MATHEMATICAL MODELING AND COMPUTATION OF MOLECULAR SOLVATION AND BINDING

By

Bao Wang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Applied Mathematics - Doctor of Philosophy

2016

ABSTRACT

MATHEMATICAL MODELING AND COMPUTATION OF MOLECULAR SOLVATION AND BINDING

By

Bao Wang

This dissertation contains a couple of results on biophysics modeling and computation, ranging from solvated molecular conformation modeling to molecular solvation and binding modeling in the solvent environment.

- We study the solvent excluded surface in Eulerian representation, provide the surface area and enclosed volume calculation, the molecular topological analysis is also addressed. We further analyze the electrostatic for the solvated molecules with the Eulerian solvent excluded surface. We show that our surface is analytical without any numerical approximation.

- We study the coarse grid Poisson Boltzmann solver. Our software enables extremely accurate numerical solution to the Poisson Boltzmann equation even at very large grid spacing. As a consequence, our software provides a reliable electrostatic calculation for the solvation and protein ligand binding related problem.

- We study the blind solvation free energy prediction problem. A hybrid of physical and statistical protocol is proposed for highly accurate solvation free energy prediction. Furthermore, to mediate the force field parametrization influence on the solvation free energy prediction, we propose a learning to rank based solvation free energy prediction paradigm.

- We explore the protein ligand binding free energy prediction and docking scoring via

the learning to rank approach. In which a learn to rank based scoring function is proposed for accurate protein ligand binding scoring.

ACKNOWLEDGMENTS

I would like to thank Professor Yongluo Cao for recommending and Professor Zhengfang Zhou for admitting me to this Ph. D. program.

My sincere thanks go to my advisor, Professor Guowei Wei, for the academic guidance and financial support. His own scientific work and insight on many of the related problems has a tremendous influence on me and has shaped the outcome of this thesis.

I want to thank Professor Yiying Tong for many stimulating discussions and encouragements during my thesis work.

I also thank Professors Peter Bates and Moxun Tang for serving on my thesis committee.

My special thank to Professor Andrea L. Bertozzi for many helpful discussions and offering a great academic position at the University of California, Los Angeles. I thank Professor Alexandre J. Chorin for the academic help in my graduate studies as well.

Further, I would thank many collaborators and friends, Zixuan Cang, Yin Cao, Chi Jin, Kedi Wu, Menglun Wang, Weicong Zhou, Dr. Wenbin Li, Professor Ming Yan, and many others. I also appreciate all the help from Metin Ozsarfati.

Last but not least, I would like to thank Huimin Yan and my parents Gaoming Wang and Taiyu Jiang for their love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In numerous physical, chemical, and biological applications, we encounter the solvation problem, in which the solvation effects play a crucial role in the whole process. These includes chemical reactions, ion channel permeations, protein ligand binding, electron transfer, signal transduction, DNA specification, transcription, post-transcription modification, gene expression, protein synthesis, etc. [228, 60, 204, 228, 60, 140, 204, 228, 62, 266]. There has been a long history of efforts paid to study the solvation effects. Many successful models have been developed for addressing the solvation effects in different applications. Naturally, based on the physical modeling of the solvent, these models are broadly classified into three categories. Molecular mechanics modeling of the solvent yields the explicit solvent models [36, 69, 188]; statistical mechanics description leads to the integral equation solvent models [91, 119, 113, 67, 110, 203]; continuum mechanics description gives the implicit solvent models [233, 232, 199, 221, 94, 296, 219]. Lately, more and more attention has been drawn for studying the solvation effects, especially for the biological and pharmaceutical sciences.

From the application point of view, one of our ultimate goals is to have a better understanding of the solvation problems and the related issues, e.g., the protein ligand binding phenomena. More importantly, the models and numerical methods developed should meet the experiments, such as the methodologies developed are capable to provide accurate molecular solvation free energy and protein ligand binding free energy prediction. This in turn can give some guidance to the experiments and is applicable to the industry. For these reasons, there are two major requirements on the models and numerical methods. First, the model of the solvation should not be heavily dependent on the force fields parametrization [37, 257]. Second, the numerical method should be accurate enough for capturing the physical meaningful results provided by the models. For instance, in the Cartesian mesh based numerical methods, the grid size influence should not affect the results too much [107, 24]. These two themes, developing force field parametrization independent models and grid size independent numerical schemes, are carried throughout this dissertation.

In chapter 2, we briefly review the classical three families of the solvation models. The pros and cons of different solvation models are shortly discussed. We provide relatively detailed discussion and mathematical description of the implicit solvation and integral equation based solvation models, both of which are more mathematically interesting in my point of view. The solvation models presented in this chapter will be the outline of the whole dissertation, and in the later chapters we will handle some detailed issues that appear in the solvation models.

In chapter 3, we present the recently proposed differential geometry based implicit solvation model. This model couples the polar and nonpolar solvation effects in a self-consistent manner. The total variation theory is employed for describing the solvent solute interface. In this model, the solvent solute interface and solute electrostatics are optimized simultaneously. It is different from the classical solvation models with a separated molecular surface modeling. We provide a systematic parameters optimization strategy in this dissertation.

Chapter 4 presents the Eulerian solvent excluded surface (ESES), which is the solvent excluded surface embedded in the Cartesian mesh, designed for the finite difference based numerical methods. Compared to the existing solvent excluded surface software, our surface is density totally independent, analytical without any approximation. The molecular surface area, volume, and molecular electrostatic analysis indicate that the state-of-the-art software MSMS [212] converges to our ESES software both qualitatively and quantitatively. Besides the surface generation, we also study the biomolecular topological structures in this chapter via the homology theory, which shows great success for understanding the topological structures of the molecules. Level set method in companion with the persistent homology theory are used for studying the persistence of the topological features associated with the ESES. This is the joint work with Beibei Liu et al.

In chapter 5, we present the numerical method for the coarse grid Poisson Boltzmann solver. Compared to all the existing Poisson Boltzmann software, we provide the most accurate and grid spacing independent reaction field energy calculation, thus meeting the basic theme of this dissertation. The solver is constructed by the following four instruments: The solvated molecular conformation modeling; Treatment of the singular charges arise from the solute molecular parametrization; Treatment of the complex geometry of the interface in the discretization of the Poisson Boltamzann equation; The evaluation of the reaction field energy, in which the coarse grid atomic central electrostatic potential evaluation is addressed. A large amount of tests, ranging from analytical tests to more than one thousand biomolecular tests, indicate our software can provide less than 0.4 % error for the electrostatic calculation for studying the solvation effects, even at the grid size 1.1 Å. A further study of the electrostatic binding free energy by our software demonstrates the accuracy for studying the protein ligand binding. This work is provided as a free online sever for the electrostatic analysis of the small molecule and biomolecules.

The blind solvation free energy prediction problem is considered in both chapters 6 and 7. In chapter 6, we proposed a hybrid physical and statistical model for the solvation free energy prediction, in which the solvation free energy is modeled as the summation of two isolated components, polar and nonpolar energies. The Poisson model and its polarizable version are utilized for modeling the polar solvation free energy. The statistical model is adopted for the nonpolar solvation free energy modeling, in which we assume the same class of molecules admit the same set of parameters in the nonpolar solvation free energy function. Motivated by the work in chapter 6, we propose a unified and force field parametrization less sensitive approach for the solvation free energy prediction in chapter 7, learning to rank based solvation prediction. Instead, the solvation free energy itself is regarded as a unity entry. The basic assumption now is that similar molecules take close solvation free energies, which is assumed to be a function of the molecular descriptors. To this end, we first employ the learning to rank method for finding the neighbor molecules to the target molecule, then utilize the neighbor information for training a linear function, and further to predict the solvation free energy of the target molecule.

In chapter 8, we extend the learning to rank based solvation prediction to a protein ligand binding free energy prediction. We propose a learning to rank based scoring function for accurate protein ligand binding affinity prediction. Our scoring function can be regarded as a hybrid force field and knowledge method. A large amount of numerical results demonstrate the accuracy of the proposed scoring approach.

Finally, we summarize the contribution of this dissertation in chapter 9.

Chapter 2

Solvation Models

Solvent models are a variety of methods to account for the behavior of solvated condensed phases, which allows simulation of chemical reactions and biological processes that take place in the solvated phases [222]. Such solvation incorporated simulations allow better predictions and improved understanding of the physical processes. The various solvation models can be generally classified based on the physical description of the solvent molecules: Explicit solvent models model the solvent molecule at the atomic level [36, 69, 188]; implicit solvent models model the solvent simply as a dielectric continuum [233, 232, 199, 221, 94, 296, 219]; the integral equation based solvation models model the solvent distribution based on the statistical mechanics theory [91, 119, 113, 67, 110, 203]. Implicit models are generally computationally efficient and can provide a reasonable description of the solvent behavior, but fail to account for the local fluctuations in solvent density around a solute molecule. The density fluctuation behavior is due to solvent ordering around a solute and is particularly prevalent when water is considered as the solvent. Explicit models are often less computationally economical, but can provide a physical spatially resolved description of the solvent. However, many of these explicit models are computationally demanding and can fail to reproduce some experimental results, often due to certain fitting methods and parametrization. Integral equation based solvation models mediate the pros and cons of both implicit and explicit solvent models. In this chapter we will provides a brief review of the different level of solvation models, with emphasis on the implicit solvent models and integral equation based solvation models.

2.1 Explicit Solvent Models

Explicit solvent models treat explicitly, i.e., the coordinates, and usually at least some of the solvent molecular degrees of freedom, are included in the solvent model. This model provides the most realistic modeling of the solvent solute interaction among different levels of solvation models, especially when the long range electrostatics interaction are dealt with the Ewald summation or Fast Multipole Method (FMM). These models generally occur in the application of molecular mechanics (MM), molecular dynamics (MD), and Monte Carlo (MC) simulations. These simulations often employ molecular mechanics force fields which are generally empirical, the force fields are usually parameterized based on a higher level theory or experimental data [55, 125].

The explicit solvent model gives the most detailed description of the solvent, and in turn it is extremely computationally expensive.

2.2 Implicit Solvent Models

2.2.1 Introduction

Water has many chemically and biologically necessary properties, one of which is a dielectric. As a dielectric, water screens (lessens) electrostatic interactions between charged particles. Water can therefore be crudely modeled as a dielectric continuum. In this manner, the electrostatic forces of a biological system can be expressed as a system of differential equations which can be solved for the electric field caused by a collection of charges. Implicit solvent models are a class of important solvation models, implicit due to the continuum description of the solvent. It is generally believed to be the best compromise between accuracy and efficiency.

2.2.2 Poisson Boltzmann (PB) Model

The Poisson Boltzmann equation (PBE), which can be formulated as

$$-\nabla \cdot (\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})) = \rho_m(\mathbf{r}) + \sum_i c_i q_i \cdot \exp\left[-\frac{q_i\phi(\mathbf{r})}{k_B T}\right] \tag{2.2.1}$$

is a nonlinear equation which solves for the electrostatic field, $\phi(\mathbf{r})$, based on the position dependent permittivity function $\epsilon(\mathbf{r})$, the solute charge distribution $\rho_m(\mathbf{r})$, and the bulk charge density c_i of ion q_i. This equation exactly solves for the electrostatic field of a charge distribution in a dielectric. Mathematically speaking, PBE is an elliptic interface problem with complex interface geometry and singular source. The complex interface comes from the complex shape of molecules in the solvent, and the singular source is due to the solute charge distribution [76].

The application of the PB model varies in many different scientific fields. It is also known in electrochemistry as Gouy Chapman theory; in solution chemistry as Debye Hückel theory; in colloid chemistry as Derjaguin Landau Verwey Overbeek (DLVO) theory. Only minor modifications are necessary to apply the PBE to various interfacial models, making it a highly useful tool in determining electrostatic potential at surfaces [29].

