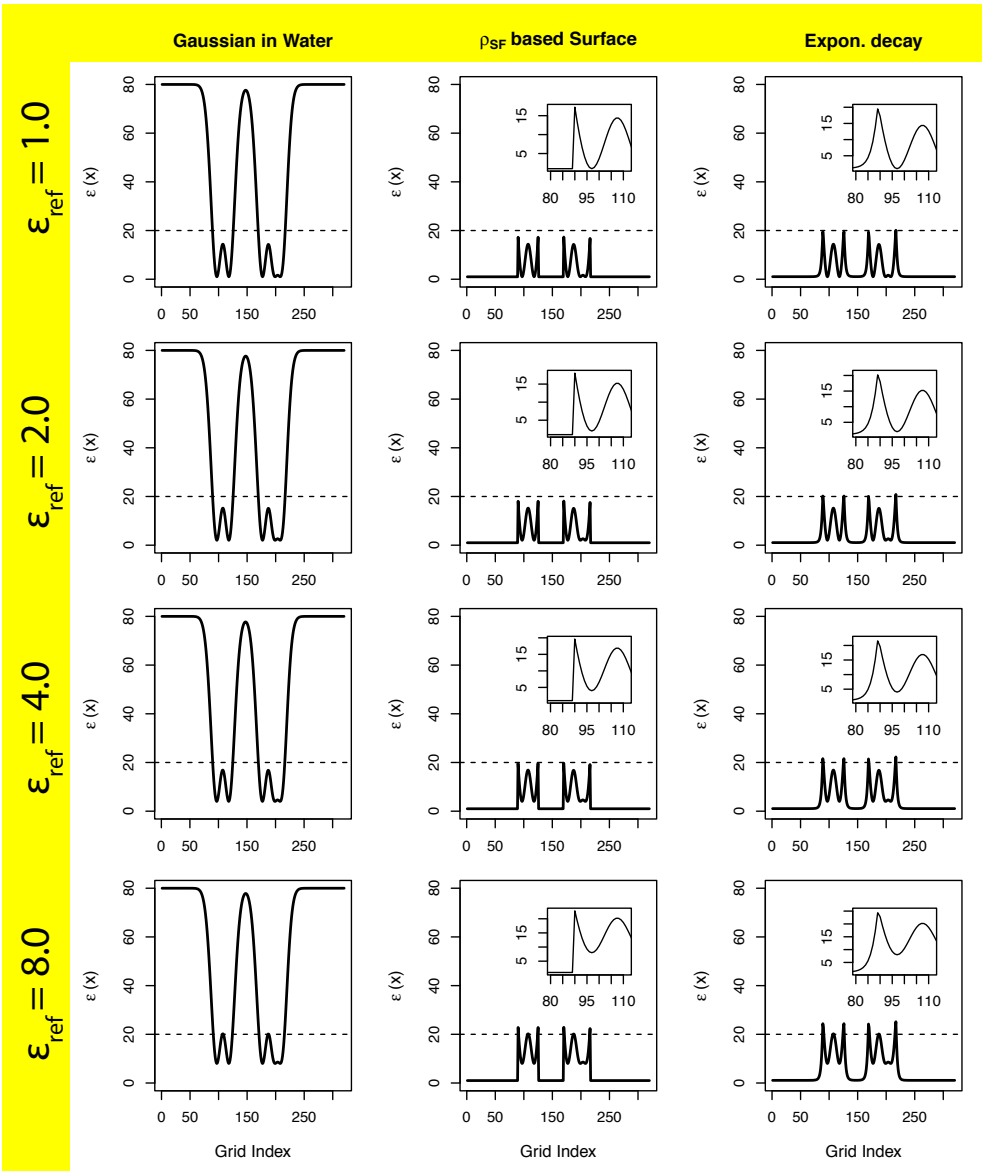

Gaussian in Water
ρSF based Surface
Expon. decay
εref = 1.0
εref = 2.0
εref = 4.0
εref = 8.0
ε (x)
Grid Index

Figure *A. 1: 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* ε*ref (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.*

Here, ε'(r) is the dielectric value of a 3D point when the solute is present in vacuum and ε(r) is the dielectric value assigned to that point when the protein's presence in solvent was modelled. 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 ρ_{SF}. The schematic for these modifications can be visualized in **Figure** A. 1.

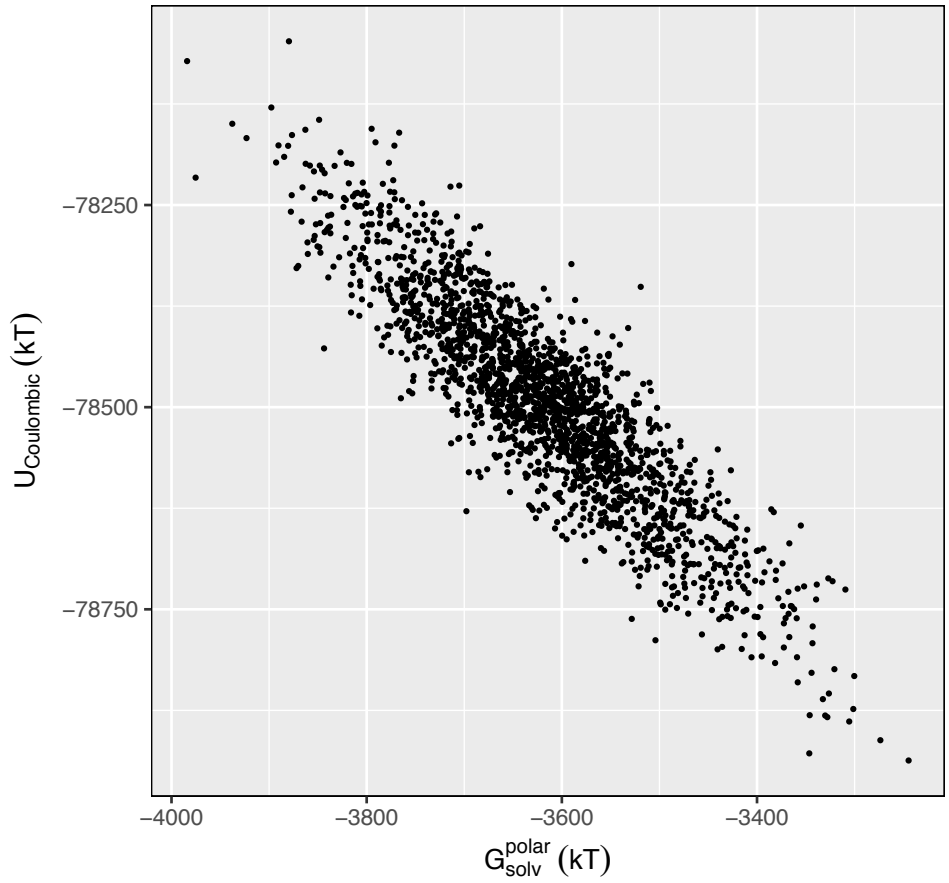


***Figure** A. 2: The inversely proportional relation of the Coulombic interaction energy and the Polar component of the solvation free energy. The data presented here pertains to the 2000 configurations of the protein 2NVH sampled from 20ns MD trajectory performed under NPT conditions. Both, the Coulombic energy ($U_{coulombic}$) and the polar solvation energy (G_{solv}^{polar}) were computed using Delphi using an internal dielectric value of 1 and solvent dielectric of 80.*

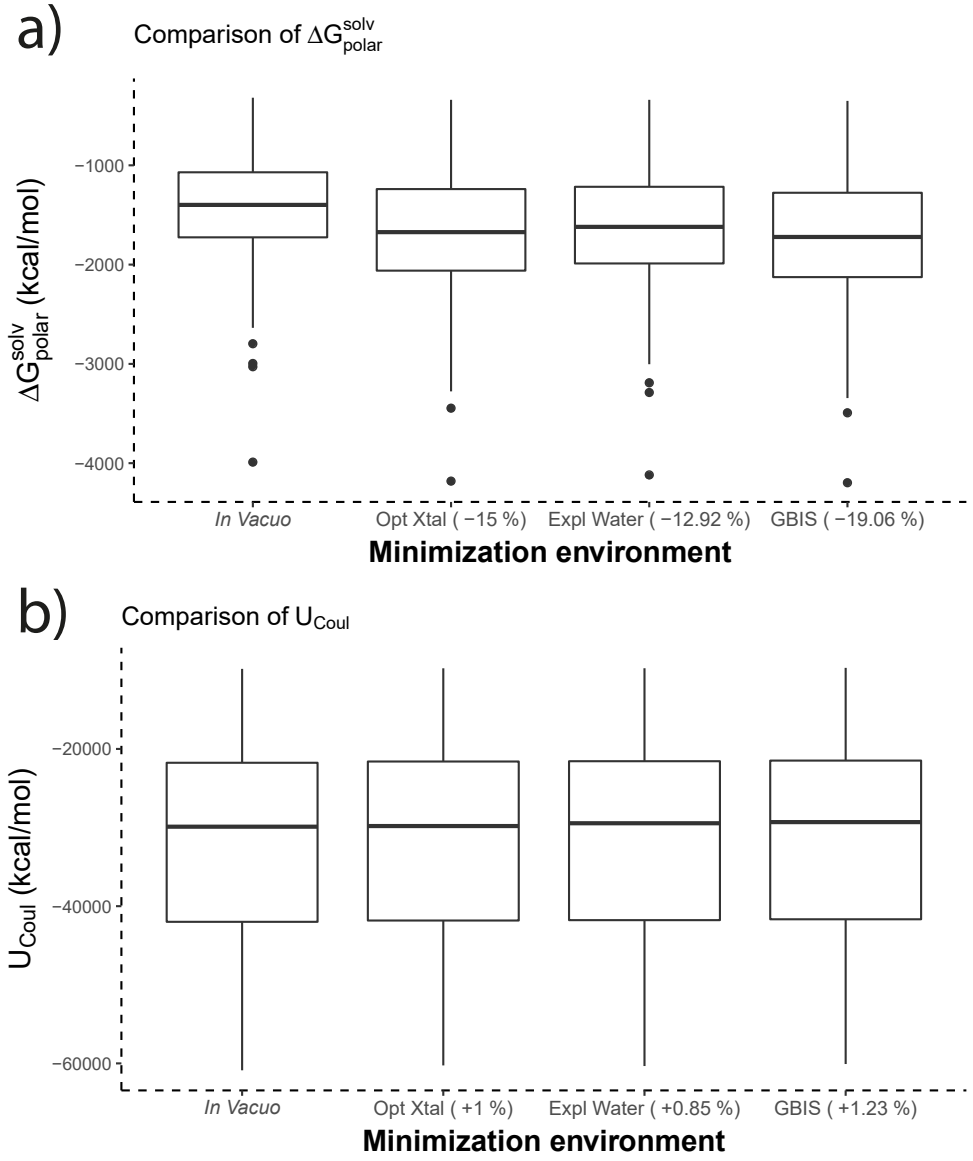


***Figure** A. 3: Comparison of (a) ΔG^{solv}_{polar} and (b) U_{coul} of the 74 proteins minimized in different environments. The percent difference of the mean of the corresponding energy components w.r.t to that of the in vacuo minimized structure are also noted in parentheses. Legend: "**In Vacuo**" – structure minimized in vacuum, "**Opt Xtal**" – optimized crystal structure, "**Expl Water**" – Explicit solvent (TIP3P) and "**GBIS**" – Generalized Born Implicit solvent.*

A.5 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.

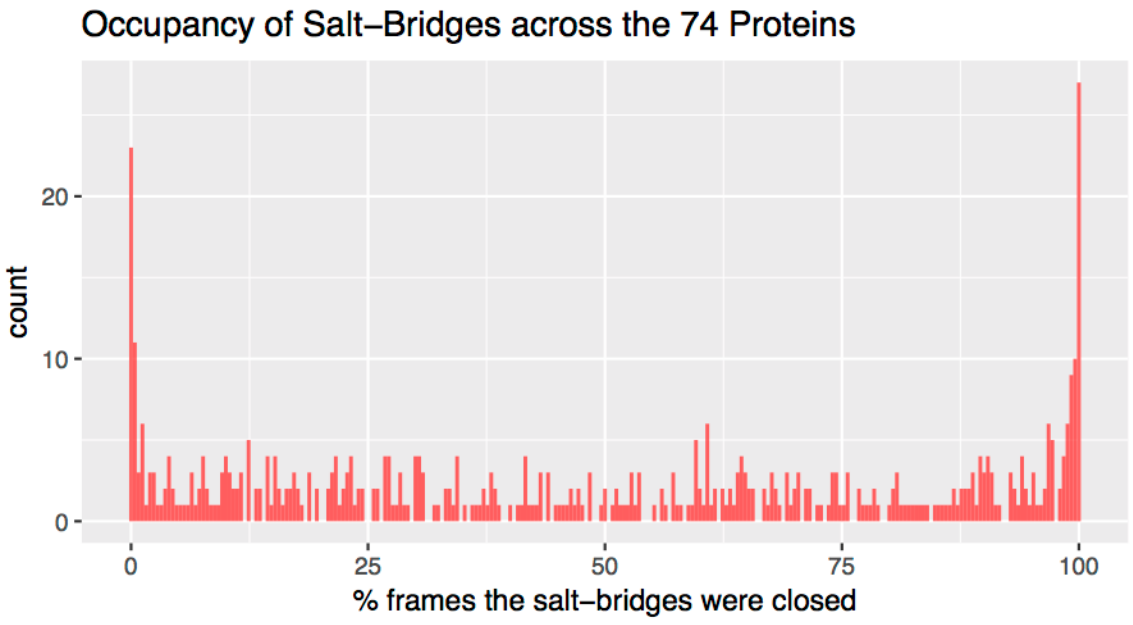


***Figure** A. 4: Histogram showing the distribution of occupancies of all the salt-bridges that were identified across all the 74 proteins.*

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** A. 4. 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).

A.6 Changing Polar solvation free energy with internal dielectric distribution

As one changes the protein dielectric (ϵ_{in} for the traditional 2-dielectric model or ϵ_{ref} for the Gaussian-based smooth dielectric model), the value of the ΔG_{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 ε_{in} values were set to 1, 2, 4 and 8 for the traditional dielectric model (TRAD) and the same values were used for ε_{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** A. 5**.**

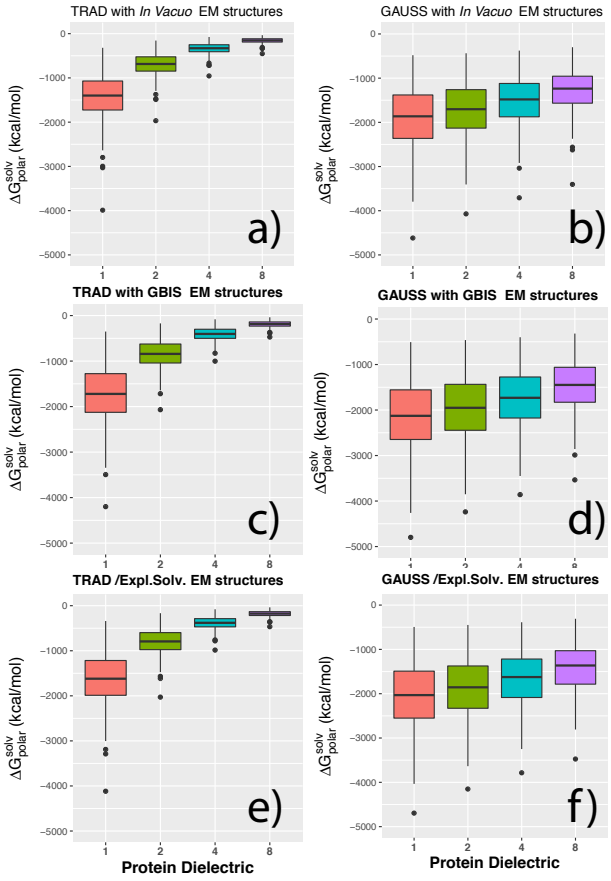


Figure *A. 5: Boxplots showing the distribution of* ΔG_{polar}^{solv} *from 74 proteins for each of the internal dielectric values (1,2,4,8) when applied using the traditional dielectric model (left) and the Gaussian-based dielectric model (right). The top panel (a, b) show the trend for in vacuo minimized structures and the middle (c, d) and bottom (e, f) show that for structures minimized in GBIS and explicit solvent, respectively.*

A.7 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 A. 6 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.