2.2.3 Generalized Born (GB) Model

Even though the PB model calculates the electrostatic in the solvent medium exactly, it is very computationally expensive. The Generalized Born (GB) equation provides an approximation of the PBE, and offers a fast calculation of the electrostatic field in the solvent environment. The GB model models atoms as charged spheres whose internal dielectric is lower than that of the environment. The screening which each atom, i, experiences is determined by the local environment: the more atom i is surrounded by other atoms, the less it's electrostatics will be screened since it is more surrounded by low dielectric; this property is called one atom descreening another. Different GB models calculate atomic descreening in different approaches. Descreening is used to calculate the Born radius, α_i, of each atom. The Born radius of an atom measures the degree of descreening. A large Born radius represents small screening (strong electric field) as if the atom were in vacuum. A small Born radius represents large screening (weak electric field) as if the atom were in bulk water. We will give a short review of the basic ideas behind the GB model, more detailed theory can be found in the works [268, 191].

2.2.3.1 Generalized Born Equations

In a GB simulation, the total electrostatic force on an atom i, is the difference between the net Coulombic force and GB force in the atom i, where the GB force is contributed from the nearby atoms:

$$\mathbf{F}_i = \mathbf{F}_i^{\text{Coulomb}} - \mathbf{F}_i^{\text{GB}},$$

where the electrostatic forces are contributed by other nearby atoms within a cutoff.

The GB force on atom i is the derivative of the total GB energy with respect to relative atom distance r_{ij}, i.e.,

$$\begin{aligned}\mathbf{F}_i^{\mathrm{GB}} &= -\sum_j \left[\frac{dE_T^{\mathrm{GB}}}{dr_{ij}}\right]\hat{r}_{ij} \qquad (2.2.2)\\ &= -\sum\left[\sum_k \frac{\partial E_T^{GB}}{\partial \alpha_k}\frac{d\alpha_k}{dr_{ij}} + \frac{\partial E_{ij}^{GB}}{\partial r_{ij}}\right]\hat{r}_{ij}\\ &= -\sum_j \left[\frac{\partial E_T^{GB}}{\partial \alpha_i}\frac{d\alpha_i}{dr_{ij}} + \frac{\partial E_T^{GB}}{\partial \alpha_j}\frac{d\alpha_j}{dr_{ij}} + \frac{\partial E_{ij}^{GB}}{\partial r_{ij}}\right]\hat{r}_{ij},\end{aligned}$$

where the partial derivative are included since the Born radius, α, is a function of all relative atom distances.

The total GB energy of the system is

$$E_T^{GB} = \sum_i \sum_{j>i} E_{ij}^{GB} + \sum_i E_{ii}^{GB}, \qquad (2.2.3)$$

where E_{ii}^{GB} is the Born radius dependent self energy of atom i, and the GB energy between atoms i and j is given by

$$E_{ij}^{GB} = -k_e D_{ij}\frac{q_i q_j}{f_{ij}},$$

the dielectric term D_{ij} is given by

$$D_{ij} = \left(\frac{1}{\epsilon_m} - \frac{\exp(-\kappa f_{ij})}{\epsilon_s}\right),$$

and the GB function is given by

$$f_{ij} = \sqrt{r_{ij} + \alpha_i\alpha_j \exp\left(\frac{-r_{ij}^2}{4\alpha_i\alpha_j}\right)}.$$

The constants referred in the above equations are listed below

- $k_e = 332.063711$ kcal Å$/e^2$ is the Coulomb constant.
- ϵ_s, dielectric constant of solvent.
- ϵ_m, dielectric constant of solute.
- ϵ_0, dielectric constant of the vacuum.
- κ, the Debye screening length, calculated from ion concentration based on $\kappa^{-1} = \sqrt{\frac{\epsilon_0 \epsilon_p kT}{2N_A e^2 T}}$.

2.2.4 Polarizable Continuum Model

Polarizable Continuum Model (PCM) is another class of implicit solvent models, the key components of this class of models are the *Ab Initio* charge calculation through different level of quantum mechanics theories, and the incorporation of the solvent solute polarization effects through a self consistent coupling of the electron density and electrostatic equations [238, 239, 71].

Consider the solute molecule M in the solvent, in the electrostatic solvent-solute interaction where the charge distribution ρ_{M} of the solute inside the cavity polarizes the dielectric continuum, which in turn, polarizes the solute charge distribution. The PCM model nests the classical electrostatic problem into a quantum mechanical framework to study polarization effects. Let H_{M} be the Hamiltonian of the solute M in solvent, which depends on the coordinates of the N_{el} electrons: $\mathbf{q} = \{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{N_{\mathrm{el}}}\}$, and on the coordinates of the N_{nuc} nuclei: $\mathbf{Q} = \{\mathbf{Q}_1, \mathbf{Q}_2, \cdots, \mathbf{Q}_{N_{\mathrm{nuc}}}\}$, and let H_{M}^0 be the corresponding Hamiltonian with Born-Oppenhemier approximation in vacuum which has the same dependence. In the PCM model, we have

$$H_{\mathrm{M}}(\mathbf{q}, \mathbf{Q}) = H_{\mathrm{M}}^{0}(\mathbf{q}, \mathbf{Q}) + V_{\mathrm{int}}, \tag{2.2.4}$$

where V_{int} is the solute solvent interaction potential.

The related Schrodinger equation is

$$H_{\mathrm{M}}(\mathbf{q}, \mathbf{Q})\Psi^{f}(\mathbf{q}, \mathbf{Q}) = E^{f}(\mathbf{Q})\Psi^{f}(\mathbf{q}, \mathbf{Q}), \tag{2.2.5}$$

here the superscripts f indicates the solution was obtained iteratively. All the relevant information about the solvent effects on the solute M is contained in the eigenvalue E^{f} and in the wave function Ψ^{f}.

The charge of the solute molecule is the sum of a discrete nuclear charge distribution and the electron density function

$$\rho_{\mathrm{M}}(\mathbf{q}, \mathbf{Q}) = \rho_{\mathrm{nuc}}(\mathbf{q}, \mathbf{Q}) + \rho_{\mathrm{el}}(\mathbf{q}, \mathbf{Q}), \tag{2.2.6}$$

where the nuclei and electron charge are given, respectively, by

$$\rho_{\mathrm{nuc}}(\mathbf{q}, \mathbf{Q}) = \sum_{\alpha} Z_{\alpha}\delta(\mathbf{r} - \mathbf{Q}_{\alpha}), \tag{2.2.7}$$

$$\rho_{\mathrm{el}}(\mathbf{q}, \mathbf{Q}) = -\int |\Psi^{f}(\mathbf{q}, \mathbf{Q})|^{2} d\mathbf{q}_{1} \cdots d\mathbf{q}_{N_{\mathrm{el}}}, \tag{2.2.8}$$

where Z_{α} is a nuclear charge and the index α runs over all the nuclei of M.

In Eq. (2.2.4), the interaction between the solute and solvent V_{int}, which depends on $\mathbf{q}$,

$\mathbf{Q}$, and a thermally averaged distribution function of the solvent molecules, g_{S}, is given by

$$V_{\mathrm{int}} = V_{\mathrm{int}}(\mathbf{q}, \mathbf{Q}, g_{\mathrm{S}}), \tag{2.2.9}$$

in the basic continuum model, the interaction term is reduced to its classical electrostatic component

$$V_{\mathrm{int}} := V_{\sigma}(\mathbf{q}, \mathbf{Q}, \rho_{\mathrm{M}}) = \sum_{\alpha} Z_{\alpha} \Phi_{\sigma}(\mathbf{Q}_{\alpha}) - \sum_{i} \Phi_{\sigma}(\mathbf{q}_{i}), \tag{2.2.10}$$

where $\Phi_{\sigma}(\mathbf{r})$ is the value of the electrostatic potential field generated by the polarized dielectric at the position $\mathbf{r}$.

The solvent solute interaction contribution to the total energy E^{f} is given by

$$W_{\mathrm{MS}} = \int \Psi^{f} V_{\sigma} \Psi^{f} d\mathbf{q}_{1} \cdots d\mathbf{q}_{N_{\mathrm{el}}} = \int \rho_{\mathrm{M}}(\mathbf{r}) \Phi_{\sigma}(\mathbf{r}) dr^{3}, \tag{2.2.11}$$

where the integration takes over the whole solute solvent space.

2.2.4.1 Quantum Mechanical Problem

To obtain the nuclear and electron charges distribution in the PCM framework, we should solve the following Schrodinger equation with the inclusion of the solute solvent interaction

$$H_{\mathrm{eff}} \Psi = E \Psi, \tag{2.2.12}$$

where the effective Hamiltonian H_{eff} is given by

$$H_{\mathrm{eff}} = H_{\mathrm{M}}^{0} + V_{\mathrm{int}}. \tag{2.2.13}$$

2.2.4.2 Electrostatic Problem

The solute-solvent interaction is obtained via solving the classical electrostatic problem. Consider the Poisson equation

$$-\nabla \cdot (\epsilon(\mathbf{r})\nabla\Phi(\mathbf{r})) = 4\pi\rho_{\mathrm{M}}(\mathbf{r}), \tag{2.2.14}$$

where the permittivity function is given by

$$\epsilon(\mathbf{r}) = \begin{cases} 1, & \mathbf{r} \in \Omega^{\mathrm{s}} \\ \epsilon, & \mathbf{r} \in \Omega^{\mathrm{m}} \end{cases} \tag{2.2.15}$$

where Ω^{s} and Ω^{m} represent the solvent and solute domains, respectively. ϵ is the dielectric constant in medium.

Due to the solvent effects, the electrostatic potential Φ can be decomposed as

$$\Phi(\mathbf{r}) = \Phi_{\mathrm{M}}(\mathbf{r}) + \Phi_{\sigma}(\mathbf{r}), \tag{2.2.16}$$

where

- $\Phi_{\mathrm{M}}(\mathbf{r})$ is the electrostatic potential generated by the charge distribution ρ_{M}, which has the analytical expression as the convolution of the Green's function with the charge distribution ρ_{M}.

- $\Phi_{\sigma}(\mathbf{r})$ is the reaction field potential generated by the polarization of the dielectric medium.

The Poisson equation Eq. (2.2.14) subjects to the following far field boundary conditions

$$\lim_{r\to\infty} r\Phi(\mathbf{r}) = \alpha, \quad \lim_{r\to\infty} r^2\Phi(\mathbf{r}) = \beta, \tag{2.2.17}$$

with finite values for α and β.

At the solute solvent contact part, the following interface conditions constraints should be added to Eq. (2.2.14)

$$[\Phi] = \Phi_{\text{in}} - \Phi_{\text{out}} = 0, \quad [\epsilon(\mathbf{r})\Phi_{\mathbf{n}}] = \left(\epsilon(\mathbf{r})\frac{\partial\Phi}{\partial\mathbf{n}}\right)_{\text{in}} - \left(\epsilon(\mathbf{r})\frac{\partial\Phi}{\partial\mathbf{n}}\right)_{\text{out}}, \tag{2.2.18}$$

where $\mathbf{n}$ is the normal direction pointing outward the solute domain, and the subscripts ”in” and ”out” represent for the limit from the inside solute domain and outside one.

2.2.4.3 Dielectric Polarizable Continuum Model (DPCM)

There are various versions of the PCM, for instance, Dielectric PCM (DPCM) [238], Conductor-like PCM (C-PCM) [54], Integral Equation Formalism PCM (IEFPCM) [71], wavelet PCM [242] et al. It is very difficult to review all these PCM models here. In this part, we specifically give a short review of the DPCM model, the basic idea is called the Apparent Surface Charge (ASC) similar to the induced surface charge approach for solving the Poisson Boltzmann equation [154, 94, 209].