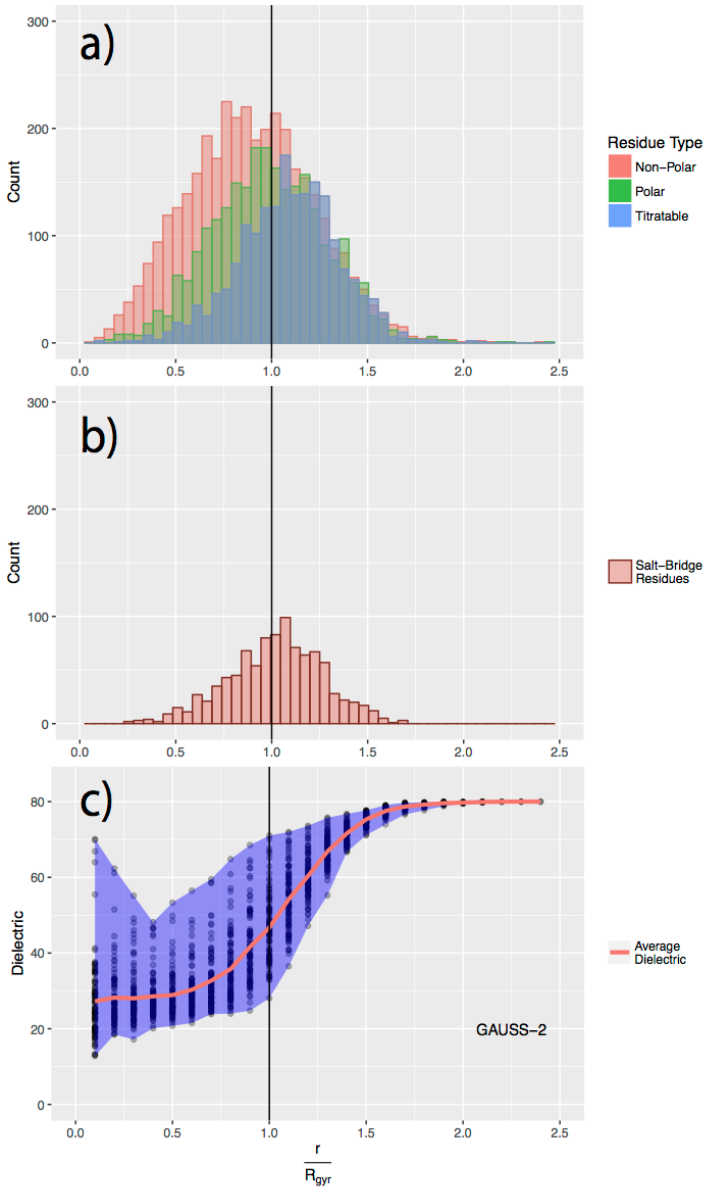


Figure *A. 6: Plots showing the average dielectric value (c) obtained from the Gaussian-based dielectric model features at different regions of the proteins in terms of the population of the non-polar, polar and titratable residues (a). In addition, the salt-bridge forming titratable residues are also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

A.8 Effect of grid-resolution on neighbors identified by the grid-based algorithm

The grid-based algorithm presented in the work is designed to identify the pair of atoms that overlap in space. That a neighbor is a "true" neighbor was verified by examining if the distance between their centers was less than or equal to the sum of their atomic radii plus some allowance (equation 12 in the main article). If a neighbor is found using both methods, we have a "True positive" case. But if it is not a neighbor based on the distance criteria but based on the grid-based algorithm, it is a "False Positive" case. The opposite renders a "False Negative" case.

For the purposes of validation of the algorithm, we computed the percentage of "False negative" and "False positive" cases. The former tells the number of neighbors wrongly discarded. The latter population will provide an assessment of the extra number of neighbors found using the grid-based algorithm compared to the distance-based method. The former can have an effect on the accuracy of the computed volumes and surface areas because of the absence of certain pairs which should have been actually present (see **Figure** 3(f) in the main article). The latter, however, only increases the number of neighbors that will be dealt with while computing volume overlaps and other related terms using the Gaussian model's

formulation. The more the number of extra pairs identified in the process, more the amount of time taken to complete all the computation.

Grid-resolution was found to have an effect on the number of "False negative" cases (see **Figure** A. 7) and is likely the cause for increased runtime (**Figure** 5.5.1 in the main article). With increase in the number of grids/Å, the population of the "False positive" cases grows.

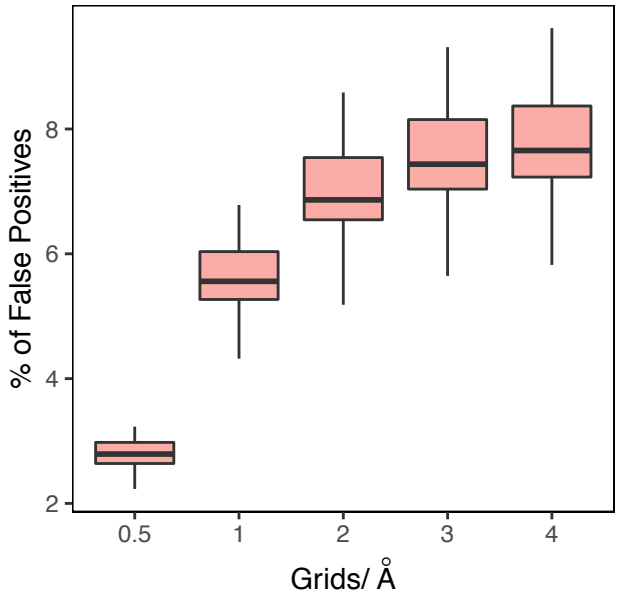


Figure *A. 7: Percentage of atom pairs estimated to be overlapping in space using the grid-based algorithm and not using the distance criteria (also termed as the "False positives") plotted as a function of the scale or grid-resolution (grids/Å). For each value of grids/Å, the distribution of the percent of "False positives" is obtained using the data from our library of 74 proteins. Each distribution is represented using a boxplot, the design of which represents the median and the inter-quartile range of the distribution.*

A.9 R_{offset} to obtain the best match with respect to the solvent excluded volumes (SEVs):

The value of R_{offset} input to the modified Gaussian model[64] was systematically varied from 0.0 to 1.2 Å in steps of 0.1Å to seek the one that offered the least percent difference from the SEV computed by the package 3V[183] using a solvent probe of radius 1.4Å. A non-zero offset is expected to increase the volume occupied by each atom artificially and by doing so, make the Gaussian model account for the small crevices and interstitial regions in the structure in the solute volume. For each R_{offset}, the volume computed by the modified Gaussian model was compared with SEV and the goodness of agreement was quantified by the slope and intercept of a linear regression, a correlation coefficient (R^2) and root mean square relative

difference (RMSRD). For each R_{offset}, the values of these quantities are listed in Table A. 1.