For systems composed by regions at constant isotropic permittivity, the polarization vector is given by the gradient of the total potential

$$\mathbf{P}_i(\mathbf{r}) = -\frac{\epsilon - 1}{4\pi}\nabla\Phi(\mathbf{r}), \tag{2.2.19}$$

where ϵ is the dielectric constant of the solute region, and we assume the solvent dielectric constant to be 1.

At the boundary of two regions i and j, there is an apparent surface charge distribution given by

$$\sigma_{ij} = -(\mathbf{P}_j - \mathbf{P}_\mathbf{i}) \cdot \mathbf{n}_{ij}, \tag{2.2.20}$$

where $\mathbf{n}_{ij}$ is the unit vector at the boundary surface pointing from medium i to medium j.

Thus the surface charge for the PCM model is given by

$$\sigma = -\mathbf{P} \cdot \mathbf{n} = \frac{\epsilon - 1}{4\pi} \nabla \Phi_{\text{out}} \cdot \mathbf{n} = \frac{\epsilon - 1}{4\pi\epsilon} \nabla \Phi_{\text{in}} \cdot \mathbf{n}, \tag{2.2.21}$$

Since $\Phi = \Phi_{\text{M}} + \Phi_{\text{S}}$, thus the surface charge can be further expressed as

$$\sigma = \frac{\epsilon - 1}{4\pi\epsilon} \frac{\partial(\Phi_{\text{M,in}} - \Phi_{\text{S,in}})}{\partial \mathbf{n}},$$

therefore, the reaction field potential due to the solvent polarization can be expressed as

$$\Phi_\sigma(\mathbf{r}) = \int_\Gamma \frac{\sigma(\mathbf{s})}{|\mathbf{r} - \mathbf{s}|} ds^2, \tag{2.2.22}$$

where Γ is the surface of the solute molecule.

2.3 Integral Equation Based Solvation Models

Compare to the explicit solvent theory, the continuum simplifies the description of the solvent, and reduces the complexity of the solvation model dramatically. Such implicit solvent models are often useful, but have a variety of limitations: they drastically average the response of water dipoles and ions to the fields created by solutes, miss most effects of atomic and molecular sizes, collapse all ion effects into a single ionic strength parameter, and fail to account for non-electrostatic aspects of solvation. In many situations these mean field approximation may be too severe [222]. There is a long-studied alternative approach to understand the equilibrium properties of water and ions, based on the integral equation approach of Ornstein and Zernike. These ideas were originally applied to atomic liquids, and have been extended to molecular solvents such as water by a variety of methods, most notably via the Reference Interaction Site Model (RISM) [91, 119, 113, 67, 110, 203].

2.3.1 Ornstein-Zernike Equation

In homogeneous fluids, the spatial number density distribution, $\rho(\mathbf{r})$, is uniform and provides little information. In contrast, the number density of a particle, 2, relative to a fixed particle, 1, contains a wealth of information and, in the grand canonical ensemble, is given by

$$\rho^{(2)}(\mathbf{1},\mathbf{2}) = \frac{1}{\Xi}\sum_{N=2}^{\infty}\frac{Z^N}{(N-2)!}\int \exp[-\beta V_N] d\mathbf{r}^{N-2}, \tag{2.3.1}$$

where the position and orientation of molecular species are denoted by boldface numbers, e.g., $\mathbf{1} := (\mathbf{r}_1, \Omega_1)$, $\beta = \frac{1}{k_B T}$ with k_B be the Boltzmann constant and T the absolute temperature, Ξ is the grand partition function, Z^N and V_N are the N-particle partition function and potential.

For homogeneous fluids, the 2-particle density distribution is related to the Pair Distribution Function (PDF), $g(\mathbf{1},\mathbf{2})$, through the equation

$$\rho^{(2)}(\mathbf{1},\mathbf{2}) = \rho_1\rho_2 g(\mathbf{1},\mathbf{2}), \tag{2.3.2}$$

where ρ_1 and ρ_2 are the bulk number densities of particles 1 and 2, respectively. When the orientational dependence is averaged out, $g(r)$ is known as the Radial Distribution Function (RDF). Alternatively, the PDF is also related to the potential of mean force, $w(\mathbf{1},\mathbf{2})$

$$h(\mathbf{1},\mathbf{2}) + 1 = g(\mathbf{1},\mathbf{2}) = \exp[-\beta w(\mathbf{1},\mathbf{2})], \tag{2.3.3}$$

where $h(\mathbf{1},\mathbf{2})$ is the Total Correlation Function (TCF).

For homogeneous, multi-component, molecular liquids, the OZ integral equation is given by

$$h_{ij}(\mathbf{1},\mathbf{2}) = c_{ij}(\mathbf{1},\mathbf{2}) + \sum_k \rho_k \int c_{ik}(\mathbf{1},\mathbf{3}) h_{kj}(\mathbf{3},\mathbf{2}) d\mathbf{3}, \tag{2.3.4}$$

where we denote the molecular species by i, j and k, $c(\mathbf{1},\mathbf{2})$ is the Direct Correlation Function (DCF) and the integration is performed over all space. Physically, we can interpret the TCF as the sum of contributions from the direct interaction of the two particles (DCF) plus the interactions mediated by the surrounding particles (the right hand convolution).

As both h and c are unknown functions, a second, closure equation is required to find a solution.

2.3.2 Closures

The most general case for the closure equation is

$$g_{\alpha\gamma} = \exp\{-\beta u_{\alpha\gamma} + h_{\alpha\gamma} - c_{\alpha\gamma} + b_{\alpha\gamma}\}, \tag{2.3.5}$$

where b is the bridge function and we have dropped the functional arguments for brevity and generality.

In the Hyper-Netted Chain (HNC) approximation, the bridge function is set to zero, which yields the following HNC closure

$$g_{\alpha\gamma} = \exp\{-\beta u_{\alpha\gamma} + h_{\alpha\gamma} - c_{\alpha\gamma}\}, \tag{2.3.6}$$

this has been found to produce very good results for ionic and polar systems. It also has an exact, closed form expression for the chemical potential when coupled with RISM theory. However, it does have drawbacks, including thermodynamic inconsistencies, poor results for neutral systems, difficulties with particle size asymmetries and difficulties converging solutions [216].

To address the issue of convergence, Kovalenko and Hirata developed a partially linearized closure. Regions of enhanced density were linearized, avoiding the exponential density response for strong potential interactions. This linearization was later generalized to a Taylor series

$$g_{\alpha\gamma} = \begin{cases} \exp\{t^*_{\alpha\gamma}\}, & \text{for } t^*_{\alpha\gamma} < 0 \\ \sum_{i=0}^{n} \frac{(t^*_{\alpha\gamma})^i}{i!}, & \text{for } t^*_{\alpha\gamma} \geq 0, \end{cases} \tag{2.3.7}$$

where

$$t^*_{\alpha\gamma} = -\beta u_{\alpha\gamma} + h_{\alpha\gamma} - c_{\alpha\gamma}.$$

2.3.3 1D-RISM

Most modern biomolecular force fields use interaction site models in which a molecule is composed of a number of sites, typically atoms, that interact in a pair-wise fashion. Such models offer a very effective way to deal with nonspherical molecules but require a practical method to apply model in Eq. (2.3.4) to molecular species with multiple sites. One approach (which, in practice, is restricted to molecules with a small number of sites) is to treat the molecules as rigid body, and site-site orientationally average the correlation functions for each site, reducing the equations to one dimension, i.e., orientational averaging is done about each site rather than, for example, averaging about the molecular center-of-mass.

In the RISM approach this is achieved by first treating the DCF as decomposable into the sum of site-site direct correlation functions

$$c(\mathbf{1},\mathbf{2}) = \sum_{\alpha_1,\gamma_2} c_{\alpha_1\gamma_2}(|\mathbf{r}_{\alpha_1} - \mathbf{r}_{\gamma_2}|), \tag{2.3.8}$$

where $c(\mathbf{1},\mathbf{2})$ is the DCF between molecules 1 and 2, as well as α and γ are interaction sites on molecules 1 and 2, respectively.

Molecules are assumed to be rigid, and their shape enters the theory through the intramolecular correlation matrix, represented in Fourier space,

$$\hat{\omega}_{\alpha\gamma}(k) = \delta_{\alpha\gamma} + (1-\delta_{\alpha\gamma})\frac{\sin(kl_{\alpha\gamma})}{kl_{\alpha\gamma}}, \tag{2.3.9}$$

where δ is the Kronecker δ-function and $l_{\alpha\gamma}$ is the distance between sites in the same type of molecule. For the same site, $\alpha = \gamma$, $l_{\alpha\alpha} = 0$ and $\hat{\omega}_{\alpha\alpha}(k) = 1$, while for the sites belong to different types of molecule $\hat{w}_{\alpha\gamma}(k) = 0$.

With the definition of the intramolecular correlation function, we can now express the molecular OZ integral equation Eq. (2.3.4) in terms of interaction sites rather than molecules.

The multi-components 1D-RISM equation can be written explicitly for molecules 1 and 2 as

$$\rho_\alpha h_{\alpha\gamma}(r)\rho_\gamma = \sum_{\lambda}^{N_{site}}\sum_{\beta}^{N_{site}} \omega_{\alpha\lambda}(r) * c_{\lambda\beta}(r) * \omega_{\beta\gamma}(r) + \sum_{\lambda}^{N_{site}}\sum_{\beta}^{N_{site}} \omega_{\alpha\lambda}(r) * c_{\lambda\beta}(r) * \rho_\beta h_{\beta\gamma}(r)\rho_\gamma,$$

where $*$ is the convolution operator and N_{site} is the total number sites from all molecular species. The above equation can be concisely written into the following compact matrix form

$$\boldsymbol{\rho}\mathbf{h}\boldsymbol{\rho} = \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\omega} + \boldsymbol{\omega} * \mathbf{c} * \boldsymbol{\rho}\mathbf{h}\boldsymbol{\rho}, \tag{2.3.10}$$

with $\boldsymbol{\rho}$ being a diagonal matrix of scalar values, $\boldsymbol{\omega}$ and $\mathbf{c}$ are matrices of radially dependent functions and all matrices are of size $N_{site} \times N_{site}$.

2.3.4 3D-RISM

For macromolecular ions, which are composed of more than a few sites, the approximation of spherically symmetric distribution functions begins to break down. One approach is to use a full 3D description of the macromolecular solute, U, while using orientationally averaged distributions for the solvent V. If the solute is at infinite dilution, Eq. (2.3.4) can be rewritten as

$$h_{ij}^{VV}(\mathbf{i},\mathbf{j}) = c_{ij}^{VV}(\mathbf{i},\mathbf{j}) + \sum_k \rho_k^V \int_\Omega c_{ik}^{VV}(\mathbf{i},\mathbf{k}) h_{kj}^{VV}(\mathbf{k},\mathbf{j}) d\mathbf{k}, \tag{2.3.11}$$

$$h_{i}^{VV}(\mathbf{1},\mathbf{i}) = c_{i}^{UV}(\mathbf{1},\mathbf{i}) + \sum_k \rho_k^V \int_\Omega c_{k}^{UV}(\mathbf{1},\mathbf{k}) h_{ki}^{VV}(\mathbf{k},\mathbf{j}) d\mathbf{k}. \tag{2.3.12}$$

In the 3D-RISM model, Eq. (2.3.11) gives the TCF of the bulk solvent, which is then used in Eq. (2.3.12) to obtain the distribution of the solvent about the solute.

2.4 Conclusion

In this chapter, we provides a short review of the three classes of the solvation models. The remaining of this dissertation focuses on solving some problems that come from the above solvation models and enrich the family of solvation models. The application of the continuum solvation models for studying the solvation and binding phenomena is also a main theme of this dissertation.