Table A. 1: Slope and intercept of the linear regression, correlation coefficient and %RMSRD quantifying the quality of the agreement provided by the volume computed by R_{offset} –based modified Gaussian model and the SEV computed using a hard-sphere probe of radius 1.4Å with 3V[183].

R_{offset} (Å)	Slope	Intercept (Å^3)	Correlation (R^2)	%RMSRD
0.0	0.70	355.27	0.999	26.7
0.1	0.71	356.56	0.999	26.3
0.2	0.72	359.51	0.999	25.3
0.3	0.73	362.77	0.999	23.7
0.4	0.75	365.06	0.999	21.4
0.5	0.78	365.63	0.999	18.5
0.6	0.82	362.45	0.999	15.0
0.7	0.86	358.84	0.999	11.0
0.8	0.90	342.49	0.999	6.6
0.9	0.95	326.22	1.000	2.3
1.0	1.00	308.74	1.000	3.5
1.1	1.05	292.86	1.000	8.2
1.2	1.10	279.82	1.000	12.9

A.10 R_{offset} to obtain the best match with respect to the volume of the interstitial regions in the solute

These R_{offset} values were also assessed in order to find the one that offered the best match with the volume of the interstitial regions, computed using the package

ProteinVolume[176]. This volume is somply the difference of the SEV and the van der Waals volume of a solute. For the Gaussian model, the equivalent is obtained by taking the difference of the volume obtained using the modified R_{offset}-based Gaussian model and the original unmodified Gaussian model. Using the same quantities used above, the goodness of agreement was quantified for each R_{offset}. For each R_{offset}, the values of these quantities are listed in Table A. 2.

Table A. 2: Slope and intercept of the linear regression, correlation coefficient and %RMSRD quantifying the quality of the agreement provided by the volume of the interstitial regions in the structure computed by by taking the difference of the volume obtained using the modified R_{offset}-based Gaussian model and the original unmodified Gaussian model and the same computed using the hard-sphere model with ProteinVolume[176].

R_{offset} (Å)	Slope	Intercept (Å^3)	Correlation (R^2)	%RMSRD
0.1	0.01	8.34	0.962	98.6
0.2	0.04	31.57	0.965	94.7
0.3	0.09	67.88	0.969	88.3
0.4	0.17	116.17	0.973	79.3
0.5	0.26	175.46	0.977	67.8
0.6	0.37	242.39	0.981	54.1
0.7	0.51	315.27	0.984	38.6
0.8	0.65	393.44	0.986	22.1
0.9	0.81	474.71	0.987	10.5
1.0	0.97	558.67	0.987	21.5
1.1	1.14	645.53	0.987	39.6
1.2	1.29	734.30	0.986	58.1

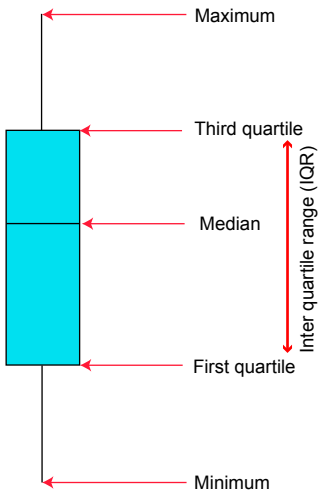
A.11 Root Mean Square Relative Difference (RMSRD)

The expression for the relative error between two sets of data, X and Y, relative to one of them (say X) with the same strength, N, is given by the following expression:

$$RMSRD = 100 * \sqrt{\frac{\sum_{i=1}^{N} \left(\frac{X_i - Y_i}{X_i}\right)^2}{N}}$$

A.12 Interpreting boxplots

Boxplots are a useful way of representing a distribution. By depicting the different quantiles for the underlying data, they provide a better sense of the distribution. Presenting the mean and the variance of a data assumes that the data is normally distributed, which, however, is not universal. Boxplots do not assume the category of the distribution of the data and can provide more information than just the mean and the variance. The figure below provides a guide to interpreting boxplots.


Maximum
Third quartile
Median
First quartile
Minimum
Inter quartile range (IQR)

A.14 Copyright permission for Chapter 5

This Agreement between Mr. Arghya Chakravorty ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4617230455008
License date	Jun 27, 2019
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Journal of Computational Chemistry
Licensed Content Title	A grid-based algorithm in conjunction with a gaussian-based model of atoms for describing molecular geometry
Licensed Content Author	Arghya Chakravorty, Emilio Gallicchio, Emil Alexov
Licensed Content Date	Jan 30, 2019
Licensed Content Volume	40
Licensed Content Issue	12
Licensed Content Pages	15
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Electronic
Portion	Full article
Will you be translating?	No
Title of your thesis / dissertation	MODELING ELECTROSTATICS AND GEOMETRICAL QUANTITIES IN MOLECULAR BIOPHYSICS USING A GAUSSIAN-BASED MODEL OF ATOMS
Expected completion date	Aug 2019
Expected size (number of pages)	240

Requestor Location	Mr. Arghya Chakravorty Clemson University 110 Kinard Lab CLEMSON, SC 29634 United States Attn: Mr. Arghya Chakravorty
Publisher Tax ID	EU826007151
Total	0.00 USD

REFERENCES

1. Arteca, G.A., C.T. Reimann, and O. Tapia, Proteins in vacuo: denaturing and folding mechanisms studied with computer-simulated molecular dynamics. Mass Spectrom Rev, 2001. 20(6): p. 402-22.
2. Stigter, D., D.O. Alonso, and K.A. Dill, Protein stability: electrostatics and compact denatured states. Proceedings of the National Academy of Sciences, 1991. 88(10): p. 4176-4180.
3. Spencer, D.S., et al., Effects of pH, Salt, and Macromolecular Crowding on the Stability of FK506-binding Protein: An Integrated Experimental and Theoretical Study. Journal of Molecular Biology, 2005. 351(1): p. 219-232.
4. Takahashi, T., Significant role of electrostatic interactions for stabilization of protein assemblies. Adv Biophys, 1997. 34: p. 41-54.
5. Dong, F. and H.-X. Zhou, Electrostatic contribution to the binding stability of protein-protein complexes. Proteins: Structure, Function, and Bioinformatics, 2006. 65(1): p. 87-102.
6. lee, L.P., Optimization of binding electrostatics: Charge complementarity in the barnase-barstar protein complex. Protein Science, 2001. 10(2): p. 362-377.
7. Kundrotas, P.J. and E. Alexov, Electrostatic Properties of Protein-Protein Complexes. Biophysical Journal, 2006. 91(5): p. 1724-1736.
8. Sheinerman, F., Electrostatic aspects of protein–protein interactions. Current Opinion in Structural Biology, 2000. 10(2): p. 153-159.
9. Gilson, M.K. and B. Honig, Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. Proteins: Structure, Function, and Genetics, 1988. 4(1): p. 7-18.
10. Strickler, S.S., et al., Protein Stability and Surface Electrostatics: A Charged Relationship†. Biochemistry, 2006. 45(9): p. 2761-2766.
11. Levy, Y., J.N. Onuchic, and P.G. Wolynes, Fly-Casting in Protein−DNA Binding: Frustration between Protein Folding and Electrostatics Facilitates Target Recognition. Journal of the American Chemical Society, 2007. 129(4): p. 738-739.

12. Spinozzi, F., P. Mariani, and M.G. Ortore, Proteins in binary solvents. Biophys Rev, 2016. 8(2): p. 87-106.
13. Onufriev, A.V. and E. Alexov, Protonation and pK changes in protein-ligand binding. Q Rev Biophys, 2013. 46(2): p. 181-209.
14. Petukh, M., S. Stefl, and E. Alexov, The role of protonation states in ligand-receptor recognition and binding. Curr Pharm Des, 2013. 19(23): p. 4182-90.
15. Talley, K. and E. Alexov, On the pH-optimum of activity and stability of proteins. Proteins, 2010. 78(12): p. 2699-706.
16. Alexov, E., Numerical calculations of the pH of maximal protein stability. The effect of the sequence composition and three-dimensional structure. Eur J Biochem, 2004. 271(1): p. 173-85.
17. Zhou, H.X., Brownian dynamics study of the influences of electrostatic interaction and diffusion on protein-protein association kinetics. Biophysical Journal, 1993. 64(6): p. 1711-1726.
18. Lopez-Garcia, J.J., J. Horno, and C. Grosse, Suspended particles surrounded by an inhomogeneously charged permeable membrane. Solution of the Poisson-Boltzmann equation by means of the network method. J Colloid Interface Sci, 2003. 268(2): p. 371-9.
19. Schmit, J.D., S. Whitelam, and K. Dill, Electrostatics and aggregation: How charge can turn a crystal into a gel. The Journal of Chemical Physics, 2011. 135(8): p. 085103.
20. Cen, G.-J., C.-C. Chang, and C.-Y. Wang, Optimizing electroosmotic flow in an annulus from Debye Hückel approximation to Poisson–Boltzmann equation. RSC Advances, 2017. 7(12): p. 7274-7286.
21. Roux, B. and T. Simonson, Implicit solvent models. Biophys Chem, 1999. 78(1-2): p. 1-20.
22. Tan, C., L. Yang, and R. Luo, How Well Does Poisson−Boltzmann Implicit Solvent Agree with Explicit Solvent? A Quantitative Analysis. The Journal of Physical Chemistry B, 2006. 110(37): p. 18680-18687.
23. Fogolari, F., A. Brigo, and H. Molinari, The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. Journal of Molecular Recognition, 2002. 15(6): p. 377-392.
24. Bashford, D. and D.A. Case, Generalized born models of macromolecular solvation effects. Annu Rev Phys Chem, 2000. 51: p. 129-52.

25. Chiba, M., D.G. Fedorov, and K. Kitaura, Polarizable continuum model with the fragment molecular orbital-based time-dependent density functional theory. Journal of Computational Chemistry, 2008. 29(16): p. 2667-2676.
26. Improta, R., et al., A state-specific polarizable continuum model time dependent density functional theory method for excited state calculations in solution. J Chem Phys, 2006. 125(5): p. 054103.
27. Feig, M. and C.L. Brooks, 3rd, Recent advances in the development and application of implicit solvent models in biomolecule simulations. Curr Opin Struct Biol, 2004. 14(2): p. 217-24.
28. Onufriev, A., Implicit Solvent Models in Molecular Dynamics Simulations: A Brief Overview. 2008. 4: p. 125-137.
29. Onufriev, A., D. Bashford, and D.A. Case, Modification of the Generalized Born Model Suitable for Macromolecules. The Journal of Physical Chemistry B, 2000. 104(15): p. 3712-3720.
30. Sharp, K.A. and B. Honig, Electrostatic interactions in macromolecules: theory and applications. Annu Rev Biophys Biophys Chem, 1990. 19: p. 301-32.
31. Rocchia, W., Poisson-boltzmann equation boundary conditions for biological applications. Mathematical and Computer Modelling, 2005. 41(10): p. 1109-1118.
32. Holst, M., The Poisson-Boltzmann Equation.
33. Decherchi, S., et al., Between algorithm and model: different Molecular Surface definitions for the Poisson-Boltzmann based electrostatic characterization of biomolecules in solution. Commun Comput Phys, 2013. 13: p. 61-89.
34. Connolly, M., Solvent-accessible surfaces of proteins and nucleic acids. Science, 1983. 221(4612): p. 709-713.
35. Richards, F.M., Areas, Volumes, Packing, and Protein Structure. Annual Review of Biophysics and Bioengineering, 1977. 6(1): p. 151-176.
36. Connolly, M.L., Analytical molecular surface calculation. Journal of Applied Crystallography, 1983. 16(5): p. 548-558.
37. Li, L., C. Li, and E. Alexov, On the Modeling of Polar Component of Solvation Energy using Smooth Gaussian-Based Dielectric Function. J Theor Comput Chem, 2014. 13(3): p. 1440002-1440016.

38. Cai, Q., et al., Dielectric boundary force in numerical Poisson–Boltzmann methods: Theory and numerical strategies. Chemical Physics Letters, 2011. 514(4-6): p. 368-373.
39. Geng, W. and G.W. Wei, Multiscale molecular dynamics using the matched interface and boundary method. J Comput Phys, 2011. 230(2): p. 435-457.
40. Klapper, I., et al., Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: Effects of ionic strength and amino-acid modification. Proteins: Structure, Function, and Genetics, 1986. 1(1): p. 47-59.
41. Nicholls, A. and B. Honig, A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. Journal of Computational Chemistry, 1991. 12(4): p. 435-445.
42. You, T.J. and S.C. Harvey, Finite element approach to the electrostatics of macromolecules with arbitrary geometries. Journal of Computational Chemistry, 1993. 14(4): p. 484-501.
43. Boschitsch, A.H., M.O. Fenley, and H.-X. Zhou, Fast Boundary Element Method for the Linear Poisson−Boltzmann Equation. The Journal of Physical Chemistry B, 2002. 106(10): p. 2741-2754.
44. Bordner, A.J. and G.A. Huber, Boundary element solution of the linear Poisson-Boltzmann equation and a multipole method for the rapid calculation of forces on macromolecules in solution. J Comput Chem, 2003. 24(3): p. 353-67.
45. Cai, Q., et al., Dielectric Boundary Forces in Numerical Poisson-Boltzmann Methods: Theory and Numerical Strategies. Chem Phys Lett, 2011. 514(4-6): p. 368-373.
46. Gilson, M.K., et al., Computation of electrostatic forces on solvated molecules using the Poisson-Boltzmann equation. The Journal of Physical Chemistry, 1993. 97(14): p. 3591-3600.
47. Xiao, L., et al., Electrostatic forces in the Poisson-Boltzmann systems. The Journal of Chemical Physics, 2013. 139(9): p. 094106.
48. Pang, X. and H.X. Zhou, Poisson-Boltzmann Calculations: van der Waals or Molecular Surface? Commun Comput Phys, 2013. 13(1): p. 1-12.
49. Lee, M.S., et al., New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. Journal of Computational Chemistry, 2003. 24(11): p. 1348-1356.

50. Wen, E.Z., et al., Enhanced ab initio protein folding simulations in Poisson-Boltzmann molecular dynamics with self-guiding forces. J Mol Graph Model, 2004. 22(5): p. 415-24.
51. Wang, C., L. Xiao, and R. Luo, Numerical interpretation of molecular surface field in dielectric modeling of solvation. J Comput Chem, 2017. 38(14): p. 1057-1070.
52. Wang, J., et al., Quantitative analysis of Poisson-Boltzmann implicit solvent in molecular dynamics. Phys Chem Chem Phys, 2010. 12(5): p. 1194-202.
53. Ikura, T., Y. Urakubo, and N. Ito, Water-mediated interaction at a protein–protein interface. Chemical Physics, 2004. 307(2-3): p. 111-119.
54. Shin, S. and A.P. Willard, Characterizing Hydration Properties Based on the Orientational Structure of Interfacial Water Molecules. J Chem Theory Comput, 2018.
55. Barnes, R., et al., Spatially Heterogeneous Surface Water Diffusivity around Structured Protein Surfaces at Equilibrium. J Am Chem Soc, 2017. 139(49): p. 17890-17901.
56. Yang, J., et al., Mapping Hydration Dynamics around a beta-Barrel Protein. J Am Chem Soc, 2017. 139(12): p. 4399-4408.
57. Zhao, S. and G.W. Wei, Matched interface and boundary (MIB) for the implementation of boundary conditions in high-order central finite differences. Int J Numer Methods Eng, 2009. 77(12): p. 1690-1730.
58. Xia, K., M. Zhan, and G.W. Wei, MIB method for elliptic equations with multi-material interfaces. J Comput Phys, 2011. 230(12): p. 4588-4615.
59. Chen, D., et al., MIBPB: a software package for electrostatic analysis. J Comput Chem, 2011. 32(4): p. 756-70.
60. Xia, K., M. Zhan, and G.W. Wei, MIB Galerkin method for elliptic interface problems. J Comput Appl Math, 2014. 272: p. 195-220.
61. Cheng, L.T., et al., Coupling the Level-Set Method with Molecular Mechanics for Variational Implicit Solvation of Nonpolar Molecules. J Chem Theory Comput, 2009. 5(2): p. 257-266.
62. Virtanen, J.J., et al., Modeling the hydration layer around proteins: HyPred. Biophys J, 2010. 99(5): p. 1611-9.
63. Li, L., et al., On the Dielectric "Constant" of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi. J Chem Theory Comput, 2013. 9(4): p. 2126-2136.