Chapter 3

Self Consistent Coupling of Polar and Nonpolar Solvation Free Energy

The general idea of implicit solvent models is to treat the solvent as a dielectric continuum and describe the solute with atomistic detail [62, 219, 112, 211, 126]. The total solvation free energy is decomposed into nonpolar and polar parts. There is a wide variety of ways to carry out this decomposition. For example, nonpolar energy contributions can be modeled in two stages: the work of displacing solvent when adding a rigid solute to the solvent and the dispersive nonpolar interactions between the solute atoms and surrounding solvent. The polar part is due to the electrostatic interactions and can be approximated by Generalized Born (GB) [64, 11, 241, 191, 84, 296, 138, 236, 178, 38, 99], Polarizable Continuum Model (PCM) [237]and Poisson-Boltzmann (PB) models [145, 76, 219, 62, 293, 6, 295]. Among them, GB models are heuristic approaches to polar solvation energy analysis. PCMs resort to quantum mechanical calculations of induced solute charges. PB methods can be formally derived from Maxwell equations and statistical mechanics for electrolyte solutions [19, 182, 111] and therefore offer the promise of handling large biomolecules with sufficient accuracy and robustness [61, 190, 11].

Conceptually, the separation between continuum solvent and the discrete (atomistic) solute introduces an interface definition. This definition may take the form of analytical functions [100, 98, 99] or nonsmooth boundaries dividing the solute-solvent domains. The van der Waals surface, solvent accessible surface [147], and Molecular Surface (MS) [206] are devised for this purpose and have found their success in biophysical calculations [225, 161, 58, 142, 20, 68, 120, 155]. It has been noticed that the performance of implicit solvent models is very sensitive to the interface definition [65, 66, 186, 231]. This comes as no surprise because many of these popular interface definitions are *ad hoc* divisions of the solute and solvent domains based on rigid molecular geometry and neglecting solute-solvent energetic interactions. Additionally, geometric singularities [51, 212] associated with these surface definitions incur enormous computational instability [295, 278, 279] and lead to conceptual difficulty in interpreting the sharp interface [45].

The Differential Geometry (DG) theory of surfaces [275] and associated geometric Partial Differential Equations (PDEs) provide a natural description of the solvent-solute interface. In 2005, Wei and his collaborators introduced curvature-controlled PDEs for generating molecular surfaces for solvation analysis [271]. The first variational solvent-solute interface, namely, the Minimal Molecular Surface (MMS), was constructed in 2006 by Wei and coworkers based on the DG theory of surfaces [13, 14, 15]. MMSs are constructed by solving the mean curvature flow, or the Laplace-Beltrami flow, and have been applied to the calculation of electrostatic potentials and solvation free energies [46, 15]. This approach was generalized to potential-driven geometric flows, which admits physical interactions, for the surface generation of biomolecules in solution [12]. While our approaches were employed and/or modified by many others [47, 281, 284, 285] for molecular surface and solvation analysis, our geometric PDE [271] and variational surface models [13, 15, 12] are, to our knowledge, the first of their kind for solvent-solute interface and solvation modeling.

Since the surface area minimization is equivalent to the minimization of surface free energies, due to a constant surface tension, this approach can be easily incorporated into the variational formulation of the PB theory [218, 92] to result in DG-based full solvation models [43, 269], following a similar approach by Dzubiella *et al* [70, 291]. The DG-based solvation models have been implemented in the Eulerian formulation, where the solvent-solute interface is embedded in the three-dimensional (3D) Euclidean space and behaves like a smooth characteristic function [43]. The resulting interface and associated dielectric function vary smoothly from their values in the solute domain to those in the solvent domain and are computationally robust. An alternative implementation is the Lagrangian formulation [44] in which the solvent-solute boundary is extracted as a sharp surface at a given isovalue and subsequently used in the solvation analysis, including nonpolar and polar modeling.

One major advantage of the DG based solvation model is that it enables the synergistic coupling between the solute and solvent domains via the variation procedure. As a result, the DG based solvation model is able to significantly reduce the number of free parameters that users must “fit” or adjust in applications to real-world systems [235]. It has been demonstrated that physical parameters, i.e., pressure and surface tension obtained from experimental data, can be directly employed in the DG-based solvation models for accurate solvation energy prediction [59]. Another advantage of the DG based solvation model is that it avoids the use of *ad hoc* surface definitions and its interfaces, particularly ones generated from the Eulerian formulation [43], are free of troublesome geometric singularities that commonly occur in conventional solvent-accessible and solvent-excluded surfaces [52, 212]. As a result, the DG based solvation model bypasses the sophisticated interface techniques required for solving the PB equation [278, 279, 90]. In particular, the smooth solvent-solute interface obtained from the Eulerian formulation [43] can be directly interpreted as the physical solvent-solute boundary profile. Additionally, the resulting smooth dielectric boundary can also have a straightforward physical interpretation. The other advantage of the DG based solvation model is that it is natural and easy to incorporate the Density Functional Theory (DFT) in its variational formulation. Consequently, it is able to reevaluate and reassign the solute charge induced by solvent polarization effect during the solvation process [45]. The resulting total energy optimization process recreates or resembles the solvent-solute interactions, i.e., polarization, dispersion, and polar and nonpolar coupling in a realistic solvation process. Recently, DG based solvation model has been extended to DG based multiscale models for non-equilibrium processes in biomolecular systems [269, 273, 272, 40, 41]. These models recover the DG based solvation model at the equilibrium [273].

Recently, we have demonstrated [46] that the DG based nonpolar solvation model is able to outperform many other methods [82, 246, 202] in solvation energy predictions for a large number nonpolar molecules. The Root Mean Square Error (RMSE) of our predictions was below 0.4 kcal/mol, which clearly indicates the potential power of the DG based solvation formulation. However, the DG based full solvation model has not shown a similar superiority in accuracy, although it works very well [43, 44]. Having so many aforementioned advantages, the DG based solvation models ought to outperform other methods with a similar level of approximations. One obstacle that hinders the performance of our DG based *full* solvation model is the numerical instability in solving two strongly coupled and highly nonlinear PDEs, namely, the Generalized Laplace-Beltrami (GLB) equation and the generalized PB (GPB) equation. To avoid such instability, a strong parameter constraint was applied to the nonpolar part in our earlier work [43, 44], which results in the reduction of our model accuracy.

The objective of the present work is to explore a better parameter optimization of the DG based solvation models. A pair of conditions is prescribed to ensure the physical solution of the GLB equation, which leads to the well-posedness of the GPB equation. Such a well-posedness in turn renders the stability of solving the GLB equation. The stable solution of the coupled GLB and GPB equation enables us to optimize the model parameters and produce the highly accurate prediction of solvation free energies. Some of the best results are obtained in the solvation free energy prediction of more than a hundred molecules of both polar and nonpolar types.

The rest of this chapter is organized as the follows. To establish the notation and facilitate further development, we present a brief review of the DG based solvation models in Section 3.1. By using the variational principle, we derive the coupled GLB and GPB equations. Necessary boundary conditions and initial values are prescribed to make this coupled system well-posed. Section 3.2 is devoted to parameter learning algorithms. We develop a protocol to stabilize the iterative solution process of coupled nonlinear PDEs. We introduce perturbation and convex optimization methods to ensure stability of the numerical solution of the GLB equation in coupling with the GPB equation. The newly achieved stability in solving the coupled PDEs leads to an appropriate optimization of solvation free energies with respect to our model parameters. In Section 3.3, we show that for more than a hundred of compounds of various types, including both polar and nonpolar molecules, the present DG solvation model offers some of the most accurate solvation free energy prediction with the overall RMSE of 0.5 kcal/mol.

3.1 The DG based solvation model

The free energy functional for the DG based full solvation model can be expressed as [270, 43, 44]

$$\begin{aligned} G[S,\Phi] = \int \Big\{ & \gamma|\nabla S| + pS + (1-S)U + S\left[-\frac{\epsilon_m}{2}|\nabla\Phi|^2 + \Phi\ \rho_m\right] \\ & +(1-S)\left[-\frac{\epsilon_s}{2}|\nabla\Phi|^2 - k_BT\sum_\alpha \rho_{\alpha 0}\left(e^{-\frac{q_\alpha\Phi}{k_BT}} - 1\right)\right]\Big\} d\mathbf{r}, \quad \mathbf{r}\in\mathbb{R}^3 \end{aligned} \qquad (3.1.1)$$

where γ is the surface tension, p is the hydrodynamic pressure difference between solvent and solute, and U denotes the solvent-solute non-electrostatic interactions represented by the semi-discrete and semi-continuum Lennard-Jones potentials in the present work. Here $0 \le S \le 1$ is a hypersurface or simply surface function that characterizes the solute domain and embeds the 2D surface in $\mathbb{R}^3$, whereas $1-S$ characterizes the solvent domain [43]. One may consider S as the position-dependent volume fraction of the solute. Additionally, Φ is the electrostatic potential and ϵ_s and ϵ_m are the dielectric constants of the solvent and solute, respectively. Here k_B is the Boltzmann constant, T is the temperature, $\rho_{\alpha 0}$ denotes the reference bulk concentration of the αth solvent species, and q_α denotes the charge valence of the αth solvent species, which is zero for an uncharged solvent component. We use ρ_m to represent the charge density of the solute. The charge density is often modeled by a point charge approximation

$$\rho_m = \sum_j^{N_m} Q_j \delta(\mathbf{r}-\mathbf{r}_j),$$

where Q_j denoting the partial charge of the jth atom in the solute. Alternatively, the charge density computed from the DFT, which changes during the iteration or energy optimization, can be directly employed as well [45].

In Eq. (3.1.1), the first three terms consist of the so called nonpolar solvation free energy functional while the last two terms form the polar one. After the variation with respect to S, we obtain an elliptic equation for the surface function S

$$\nabla \cdot \left(\gamma \frac{\nabla S}{|\nabla S|} \right) + V = 0, \tag{3.1.2}$$

where the potential driven term is given by

$$V = -p + U + \frac{\epsilon_m}{2} |\nabla \Phi|^2 - \Phi \ \rho_m - \frac{\epsilon_s}{2} |\nabla \Phi|^2 - k_B T \sum_\alpha \rho_{\alpha 0} \left(e^{-\frac{q_\alpha \Phi}{k_B T}} - 1 \right).$$

It is a standard procedure to seek the solution of Eq. (3.1.2) by converting it into a parabolic equation [12]. As such, we construct the following Generalized Laplace-Beltrami (GLB) equation [43, 44].

$$\frac{\partial S}{\partial t} \ = \ |\nabla S| \left[\nabla \cdot \left(\gamma \frac{\nabla S}{|\nabla S|} \right) + V \right]. \tag{3.1.3}$$

here we utilized the method proposed by Marquina and Osher to tame the direct steepest descent marching 3.1.2[166].

As in the nonpolar case, solving the generalized Laplace-Beltrami equation (3.1.3) generates the solvent-solute interface through the surface function S.

Additionally, variation with respect to Φ gives rise to the generalized Poisson-Boltzmann (GPB) equation:

$$-\nabla \cdot (\epsilon(S) \nabla \Phi) = S \rho_m + (1 - S) \sum_\alpha q_\alpha \rho_{\alpha 0} e^{-\frac{q_\alpha \Phi}{k_B T}}, \tag{3.1.4}$$

where $\epsilon(S) = (1 - S)\epsilon_s + S\epsilon_m$ is the generalized permittivity function. As shown in our earlier work [270, 43], $\epsilon(S)$ is a smooth dielectric function gradually varying from ϵ_m to ϵ_s. Thus, the solution procedure of the GPB equation avoids many numerical difficulties of solving elliptic equations with discontinuous coefficients [286, 295, 294, 280, 279] in the standard PB equation.