64. Grant, J.A. and B.T. Pickup, A Gaussian Description of Molecular Shape. The Journal of Physical Chemistry, 1995. 99(11): p. 3503-3510.
65. Roux, B.t. and T. Simonson, Implicit solvent models. Biophysical Chemistry, 1999. 78(1-2): p. 1-20.
66. Zhou, S., et al., Variational Implicit Solvation with Poisson–Boltzmann Theory. Journal of Chemical Theory and Computation, 2014. 10(4): p. 1454-1467.
67. Weeks, J.D., D. Chandler, and H.C. Andersen, Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids. The Journal of Chemical Physics, 1971. 54(12): p. 5237-5247.
68. Tan, C., Y.-H. Tan, and R. Luo, Implicit Nonpolar Solvent Models. The Journal of Physical Chemistry B, 2007. 111(42): p. 12263-12274.
69. Gallicchio, E., L.Y. Zhang, and R.M. Levy, The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. Journal of Computational Chemistry, 2002. 23(5): p. 517-529.
70. Li, L., et al., DelPhi: a comprehensive suite for DelPhi software and associated resources. BMC Biophysics, 2012. 5(1): p. 9.
71. Li, C., et al., DelPhi Suite: New Developments and Review of Functionalities. Journal of Computational Chemistry, 2019.
72. Wang, L., L. Li, and E. Alexov, pKa predictions for proteins, RNAs, and DNAs with the Gaussian dielectric function using DelPhi pKa. Proteins: Structure, Function, and Bioinformatics, 2015. 83(12): p. 2186-2197.
73. Peng, Y. and E. Alexov, Computational investigation of proton transfer, pKa shifts and pH-optimum of protein-DNA and protein-RNA complexes. Proteins: Structure, Function, and Bioinformatics, 2017. 85(2): p. 282-295.
74. Wang, L., M. Zhang, and E. Alexov, DelPhiPKa web server: predicting pKaof proteins, RNAs and DNAs. Bioinformatics, 2016. 32(4): p. 614-615.
75. Petukh, M., M. Li, and E. Alexov, Predicting Binding Free Energy Change Caused by Point Mutations with Knowledge-Modified MM/PBSA Method. PLoS Comput Biol, 2015. 11(7): p. e1004276.
76. Petukh, M., L. Dai, and E. Alexov, SAAMBE: Webserver to Predict the Charge of Binding Free Energy Caused by Amino Acids Mutations. International Journal of Molecular Sciences, 2016. 17(4): p. 547.

77. Peng, Y., et al., Predicting protein–DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver. Bioinformatics, 2018. 34(5): p. 779-786.
78. Getov, I., M. Petukh, and E. Alexov, SAAFEC: Predicting the Effect of Single Point Mutations on Protein Folding Free Energy Using a Knowledge-Modified MM/PBSA Approach. International Journal of Molecular Sciences, 2016. 17(4): p. 512.
79. Chakravorty, A., et al., A New DelPhi Feature for Modeling Electrostatic Potential around Proteins: Role of Bound Ions and Implications for Zeta-Potential. Langmuir, 2017. 33(9): p. 2283-2295.
80. Petukh, M., T. Kimmet, and E. Alexov, BION web server: predicting non-specifically bound surface ions. Bioinformatics, 2013. 29(6): p. 805-806.
81. Li, L., A. Chakravorty, and E. Alexov, DelPhiForce, a tool for electrostatic force calculations: Applications to macromolecular binding. J Comput Chem, 2017. 38(9): p. 584-593.
82. Li, L., et al., DelPhiForce web server: electrostatic forces and energy calculations and visualization. Bioinformatics, 2017. 33(22): p. 3661-3663.
83. Li, L., J. Alper, and E. Alexov, Multiscale method for modeling binding phenomena involving large objects: application to kinesin motor domains motion along microtubules. Sci Rep, 2016. 6: p. 23249.
84. Li, L., et al., Forces and Disease: Electrostatic force differences caused by mutations in kinesin motor domains can distinguish between disease-causing and non-disease-causing mutations. Scientific Reports, 2017. 7(1).
85. Chakravorty, A., et al., Gaussian-Based Smooth Dielectric Function: A Surface-Free Approach for Modeling Macromolecular Binding in Solvents. Frontiers in Molecular Biosciences, 2018. 5.
86. Li, L., et al., On the Dielectric "Constant" of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi. Journal of Chemical Theory and Computation, 2013. 9(4): p. 2126-2136.
87. Wang, W., et al., Biomolecular Simulations: Recent Developments in Force Fields, Simulations of Enzyme Catalysis, Protein-Ligand, Protein-Protein, and Protein-Nucleic Acid Noncovalent Interactions. Annual Review of Biophysics and Biomolecular Structure, 2001. 30(1): p. 211-243.

88. Sinha, S.K., S. Chakraborty, and S. Bandyopadhyay, Thickness of the Hydration Layer of a Protein from Molecular Dynamics Simulation. The Journal of Physical Chemistry B, 2008. 112(27): p. 8203-8209.
89. Li, L., et al., DelPhi: a comprehensive suite for DelPhi software and associated resources. BMC Biophys, 2012. 5: p. 9-20.
90. Hazra, T., et al., 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, 2019.
91. Hartley, R.W., Barnase and barstar: two small proteins to fold and fit together. Trends in Biochemical Sciences, 1989. 14(11): p. 450-454.
92. Janin, J., The kinetics of protein-protein recognition. Proteins: Structure, Function, and Genetics, 1997. 28(2): p. 153-161.
93. Hoefling, M. and K.E. Gottschalk, Barnase-Barstar: from first encounter to final complex. J Struct Biol, 2010. 171(1): p. 52-63.
94. Zhou, H.X., Disparate ionic-strength dependencies of on and off rates in protein-protein association. Biopolymers, 2001. 59(6): p. 427-33.
95. Bertonati, C., B. Honig, and E. Alexov, Poisson-Boltzmann calculations of nonspecific salt effects on protein-protein binding free energies. Biophys J, 2007. 92(6): p. 1891-9.
96. Jia, Z., et al., Treating ion distribution with Gaussian-based smooth dielectric function in DelPhi. Journal of Computational Chemistry, 2017: p. 1974–1979.
97. Bereiter-Hahn, J., Mechanics of crawling cells. Medical Engineering & Physics, 2005. 27(9): p. 743-753.
98. Jo, S., et al., CHARMM-GUI Membrane Builder for Mixed Bilayers and Its Application to Yeast Membranes. Biophysical Journal, 2009. 97(1): p. 50-58.
99. Peng, Y., et al., Predicting protein–DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver. Bioinformatics, 2017.
100. Peng, Y., et al., Predicting protein-DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver. Bioinformatics, 2017.
101. Pahari, S., et al., DelPhiPKa: Including salt in the calculations and enabling polar residues to titrate. Proteins: Structure, Function, and Bioinformatics, 2018. 86(12): p. 1277-1283.

102. Srinivasan, J., et al., Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices. Journal of the American Chemical Society, 1998. 120(37): p. 9401-9409.
103. Kollman, P.A., et al., Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res, 2000. 33(12): p. 889-897.
104. Grant, J.A., B.T. Pickup, and A. Nicholls, A smooth permittivity function for Poisson-Boltzmann solvation methods. Journal of Computational Chemistry, 2001. 22(6): p. 608-640.
105. Warshel, A. and S.T. Russell, Calculations of electrostatic interactions in biological systems and in solutions. Q Rev Biophys, 1984. 17(3): p. 283-422.
106. Warshel, A., et al., Modeling electrostatic effects in proteins. Biochim Biophys Acta, 2006. 1764(11): p. 1647-76.
107. Goh, G.B., B. García-Moreno E, and C.L. Brooks, The High Dielectric Constant of Staphylococcal Nuclease Is Encoded in Its Structural Architecture. Journal of the American Chemical Society, 2011. 133(50): p. 20072-20075.
108. Simonson, T. and C.L. Brooks, Charge Screening and the Dielectric Constant of Proteins: Insights from Molecular Dynamics. Journal of the American Chemical Society, 1996. 118(35): p. 8452-8458.
109. Luise, A., M. Falconi, and A. Desideri, Molecular dynamics simulation of solvated azurin: correlation between surface solvent accessibility and water residence times. Proteins, 2000. 39(1): p. 56-67.
110. Makarov, V.A., et al., Residence times of water molecules in the hydration sites of myoglobin. Biophys J, 2000. 79(6): p. 2966-74.
111. Quillin, M.L., P.T. Wingfield, and B.W. Matthews, Determination of solvent content in cavities in IL-1beta using experimentally phased electron density. Proceedings of the National Academy of Sciences, 2006. 103(52): p. 19749-19753.
112. Ponder, J.W. and D.A. Case, Force fields for protein simulations. Adv Protein Chem, 2003. 66: p. 27-85.
113. Jorgensen, W.L., et al., Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics, 1983. 79(2): p. 926-935.

114. Darden, T., D. York, and L. Pedersen, Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. The Journal of Chemical Physics, 1993. 98(12): p. 10089-10092.
115. Abraham, M.J., et al., GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX, 2015. 1-2: p. 19-25.
116. Van Der Spoel, D., et al., GROMACS: fast, flexible, and free. J Comput Chem, 2005. 26(16): p. 1701-1718.
117. Jorgensen, W.L. and J. Tirado-Rives, The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. Journal of the American Chemical Society, 1988. 110(6): p. 1657-1666.
118. Schoenborn, B.P., A. Garcia, and R. Knott, Hydration in protein crystallography. Prog Biophys Mol Biol, 1995. 64(2-3): p. 105-19.
119. Hubbard, S.J. and J.M. Thornton, 'NACCESS', computer program. 1993.
120. Yeh, Y.-l. and C.-Y. Mou, Orientational Relaxation Dynamics of Liquid Water Studied by Molecular Dynamics Simulation. The Journal of Physical Chemistry B, 1999. 103(18): p. 3699-3705.
121. Chakravorty, A., et al., 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, 2018. 14(2): p. 1020-1032.
122. Shivakumar, D., Y. Deng, and B. Roux, Computations of Absolute Solvation Free Energies of Small Molecules Using Explicit and Implicit Solvent Model. Journal of Chemical Theory and Computation, 2009. 5(4): p. 919-930.
123. Jayaram, B., et al., Free energy calculations of ion hydration: an analysis of the Born model in terms of microscopic simulations. The Journal of Physical Chemistry, 1989. 93(10): p. 4320-4327.
124. Nicholls, A., et al., Predicting small-molecule solvation free energies: an informal blind test for computational chemistry. J Med Chem, 2008. 51(4): p. 769-779.
125. Im, W., D. Beglov, and B. Roux, Continuum Solvation Model: computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation. Computer Physics Communications, 1998. 111(1-3): p. 59-75.

126. Song, X., An inhomogeneous model of protein dielectric properties: Intrinsic polarizabilities of amino acids. The Journal of Chemical Physics, 2002. 116(21): p. 9359-9363.
127. Simonson, T. and D. Perahia, Internal and interfacial dielectric properties of cytochrome c from molecular dynamics in aqueous solution. Proc Natl Acad Sci U S A, 1995. 92(4): p. 1082-1086.
128. Voges, D. and A. Karshikoff, A model of a local dielectric constant in proteins. The Journal of Chemical Physics, 1998. 108(5): p. 2219-2227.
129. Arnold, G.E. and R.L. Ornstein, An evaluation of implicit and explicit solvent model systems for the molecular dynamics simulation of bacteriophage T4 lysozyme. Proteins: Structure, Function, and Genetics, 1994. 18(1): p. 19-33.
130. Petukh, M., L. Dai, and E. Alexov, SAAMBE: Webserver to Predict the Charge of Binding Free Energy Caused by Amino Acids Mutations. Int J Mol Sci, 2016. 17(4): p. 547-558.
131. Wang, L., et al., Using DelPhi Capabilities to Mimic Protein's Conformational Reorganization with Amino Acid Specific Dielectric Constants. Communications in Computational Physics, 2015. 13(01): p. 13-30.
132. Berman, H.M., et al., The Protein Data Bank. Nucleic Acids Res, 2000. 28(1): p. 235-242.
133. Ponder, J.W. and D.A. Case, Force Fields for Protein Simulations. 2003. 66: p. 27-85.
134. Onufriev, A., D. Bashford, and D.A. Case, Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins: Structure, Function, and Bioinformatics, 2004. 55(2): p. 383-394.
135. Swanson, J.M., R.H. Henchman, and J.A. McCammon, Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. Biophysical Journal, 2004. 86(1 Pt 1): p. 67-74.
136. Kumari, R., R. Kumar, and A. Lynn, g_mmpbsa—A GROMACS Tool for High-Throughput MM-PBSA Calculations. Journal of Chemical Information and Modeling, 2014. 54(7): p. 1951-1962.
137. Genheden, S. and U. Ryde, The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov, 2015. 10(5): p. 449-461.

138. Mukhopadhyay, A., et al., Introducing Charge Hydration Asymmetry into the Generalized Born Model. Journal of Chemical Theory and Computation, 2014. 10(4): p. 1788-1794.
139. Onufriev, A.V. and B. Aguilar, Accuracy of continuum electrostatic calculations based on three common dielectric boundary definitions. Journal of Theoretical and Computational Chemistry, 2014. 13(03): p. 1440006.
140. Alexov, E., Role of the protein side-chain fluctuations on the strength of pair-wise electrostatic interactions: comparing experimental with computed pK(a)s. Proteins, 2003. 50(1): p. 94-103.
141. Kumar, S. and R. Nussinov, Fluctuations in ion pairs and their stabilities in proteins. Proteins: Structure, Function, and Bioinformatics, 2001. 43(4): p. 433-454.
142. Jelesarov, I. and A. Karshikoff, Defining the role of salt bridges in protein stability. Methods Mol Biol, 2009. 490: p. 227-260.
143. Simonson, T. and D. Perahia, Polar fluctuations in proteins: molecular-dynamic studies of cytochrome c in aqueous solution. Faraday Discuss, 1996(103): p. 71-90.
144. Chakravorty, A., E. Gallicchio, and E. Alexov, A grid-based algorithm in conjunction with a gaussian-based model of atoms for describing molecular geometry. Journal of Computational Chemistry, 2019.
145. Levy, R.M. and E. Gallicchio, COMPUTER SIMULATIONS WITH EXPLICIT SOLVENT: Recent Progress in the Thermodynamic Decomposition of Free Energies and in Modeling Electrostatic Effects. Annual Review of Physical Chemistry, 1998. 49(1): p. 531-567.
146. Fink, A.L., Protein aggregation: folding aggregates, inclusion bodies and amyloid. Folding and Design, 1998. 3(1): p. R9-R23.
147. Dill, K.A., Dominant forces in protein folding. Biochemistry, 2002. 29(31): p. 7133-7155.
148. Frye, K.J. and C.A. Royer, Probing the contribution of internal cavities to the volume change of protein unfolding under pressure. Protein Science, 1998. 7(10): p. 2217-2222.
149. Roche, J., et al., Cavities determine the pressure unfolding of proteins. Proceedings of the National Academy of Sciences, 2012. 109(18): p. 6945-6950.