The GLB (3.1.3) and GBP (3.1.4) equations form a highly nonlinear system, in which the GLB equation is solved for the interface profile S of the solute and solvent. The interface profile determines the dielectric function $\epsilon(S)$ in the GPB equation. The GPB equation is solved for the electrostatics potential Φ that behaves as an external potential in the GLB equation. The strongly coupled system should be solved in self-consistent iterations.

For GLB equation (3.1.3), the computational domain is $\Omega/\Omega_m^{\rm vdW}$, where $\Omega_m^{\rm vdW}$ is the solute van der Waals domain given by $\Omega_m^{\rm vdW} = \bigcup_i B(r_i^{\rm vdW})$. Here $B(r_i^{\rm vdW})$ is the ith ball in the solute centered at $\mathbf{r}_i$ with van der Waals radius $r_i^{\rm vdW}$. We apply the following Dirichlet boundary condition to $S(\mathbf{r}, t)$

$$S(\mathbf{r}, t) = \begin{cases} 0, & \forall \mathbf{r} \in \partial\Omega \\ 1, & \forall \mathbf{r} \in \partial\Omega_m^{\rm vdW}. \end{cases} \tag{3.1.5}$$

The initial value of $S(\mathbf{r}, t)$ is given by

$$S(\mathbf{r}, 0) = \begin{cases} 1, & \forall \mathbf{r} \in \partial\Omega_m^{\rm ext}, \\ 0, & \text{otherwise}, \end{cases} \tag{3.1.6}$$

where $\partial\Omega_m^{\rm ext}$ is the boundary of the extended solute domain constructed by $\Omega_m^{\rm ext} = \bigcup_i B(r_i^{\rm vdW} + r^{\rm probe})$. Here $B(r_i^{\rm vdW} + r^{\rm probe})$ has an extended radius of $r_i^{\rm vdW} + r^{\rm probe}$ with $r^{\rm probe}$ being the probe radius, which is set to 1.4 Å in the present work.

For GPB equation (3.1.4), the computational domain is Ω. We set the Dirichlet boundary condition via the Debye-Hückel expression,

$$\Phi(\mathbf{r})=\sum_{i=1}^{N_m}\frac{Q_i}{\epsilon_s|\mathbf{r}-\mathbf{r}_i|}e^{-\bar{\kappa}|\mathbf{r}-\mathbf{r}_i|},\quad \forall\mathbf{r}\in\partial\Omega, \tag{3.1.7}$$

where $\bar{\kappa}$ is the modified Debye-Hückel screening function [44], which is zero if there is no salt molecule in the solvent. Note that no interface condition [278] is needed as S and $\epsilon(S)$ are smooth functions in general for $t>0$. Consequently, the resulting GBP (3.1.4) equation is easy to solve.

To compare with experimental solvation data, one needs to compute the total solvation free energy, which, in our DG based solvation model, is obtained as

$$\Delta G=\Delta G^{\mathrm{P}}+G^{\mathrm{NP}}, \tag{3.1.8}$$

where ΔG^{P} is the electrostatic solvation free energy,

$$\Delta G^{\mathrm{P}}=\frac{1}{2}\sum_{i=1}^{N_m}Q_i\left[\Phi(\mathbf{r}_i)-\Phi_h(\mathbf{r}_i)\right] \tag{3.1.9}$$

where Φ_h is the solution of the above the GPB model in a homogenous system, obtained by setting a constant permittivity function $\epsilon(\mathbf{r})=\epsilon_m$ in the whole domain Ω. The nonpolar energy G^{NP} is computed by

$$G^{\mathrm{NP}}=\int\left[\gamma|\nabla S|+pS+(1-S)U\right]d\mathbf{r}. \tag{3.1.10}$$

The DG based solvation model is formulated as a coupled GLB and GPB equation system, in which the GLB equation provides the solvent solute boundary for solving the GPB, while the GPB equation produces the external potential in the GLB equation for the evolution of the surface function S. The solution procedure for this coupled system has been discussed in our earlier work [43, 44]. Essentially, for the GLB equation, an Alternating Direction Implicit (ADI) scheme was utilized for the time integral, in conjugation with the second order finite difference method for the spatial discretization. The GPB equation was discretized by a standard second order finite difference scheme and the resulting algebraic equation system was solved by using a standard Krylov subspace method based solver [43, 44].

3.2 Parametrization methods and algorithms

To solve the above coupled equation system, a set of parameters that appeared in the GLB equation, namely, surface tension γ, hydrodynamic pressure difference p, and the product of solvent density and well depth parameter of the jth atom $\tilde{\varepsilon}_{j\alpha} \doteq \rho_\alpha \varepsilon_j$, should be predetermined. Unfortunately, this coupled system is unstable at the certain choices of parameters. Specifically, for certain V, one may have $S > 1$ or $S < 0$, which leads to unphysical $\epsilon(S)$ and unphysical solution of GPB equation (3.1.4) and thus gives rise to a divergent S. This instability can seriously reduce the model accuracy [43, 44].

For a concise description of our algorithm, we assume that there is only one solvent component (water) and denote the parameter set as:

$$\mathbf{P} = \{\gamma, p, \tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \cdots, \tilde{\varepsilon}_{N_T}\} \tag{3.2.1}$$

where N_T is the number of types of atoms in the solute molecule.

As mentioned in the previous part, the parameter set $\mathbf{P}$ used in solving the coupled PDEs should meet two requirements, namely, the stability of solving the coupled PDEs and the optimal prediction of the solvation free energy (or fitting the experimental solvation free energy in the best approach). Based on these two criteria we introduce a two-stage numerical procedure to optimize the parameter set and solve the coupled PDEs:

- Explore the stability conditions of the coupled PDEs by introducing an auxiliary system via a small perturbation;
- Optimize the parameter set by an iteratively scheme satisfying the stability constraint.

3.2.1 Stability conditions

In this part we investigate the stability conditions for the numerical solution to the coupled PDEs (3.1.3) and (3.1.4). The basic idea is to utilize a small perturbation method. It is known that omitting the external potential in the GLB equation yields the Laplace-Beltrami (LB) equation:

$$\frac{\partial S}{\partial t} = |\nabla S| \nabla \cdot \left(\gamma \frac{\nabla S}{|\nabla S|} \right) \tag{3.2.2}$$

This equation is of diffusion type and is well posed with the Dirichlet type of boundary conditions provided $\gamma > 0$. Numerically it is easy to solve Eq. (3.2.2) to yield the profile of the solvent solute boundary.

After solving the LB equation (3.2.2), we use the generated smooth profile of the solvent solute boundary to determine the permittivity function in the GPB equation. For simplicity, we consider a pure water solvent,

$$-\nabla \cdot (\epsilon(S)\nabla\Phi) = S\rho_m. \tag{3.2.3}$$

Without the external potential the system of Eqs. (3.2.2)-(3.2.3) can be solved stably by first solving the LB equation and then the GPB equation.

Motivated by the above observation, if the external potential is dominated by the mean curvature term, the stability of coupled GPB and GLB equations can be preserved. Based on numerical experiments, the Lennard-Jones interaction between the solvent and solute is usually small since this term is constrained by the nonpolar free energy in our model. In our method, we enforce the following constraint conditions to make the coupled system well-posed in the numerical sense

$$\gamma > \gamma_0 > 0, \tag{3.2.4}$$

and

$$|p| \leq \beta\gamma, \tag{3.2.5}$$

where γ_0 and β are some appropriate positive constants.

In summary, the original problem is transformed into optimizing parameters in the following system to attain the best solvation free energy fitting with experimental results:

$$\begin{cases} \frac{\partial S}{\partial t} = |\nabla S| \left[\nabla \cdot \left(\gamma \frac{\nabla S}{|\nabla S|}\right) - p + U + \frac{1}{2}\epsilon_m |\nabla\Phi|^2 - \frac{1}{2}\epsilon_s |\nabla\Phi|^2\right], \\ -\nabla \cdot (\epsilon(S)\nabla\Phi) = S\rho_m, \\ \gamma > \gamma_0 > 0, \\ |p| \leq \beta\gamma. \end{cases} \tag{3.2.6}$$

Note that the potential $\rho_m \Phi$ is omitted in the GLB equation (3.2.6), because we have already enforced the Dirichlet boundary condition in the GLB equation, while ρ_m is inside the van der Waals surface.

Remark 3.2.1. *Based on large amount of numerical tests, it is found that there is no need to enforce the constraint conditions on the parameters that appear in the Lennard-Jones term. When this term is used to fit the solvation energy with experimental results, the parameters can be bounded in a small neighborhood of 0 automatically during the fitting procedure. These parameters essentially do not affect the numerical stability.*

3.2.2 Self-consistent approach for solving the coupled PDEs

In this part, we propose a self-consistent approach to solve the coupled GLB and GPB equations for a given set of parameters. Basically, the coupled system is solved iteratively until both the electrostatic solvation free energy ΔG^{P} given in Eq. (3.1.9) and the surface function S are both converged. Here the surface function is said to be converged provided that the surface area and enclosed volume are both converged.

We present an algorithm for solving the following coupled systems:

$$-\nabla \cdot (\epsilon(S)\nabla\Phi) = S\rho_m, \tag{3.2.7}$$

and

$$\frac{\partial S}{\partial t} = |\nabla S| \left[\nabla \cdot \left(\gamma \frac{\nabla S}{|\nabla S|}\right) + V_e\right], \tag{3.2.8}$$

where V_e is the external potential which is defined as:

- **Auxiliary system:** $V_e = \frac{1}{2}(\epsilon_m - \epsilon_s)|\nabla\Phi|^2$,

- **Full system:** $V_e = -p + U + \frac{1}{2}(\epsilon_m - \epsilon_s)|\nabla\Phi|^2$.

Dirichlet boundary conditions are employed for both GPB (3.2.7) and GLB (3.2.8) equations with auxiliary and full external potentials, giving rise to a well-posed coupled system. The smooth profile of the solvent-solute boundary enables the direct use of the second order central finite difference scheme to achieve the second order convergence in discretizing the GPB equation. The biconjugate gradient scheme is used to solve the resulting algebraic equation system. The GLB equation of both the auxiliary and full systems can be solved by the central finite difference discretization of the spatial domain and the forward Euler time integrator for the time domain discretization.

Remark 3.2.2. *For the sake of simplicity, in the current work, we employed the central finite difference scheme for spatial domain discretization in both GPB and GLB equations, and forward Euler integrator for the time domain discretization of GLB equation. For stability consideration, in the discretization of the GLB equation, the discretization step size of temporal and spatial domain satisfies the Courant-Friedrichs-Lewy condition. To accelerate the numerical integration, a multigrid solver can be employed for GBP equation, and an alternating direction implicit scheme [43], which is unconditionally stable, can be utilized for the temporal integration. However, detail discussion of these accelerated schemes is beyond the scope of the present work.*

The coupled GLB and GPB equations is solved in a self-consistent manner. A dynamical coupling is needed two solve the coupled system.

Remark 3.2.3. *In solving the GLB equation, during each updating, to ensure the stability, instead of the fully update, we update it partially, i.e., the updated solution is the weighted sum of the new solution of the current GLB solution S_{new} and the old solution of the GLB* *equation in the previous step $S_{\rm old}$:*

$$S = a_1 S_{\rm new} + (1 - a_1) S_{\rm old}, \tag{3.2.9}$$

where a_1 is a constant and set to 0.5 in the present work.

3.2.3 Convex optimization for parameter learning

In this part, we present the parameter optimization scheme. In our approach, parameters start from an initial guess and then are updated sequentially until reaching the convergence. Here the convergence is measured by the RMSE between the fitted and experimental solvation free energies for a given set of molecules.

Consider the parameter optimization for a given group of molecules, $\{T_1, T_2, \cdots, T_n\}$. As discussed above the parameter set is $\mathbf{P}$. To optimize the parameter set $\mathbf{P}$, we start from GPB equation (3.2.7) and the auxiliary system of GLB equation (3.2.8) with $\gamma = 0.05$. After solving the initial coupled system by using Algorithm **??**, we obtain the following quantities for each molecule in the training set:

$$\Bigg\{\Delta \mathrm{G}^{\mathrm{P}}_{\mathrm{j}}, \mathrm{Area}_{\mathrm{j}}, \mathrm{Vol}_{\mathrm{j}}, \left(\sum_{i=1}^{N_m} \delta_i^1 \int_{\Omega_s} \left[\left(\frac{\sigma_s + \sigma_1}{||\mathbf{r} - \mathbf{r}_i||}\right)^{12} - 2\left(\frac{\sigma_s + \sigma_1}{||\mathbf{r} - \mathbf{r}_i||}\right)^{6}\right] d\mathbf{r}\right)_j, \tag{3.2.10}$$

$$\cdots, \tag{3.2.11}$$

$$\left(\sum_{i=1}^{N_m} \delta_i^{N_T} \int_{\Omega_s} \left[\left(\frac{\sigma_s + \sigma_{N_T}}{||\mathbf{r} - \mathbf{r}_i||}\right)^{12} - 2\left(\frac{\sigma_s + \sigma_{N_T}}{||\mathbf{r} - \mathbf{r}_i||}\right)^{6}\right] d\mathbf{r}\right)_j \Bigg\} \tag{3.2.12}$$

where $j = 1, 2, \cdots, n$. Here N_m and N_T denote the number of atoms and types of atoms in a specific molecule. The last few terms involve semi-discrete and semi-continuum Lennard-

Jones potentials [43]. Additionally,

$$\delta_i^j = \begin{cases} 1, & \text{if atom } i \text{ belongs to type } j, \\ 0, & \text{otherwise.} \end{cases}$$

where $i = 1, 2, \cdots, N_m$; $j = 1, 2, \cdots N_T$; $\sigma_i, i = 1, 2, \cdots, N_T$ is the atomic radius of the ith type of atoms. Therefore, atoms of the same type have a common atomic radius and fitting parameter $\tilde{\varepsilon}$.

The predicted solvation free energy for molecule j can be represented as:

$$\Delta \mathrm{G_j} = \Delta \mathrm{G_j^P} + \tilde{\varepsilon}_1 \left(\sum_{i=1}^{N_m} \delta_i^1 \int_{\Omega_s} \left[\left(\frac{\sigma_s + \sigma_1}{||\mathbf{r} - \mathbf{r}_i||} \right)^{12} - 2 \left(\frac{\sigma_s + \sigma_1}{||\mathbf{r} - \mathbf{r}_i||} \right)^6 \right] d\mathbf{r} \right)_j \tag{3.2.13}$$

$$+ \cdots + \tilde{\varepsilon}_{N_T} \left(\sum_{i=1}^{N_m} \delta_i^{N_T} \int_{\Omega_s} \left[\left(\frac{\sigma_s + \sigma_{N_T}}{||\mathbf{r} - \mathbf{r}_i||} \right)^{12} - 2 \left(\frac{\sigma_s + \sigma_{N_T}}{||\mathbf{r} - \mathbf{r}_i||} \right)^6 \right] d\mathbf{r} \right)_j \tag{3.2.14}$$

$$+ \gamma \mathrm{Area_j} + p \mathrm{Vol_j} \tag{3.2.15}$$

We denote the predicted solvation free energy for the given set of molecules as $\Delta \mathbf{G}(\mathbf{P}) \doteq \{\Delta G_1, \Delta G_2, \cdots, \Delta G_n\}$, which is a function of the parameter set $\mathbf{P}$, and denote the corresponding experimental solvation free energy as $\Delta \mathbf{G}^{\mathrm{Exp}} \doteq \left\{ \Delta G^{\mathrm{Exp1}}, \Delta G^{\mathrm{Exp2}}, \cdots, \Delta G^{\mathrm{Exp}n} \right\}$.

Then the parameter optimization problem in the coupled PDEs given by Eqs. (3.2.6) can be transformed into the following regularized and constrained optimization problem:

$$\min_{\mathbf{P}} \left(||\Delta \mathbf{G}(\mathbf{P}) - \Delta \mathbf{G}^{\mathrm{Exp}}||_2 + \lambda ||\mathbf{P}||_2 \right), \tag{3.2.16}$$

subject to

$$\gamma \geq \gamma_0, \tag{3.2.17}$$

and

$$|p| \leq \beta\gamma, \tag{3.2.18}$$

where $|| * ||_2$ is the L_2 norm of the quantity $*$ and λ is the regularization parameter chosen to be 10 in the present work to ensure the dominance of the first term and avoid over-fitting. Here γ_0 and β are set respectively to 0.05 and 0.1 in the present implementation, which guarantees the stability of the coupled system according to a large amount of numerical tests.

It is obvious that the objective function (3.2.16) in the optimization is a convex function, meanwhile the solution domain restricted by constraints (3.2.17)-(3.2.18) forms a convex domain. Therefore the optimization problem given by Eqs. (3.2.16)-(3.2.18) is a convex optimization problem, which was studied by Grant and Boyd [103, 102].

After solving the above convex optimization problem, parameter set $\mathbf{P}$ is updated and used again in solving the coupled GLB and GPB system, i.e., Eqs. (3.2.8) and (3.2.7). Repeating the above procedure, a new group of predicted solvation free energies together with a new group of parameters is obtained. This procedure is repeated until the RMSE between the predicted and experimental solvation free energies in two sequential iterations is within a given threshold.

3.2.4 Algorithm for parameter optimization and solution of the coupled PDEs

Based on the preparation made in the previous two subsections, namely, the self-consistent approach for solving the coupled GLB and GPB system and the parameter optimization, we provide the combined algorithm for the parameter optimization and solving the coupled system for a given set of molecules.

Numerically, to resolve the coupled PDEs and parameter optimization. A self consistent iteration is employed two solve the parameter optimization and coupled PDEs. In which the parameters optimization problem is solved in the outer iteration, while the coupled PDEs are solved in the inner iteration.

3.3 Numerical results

In this section we present the numerical study of the DG based solvation model using the proposed parameter optimization algorithms. We first explore the optimal solvent radius used in the van der Waals interactions. Due to the high nonlinearity, the solvent radius cannot be automatically optimized and its optimal value is obtained via searching the parameter domain. We show that for a group of molecules, there is a local minimum in the RMSE when the solvent radius is varied. The corresponding optimal solvent radius is adopted for other molecules. Additionally, we consider a large number of molecules with known experimental solvation free energies to test the proposed parameter optimization algorithms. These molecules are of both polar and nonpolar types and are divided into six groups: the SAMPL0 test set [185], the alkane, alkene, ether, alcohol and phenol types [175]. It is found that our DG based solvation model works really well for these molecules. Finally, to demonstrate the predictive power of the present DG based solvation model, we perform a five-fold cross validation [108] for alkane, alkene, ether, alcohol and phenol types of molecules. It is found that training and validation errors are of the same level, which confirms the ability of our model for the solvation free energy prediction.

The SAMPL0 molecule structural conformations are adopted from the literature with ZAP 9 radii and the OpenEye-AM1-BCC v1 charges [185]. For other molecules, structural conformations are obtained from FreeSolv [175]. Amber GAFF force field is utilized for the charge assignment [35]. The van der Waals radii as well as the atomic radii of Hydrogen, Carbon and Oxygen atoms are set to 1.2, 1.7 and 1.5 Å, respectively. The grid spacing is set to 0.25 Å in all of our calculations (discretization and integration). The computational domain is set to the bounding box of the solute molecule with an extra buffer length of 6.0 Å.

Table 3.1: The solvation free energy prediction for the SAMPL0 set. Energy is in the unit of kcal/mol.

Name	ΔG^{P}	G^{NP}	ΔG	ΔG^{Exp}[185]	Error
Glycerol triacetate	-10.60	2.53	-8.07	-8.84	-0.77
Benzyl bromide	-4.31	1.93	-2.38	-2.38	0.00
Benzyl chloride	-4.45	1.18	-3.27	-1.93	1.34
m-Bis (trifluoromethyl) benzene	-2.62	3.70	1.08	1.07	-0.01
N,N-Dimethyl-p-methoxybenzamide	-8.35	-2.22	-10.57	-11.01	-0.45
N,N-4-Trimethylbenzamide	-6.93	-3.09	-10.03	-9.76	0.27
bis-2-Chloroethyl ether	-3.73	-0.14	-3.59	-4.23	-0.64
1,1-Diacetoxyethane	-7.07	2.00	-5.07	-4.97	0.10
1,1-Diethoxyethane	-3.58	0.43	-3.15	-3.28	-0.13
1,4-Dioxane	-5.36	-0.38	-5.74	-5.05	0.69
Diethyl propanedioate	-7.07	1.40	-5.67	-6.00	-0.33
Dimethoxymethane	-4.09	1.19	-2.90	-2.93	-0.03
Ethylene glycol diacetate	-7.66	1.90	-5.76	-6.34	-0.58
1,2-Diethoxyethane	-3.64	0.45	-4.09	-3.54	0.55
Diethyl sulfide	-2.21	0.76	-1.47	-1.43	0.04
Phenyl formate	-7.10	2.08	-5.02	-4.08	0.94
Imidazole	-11.54	2.71	-8.83	-9.81	-0.98
RMSE					0.60

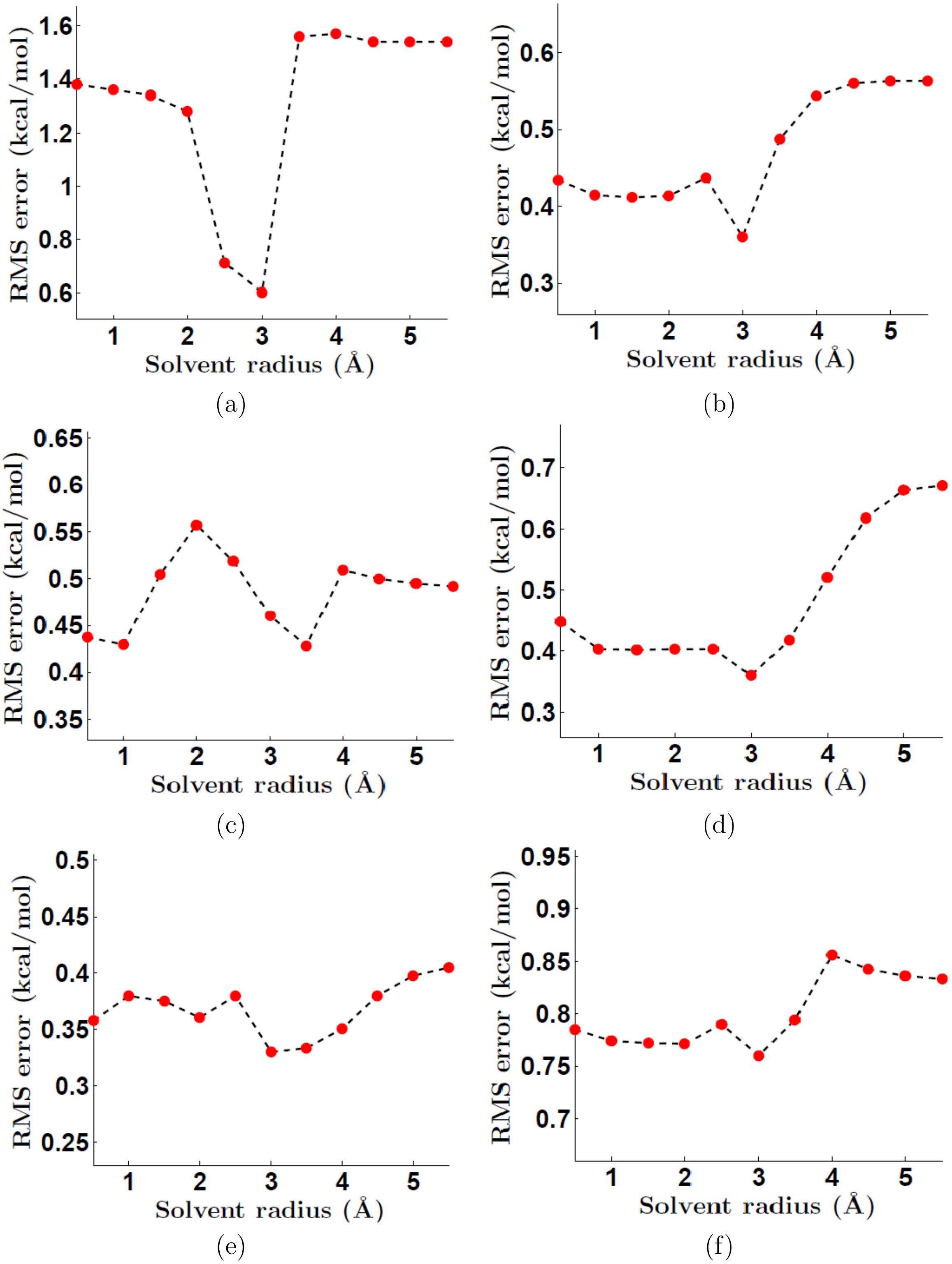


Figure 3.1: The relations between the solvent radii and the RMSEs. (a) SAMPL0 test set; (b) Alkane set; (c) Alkene set; (d) Ether set; (e) Alcohol; (f) Phenol set. Notably, there is a common local minimum at the solvent radii 3.0 Å for all test sets except for the alkene set.