150. Hermann, R.B., Theory of hydrophobic bonding. II. Correlation of hydrocarbon solubility in water with solvent cavity surface area. The Journal of Physical Chemistry, 1972. 76(19): p. 2754-2759.
151. Tannor, D.J., et al., Accurate First Principles Calculation of Molecular Charge Distributions and Solvation Energies from Ab Initio Quantum Mechanics and Continuum Dielectric Theory. Journal of the American Chemical Society, 1994. 116(26): p. 11875-11882.
152. Simonson, T. and A.T. Bruenger, Solvation Free Energies Estimated from Macroscopic Continuum Theory: An Accuracy Assessment. The Journal of Physical Chemistry, 1994. 98(17): p. 4683-4694.
153. Chothia, C., Hydrophobic bonding and accessible surface area in proteins. Nature, 1974. 248(5446): p. 338-339.
154. Gelles, J. and M.H. Klapper, Pseudo-dynamic contact surface areas: estimation of apolar bonding. Biochim Biophys Acta, 1978. 533(2): p. 465-77.
155. Chiche, L., et al., Protein model structure evaluation using the solvation free energy of folding. Proc Natl Acad Sci U S A, 1990. 87(8): p. 3240-3.
156. Reynolds, J.A., D.B. Gilbert, and C. Tanford, Empirical correlation between hydrophobic free energy and aqueous cavity surface area. Proc Natl Acad Sci U S A, 1974. 71(8): p. 2925-7.
157. Lum, K., D. Chandler, and J.D. Weeks, Hydrophobicity at Small and Large Length Scales. The Journal of Physical Chemistry B, 1999. 103(22): p. 4570-4577.
158. Gallicchio, E., M.M. Kubo, and R.M. Levy, Enthalpy−Entropy and Cavity Decomposition of Alkane Hydration Free Energies: Numerical Results and Implications for Theories of Hydrophobic Solvation. The Journal of Physical Chemistry B, 2000. 104(26): p. 6271-6285.
159. Mobley, D.L., et al., Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. Journal of Chemical Theory and Computation, 2009. 5(2): p. 350-358.
160. Kang, Y.K., G. Nemethy, and H.A. Scheraga, Free energies of hydration of solute molecules. 1. Improvement of the hydration shell model by exact computations of overlapping volumes. The Journal of Physical Chemistry, 1987. 91(15): p. 4105-4109.

161. Wagoner, J.A. and N.A. Baker, Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms. Proc Natl Acad Sci U S A, 2006. 103(22): p. 8331-6.
162. Hummer, G., Hydrophobic Force Field as a Molecular Alternative to Surface-Area Models. Journal of the American Chemical Society, 1999. 121(26): p. 6299-6305.
163. Ferrara, P., J. Apostolakis, and A. Caflisch, Evaluation of a fast implicit solvent model for molecular dynamics simulations. Proteins: Structure, Function, and Genetics, 2002. 46(1): p. 24-33.
164. Levy, R.M., et al., On the Nonpolar Hydration Free Energy of Proteins: Surface Area and Continuum Solvent Models for the Solute−Solvent Interaction Energy. Journal of the American Chemical Society, 2003. 125(31): p. 9523-9530.
165. Chen, J. and C.L. Brooks Iii, Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. Phys. Chem. Chem. Phys., 2008. 10(4): p. 471-481.
166. Gallicchio, E., K. Paris, and R.M. Levy, The AGBNP2 Implicit Solvation Model. Journal of Chemical Theory and Computation, 2009. 5(9): p. 2544-2564.
167. Barone, V., M. Cossi, and J. Tomasi, A new definition of cavities for the computation of solvation free energies by the polarizable continuum model. The Journal of Chemical Physics, 1997. 107(8): p. 3210-3221.
168. Gallicchio, E. and R.M. Levy, AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. Journal of Computational Chemistry, 2004. 25(4): p. 479-499.
169. Eisenberg, D. and A.D. McLachlan, Solvation energy in protein folding and binding. Nature, 1986. 319(6050): p. 199-203.
170. Sitkoff, D., K.A. Sharp, and B. Honig, Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. The Journal of Physical Chemistry, 1994. 98(7): p. 1978-1988.
171. Wang, J., et al., Solvation Model Based on Weighted Solvent Accessible Surface Area. The Journal of Physical Chemistry B, 2001. 105(21): p. 5055-5067.

172. Weiser, J., P.S. Shenkin, and W.C. Still, Approximate solvent-accessible surface areas from tetrahedrally directed neighbor densities. Biopolymers, 1999. 50(4): p. 373-80.
173. Cazals, F., H. Kanhere, and S. Loriot, Computing the volume of a union of balls. ACM Transactions on Mathematical Software, 2011. 38(1): p. 1-20.
174. Willard, L., et al., VADAR: a web server for quantitative evaluation of protein structure quality. Nucleic Acids Res, 2003. 31(13): p. 3316-9.
175. Kleywegt, G.J. and T.A. Jones, Detection, delineation, measurement and display of cavities in macromolecular structures. Acta Crystallographica Section D Biological Crystallography, 1994. 50(2): p. 178-185.
176. Chen, C.R. and G.I. Makhatadze, ProteinVolume: calculating molecular van der Waals and void volumes in proteins. BMC Bioinformatics, 2015. 16(1).
177. Mattson, W. and B.M. Rice, Near-neighbor calculations using a modified cell-linked list method. Computer Physics Communications, 1999. 119(2-3): p. 135-148.
178. Hess, B., et al., GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. Journal of Chemical Theory and Computation, 2008. 4(3): p. 435-447.
179. Verlet, L., Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. Physical Review, 1967. 159(1): p. 98-103.
180. Li, W.-Q., et al., Comparison research on the neighbor list algorithms: Verlet table and linked-cell. Computer Physics Communications, 2010. 181(10): p. 1682-1686.
181. Páll, S. and B. Hess, A flexible algorithm for calculating pair interactions on SIMD architectures. Computer Physics Communications, 2013. 184(12): p. 2641-2650.
182. Yip, V. and R. Elber, Calculations of a list of neighbors in Molecular Dynamics simulations. Journal of Computational Chemistry, 1989. 10(7): p. 921-927.
183. Voss, N.R. and M. Gerstein, 3V: cavity, channel and cleft volume calculator and extractor. Nucleic Acids Research, 2010. 38(Web Server): p. W555-W562.
184. Mitternacht, S., FreeSASA: An open source C library for solvent accessible surface area calculations. F1000Res, 2016. 5: p. 189.

185. Schaefer, M. and M. Karplus, A Comprehensive Analytical Treatment of Continuum Electrostatics. The Journal of Physical Chemistry, 1996. 100(5): p. 1578-1599.
186. Weiser, J., P.S. Shenkin, and W.C. Still, Optimization of Gaussian surface calculations and extension to solvent-accessible surface areas. Journal of Computational Chemistry, 1999. 20(7): p. 688-703.
187. Zhang, B., et al., Efficient gaussian density formulation of volume and surface areas of macromolecules on graphical processing units. Journal of Computational Chemistry, 2017. 38(10): p. 740-752.