3.3.1 Solvent radius

In the present semi-discrete and semi-continuum Lennard-Jones potential,

$$\tilde{\epsilon}_k \int_{\Omega_s} \left[\left(\frac{\sigma_s + \sigma_i}{||\mathbf{r} - \mathbf{r}_i||} \right)^{12} - 2 \left(\frac{\sigma_s + \sigma_i}{||\mathbf{r} - \mathbf{r}_i||} \right)^{6} \right] d\mathbf{r},$$

the positions $\mathbf{r}_i$, $(i = 1, 2, \cdots, N_m)$ are the coordinates of solute atoms, while $\mathbf{r}$ is not the position of a regular solvent atom or molecule. Since the solvent is treated as a continuum, $\mathbf{r}$ varies, in principle, continuously over the whole solvent domain. The distance $||\mathbf{r} - \mathbf{r}_i||$ is scaled by the sum of solvent radius σ_s and solute radii σ_i. Because of the explicit representation of solute atoms, solute atomic radii σ_i are set to their van der Waals radii, the radii that define the van der Waals surface, which is used for setting up the boundary condition for the GLB equation. However, the continuum treatment of the solvent prevents us to simply associate σ_s with the radius of the solvent molecule. Unlike the the fully discrete Lennard-Jones potential in explicit solvent models, the semi-discrete and semi-continuum Lennard-Jones potential in our DG based solvation model describes the "interaction" of a solute atom with an arbitrary position in the solvent domain. In numerical approximation, the arbitrary position is represented by a grid mesh. Therefore, one cannot simply take the solvent radius in the present model as the radius of individual (discrete) solvent molecules. Additionally, it is noted that the solvent radius in the present work and solvent probe radius in the Poisson-Boltzmann theory are two different concepts. In the present work, solvent radius σ_s is considered as an optimization parameter. Note that due to the nonlinear nature, this optimization cannot be carried out together or mixed with the parameter optimization discussed in the earlier section.

We utilize a brute force approach for the solvent radius selection or optimization. Six

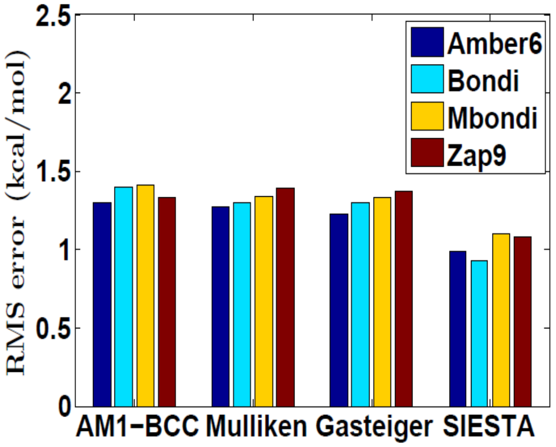


Figure 3.2: The predicted and experimental solvation free energy for the 17 molecules in the SAMPL0 test set.

sets of test examples are utilized to explore appropriate solvent radius. The SAMPL0 test set [185] is a benchmark having 17 molecules. Additionally, we consider 38 alkane, 22 alkene, 17 ether, 25 alcohol, and 18 phenol molecules. The solvent radius is varied from 0.5 Å to 5.5 Å away from van der Waals surface. Due to the fast decay property of the Lennard-Jones interactions, the above setting enables the full inclusion of the Lennard-Jones interactions in our model. Figure 3.1 depicts the RMSEs of six test sets at different solvent radii calculated from the present DG based solvation model. In Figure 3.1 (a), the result clearly demonstrates that with the increase of the solvent radius, the RMSE decreases dramatically initially. The minimum appears at 3.0 Å. The further increase of the solvent radius leads to a rapid jump in the RMSE before it stabilizes around 1.54 kcal/mol. It is noted that 3.0 Å is much larger than the commonly used solvent probe radius of 1.4 Å in Poisson-Boltzmann theory based implicit solvent models. For other five test sets, although the behavior of the RMSE differs in each case, essentially all the RMSEs have a local minimum at the solvent radius of 3 Å. Therefore, in all the following computations, the solvent radius is set to 3.0 Å.

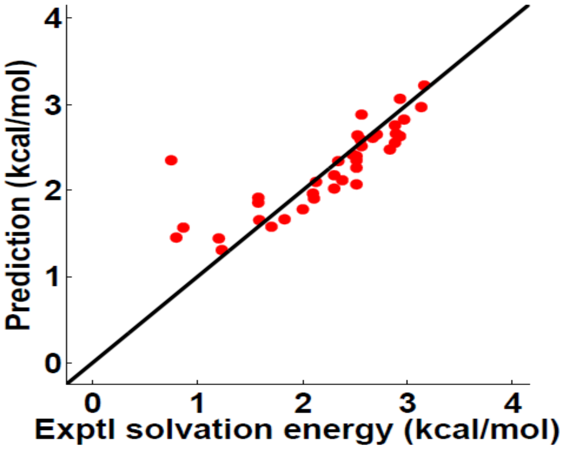


Figure 3.3: The predicted and experimental solvation free energies for 38 alkane molecules.

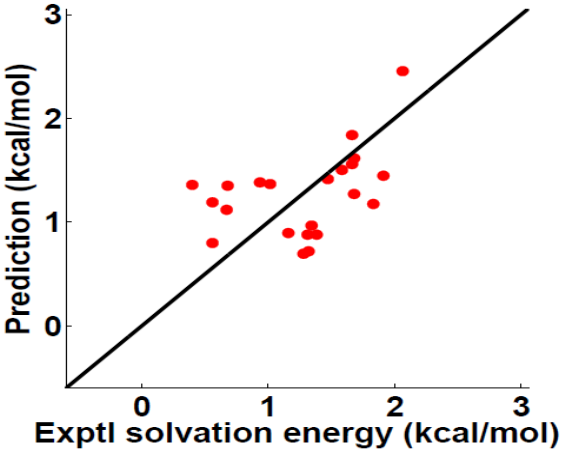


Figure 3.4: The predicted and experimental solvation free energies for 22 alkene molecules.

3.3.2 Optimization results

In this section, we illustrate the performance of our parameter optimization algorithms. First, we provide the regression results of the SAMPL0 test set [185]. Figure 3.2 shows the predicted and experimental solvation free energies based on the present model and optimization method. It is obvious that predicted solvation free energies are highly consistent with the experimental ones. The RMSE is 0.60 kcal/mol.

Table 3.1 shows the breakup of polar, non-polar and total predicted solvation free energies. The experimental values and errors are also provided [185].

Table 3.2: The solvation free energy prediction for the alkane set. All energies are in the unit of kcal/mol.

Name	ΔG^{P}	G^{NP}	ΔG	ΔG^{Exp}[175]	Error
octane	-0.13	2.89	2.76	2.88	0.12
ethane	-0.04	1.70	1.66	1.83	0.17
propane	-0.05	1.83	1.78	2.00	0.22
cyclopropane	-0.08	2.43	2.35	0.75	-1.60
isobutane	-0.07	2.09	2.02	2.30	0.28
2,2-dimethylbutane	-0.07	2.34	2.27	2.51	0.24
isopentane	-0.07	2.19	2.12	2.38	0.26
2,3-dimethylbutane	-0.07	2.41	2.34	2.34	0.00
3-methylpentane	-0.08	2.43	2.35	2.51	0.16
methylcyclopentane	-0.10	1.76	1.66	1.59	-0.07
n-butane	-0.07	2.03	1.96	2.10	0.14
isohexane	-0.09	2.49	2.40	2.51	0.11
2,4-dimethylpentane	-0.09	2.57	2.48	2.83	0.35
methylcyclohexane	-0.10	1.68	1.58	1.70	0.12
n-pentane	-0.08	2.25	2.17	2.30	0.13
hexane	-0.09	2.51	2.42	2.48	0.06
cyclohexane	-0.10	1.40	1.30	1.23	-0.07
nonane	-0.14	3.11	2.97	3.13	0.16
heptane	-0.11	2.73	2.62	2.67	0.05
cyclopentane	-0.10	1.54	1.44	1.20	-0.24
cycloheptane	-0.11	1.56	1.45	0.80	-0.65
cyclooctane	-0.12	1.69	1.57	0.86	-0.71
neopentane	-0.06	2.13	2.07	2.51	0.44
2,2,4-trimethylpentane	-0.08	2.74	2.66	2.89	0.23
3,3-dimethylpentane	-0.07	2.58	2.51	2.56	0.05
2,3-dimethylpentane	-0.08	2.72	2.64	2.52	-0.12
2,3,4-trimethylpentane	-0.08	2.96	2.88	2.56	-0.32
1,2-dimethylcyclohexane	-0.10	2.02	1.92	1.58	-0.34
3-methylhexane	-0.09	2.74	2.65	2.71	0.06
3-methylheptane	-0.11	2.94	2.83	2.97	0.14
1,4-dimethylcyclohexane	-0.11	2.02	1.91	2.11	0.20
2,2-dimethylpentane	-0.08	2.64	2.56	2.88	0.32
2-methylhexane	-0.10	2.73	2.63	2.93	0.30
decane	-0.16	3.37	3.21	3.16	-0.06
propylcyclopentane	-0.12	2.21	2.09	2.13	0.03
cis-1,2-Dimethylcyclohexane	-0.09	1.95	1.86	1.58	-0.28
2,2,5-trimethylhexane	-0.09	3.15	3.06	2.93	-0.13
pentylcyclopentane	-0.15	2.73	2.58	2.55	-0.04
RMSE					0.36

Compared to our earlier prediction [43] in which the same model is employed but the parameters were not optimized in the present manner, the RMSE decreases dramatically from previous 1.76 kcal/mol to 0.60 kcal/mol for the same test set. Note that the present RMSE (0.60 kcal/mol) is also significantly smaller than that of the explicit solvent approach (1.71 ± 0.05 kcal/mol) and that obtained by the PB based prediction (1.87 kcal/mol) under the same structure, charge and radius setting [185]. The present results confirm the efficiency of the proposed new parameter optimization algorithms and demonstrate the accuracy and power of our DG based solvation models.

Table 3.3: The solvation free energy prediction for the alkene set. All energies are in the unit of kcal/mol.

Name	$\Delta G^{\rm P}$	$G^{\rm NP}$	ΔG	$\Delta G^{\rm Exp}$[175]	Error
ethylene	-0.27	0.96	0.69	1.28	0.59
isoprene	-0.62	1.97	1.35	0.68	-0.67
but-1-ene	-0.29	1.17	0.88	1.38	0.50
butadiene	-0.56	1.75	1.19	0.56	-0.63
pent-1-ene	-0.30	1.57	1.27	1.68	0.41
prop-1-ene	-0.32	1.03	0.71	1.32	0.61
2-methylprop-1-ene	-0.37	1.26	0.89	1.16	0.27
cyclopentene	-0.37	1.17	0.79	0.56	-0.23
2-methylbut-2-ene	-0.40	1.28	0.87	1.31	0.44
2,3-dimethylbuta-1,3-diene	-0.65	2.01	1.36	0.40	-0.95
3-methylbut-1-ene	-0.27	1.45	1.18	1.83	0.65
1-methylcyclohexene	-0.38	1.50	1.11	0.67	-0.45
penta-1,4-diene	-0.53	1.91	1.38	0.93	-0.45
hex-1-ene	-0.30	1.81	1.50	1.58	0.08
hexa-1,5-diene	-0.51	1.88	1.37	1.01	-0.36
hept-1-ene	-0.33	2.17	1.84	1.66	-0.18
hept-2-ene	-0.34	1.96	1.62	1.68	0.06
4-Methyl-1-pentene	-0.26	1.71	1.45	1.91	0.46
2-methylpent-1-ene	-0.33	1.75	1.42	1.47	0.05
non-1-ene	-0.36	2.81	2.45	2.06	-0.39
trans-2-Heptene	-0.34	1.90	1.56	1.66	0.10
trans-2-Pentene	-0.30	1.26	0.96	1.34	0.38
RMSE					0.46

Additionally, we investigate the solvation free energies prediction of two families of nonpolar molecules, alkane and alkene, which were studied previous by using the DG based nonpolar solvation model [46]. In the following, we demonstrate that the present DG based full solvation model can provide the same level of accuracy in the solvation free energy prediction for alkane and alkene molecules.

Figures 3.3 and 3.4 depict the predicted and experimental solvation free energies for 38 alkane and 22 alkene molecules, respectively. Tables 3.2 and 3.3 list the polar, nonpolar, total and experimental solvation free energies for both families of solute molecules, respectively. Except for one alkane molecule, namely, cycloprotane, whose predicted error is 1.60 kcal/mol, the errors for all other molecules are within 1 kcal/mol. The RMSEs of these two families are 0.36 and 0.46 kcal/mol, respectively. This level of accuracy is similar to our earlier results obtained by using our DG based nonpolar solvation model [46], which does not involve the electrostatic (polar) model and is computationally easier to optimize.

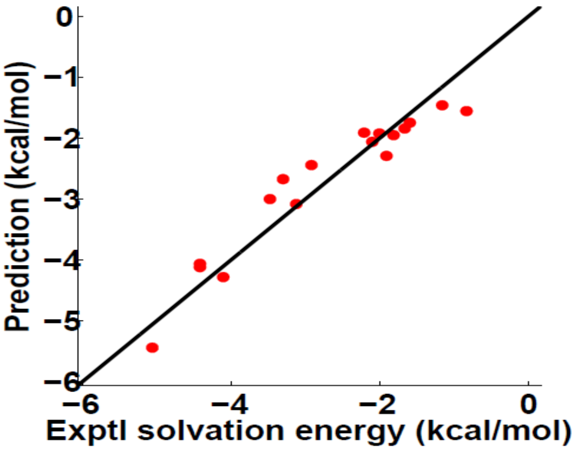


Figure 3.5: The predicted and experimental solvation free energy for the 17 ether molecules.

It is interesting to note that for both alkane and alkene molecules, the polar solvation free energy contribution is very small and the nonpolar part dominates the solvation free energy contribution, which explains why the DG based nonpolar solvation model works extremely well for the solvation free energy prediction of alkane and alkene molecules [46]. Further,

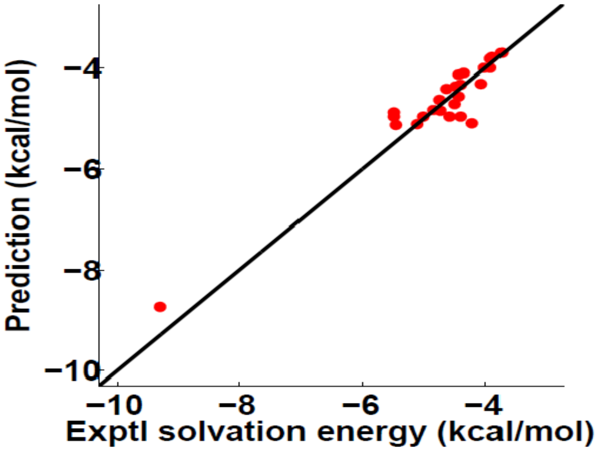


Figure 3.6: The predicted and experimental solvation free energy for the 25 alcohol molecules.

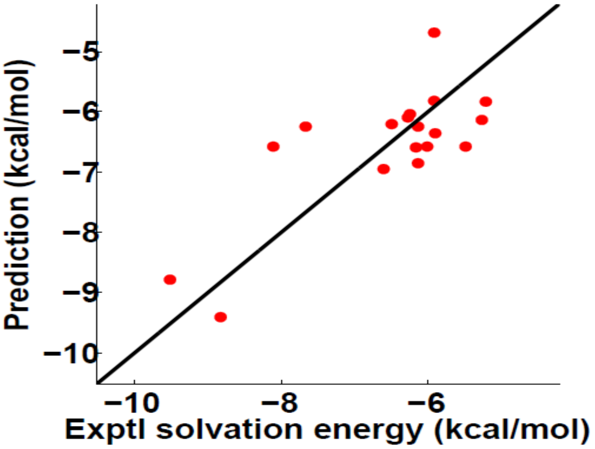


Figure 3.7: The predicted and experimental solvation free energy for the 18 phenol molecules.

note that for almost all the alkane molecules, the polar solvation free energies ΔG^{P} are of magnitude 0.01 kcal/mol, while alkene molecules have slightly larger magnitude polar free energies, which further verifies that alkene molecules has a stronger polarity than alkane molecules in general.

Finally, we analyze three classes of polar solute molecules, namely, ether, alcohol, and phenol molecules. Figures 3.5, 3.6 and 3.7 illustrate the predicted and experimental solvation free energies for 17 ether, 25 alcohol, and 18 phenol molecules, respectively. Tables 3.4, 3.5 and 3.6 list the polar, nonpolar, total and experimental solvation free energies for the

Table 3.4: The solvation free energy prediction for the ether set. All energies are in the unit of kcal/mol.

Name	ΔG^{P}	G^{NP}	ΔG	ΔG^{Exp}[175]	Error
ethoxyethane	-4.08	2.33	-1.75	-1.59	0.16
2-methyltetrahydrofuran	-4.10	1.43	-2.67	-3.30	-0.63
tetrahydrofuran	-4.36	1.36	-3.00	-3.47	-0.47
1-propoxypropane	-3.75	2.29	-1.46	-1.16	0.30
methoxymethane	-4.55	2.26	-2.29	-1.91	0.36
tetrahydropyran	-4.17	1.09	-3.07	-3.12	-0.05
1-butoxybutane	-3.88	2.33	-1.55	-0.83	0.72
trimethoxymethane	-7.57	3.51	-4.06	-4.42	-0.36
methoxyethane	-4.35	2.29	-2.06	-2.10	-0.04
1-methoxypropane	-4.08	2.24	-1.84	-1.66	0.18
2-methoxypropane	-4.12	2.20	-1.92	-2.01	-0.09
1-Ethoxypropane	-4.26	2.32	-1.94	-1.81	0.13
1,3-Dioxolane	-6.09	1.81	-4.28	-4.10	0.18
2,5-dimethyltetrahydrofuran	-3.86	1.42	-2.44	-2.92	-0.48
1,1,1-trimethoxyethane	-7.58	3.46	-4.12	-4.42	-0.30
2-methoxy-2-methyl-propane	-3.88	1.97	-1.91	-2.21	-0.30
1,4-dioxane	-7.09	1.66	-5.44	-5.06	0.38
RMSE					0.36

corresponding families of solute molecules. The RMSEs of these three families are 0.36, 0.33, and 0.76 kcal/mol, respectively.

From the results listed in Tables 3.4, 3.5 and 3.6 we note that for ether molecules, all the nonpolar energies are positive which neutralizes some polar contributions to the total solvation free energies. For the alcohol molecules, the nonpolar energies are all negative, which enhance the contributions of the polar contributions to the total solvation free energies. Since the surface part is always positive and the volume part is mostly positive, the attractive van der Waals interactions between alcohol molecules and water solvent must be very strong. Physically, there are strong solvent-solute hydrogen bonds that make alcohol molecules easily solvated. These solvent-solute interaction are described by the strong attractive van der Waals interactions in the present model. As for the phenol molecules, there is a mixed

Table 3.5: The solvation free energy prediction for the alcohol set. All energies are in the unit of kcal/mol.

Name	ΔG^{P}	G^{NP}	ΔG	ΔG^{Exp}[175]	Error
ethylene glycol	-6.98	-1.76	-8.73	-9.30	-0.57
butan-1-ol	-3.33	-1.51	-4.84	-4.72	0.12
ethanol	-3.49	-1.47	-4.96	-5.00	-0.04
methanol	-3.69	-1.41	-5.10	-5.10	0.00
propan-1-ol	-3.34	-1.48	-4.82	-4.85	-0.03
propan-2-ol	-3.26	-1.36	-4.62	-4.74	-0.12
pentan-1-ol	-3.36	-1.61	-4.97	-4.57	0.40
2-methylpropan-2-ol	-3.10	-1.27	-4.37	-4.47	-0.10
2-methylbutan-2-ol	-2.95	-1.17	-4.12	-4.43	-0.31
2-methylpropan-1-ol	-3.20	-1.50	-4.70	-4.50	0.20
butan-2-ol	-3.09	-1.32	-4.40	-4.62	-0.22
cyclopentanol	-3.20	-1.68	-4.88	-5.49	-0.61
4-methylpentan-2-ol	-2.65	-1.05	-3.69	-3.73	-0.04
cyclohexanol	-3.21	-1.92	-5.13	-5.46	-0.33
hexan-1-ol	-3.43	-1.53	-4.96	-4.40	0.56
heptan-1-ol	-3.48	-1.62	-5.09	-4.21	0.88
2-methylbutan-1-ol	-3.27	-1.29	-4.56	-4.42	0.14
cycloheptanol	-3.07	-1.89	-4.96	-5.48	-0.52
2-methylpentan-3-ol	-2.86	-0.93	-3.78	-3.88	-0.10
pentan-3-ol	-3.01	-1.08	-4.10	-4.35	-0.25
4-Heptanol	-2.90	-1.10	-3.99	-4.01	-0.02
2-methylpentan-2-ol	-2.93	-1.08	-4.00	-3.92	0.08
2,3-Dimethyl-2-butanol	-2.89	-0.93	-3.82	-3.91	-0.09
hexan-3-ol	-3.04	-1.27	-4.31	-4.06	0.25
pentan-2-ol	-3.10	-1.23	-4.33	-4.39	-0.06
RMSE					0.33

pattern for the nonpolar contributions.

The above study of a large variety of molecules indicates that the DG based solvation model together with the proposed parameter optimization algorithms can provide very accurate predictions of solvation free energies for both polar and nonpolar solute molecules.

Table 3.6: The solvation free energy prediction for the phenol set. All energies are in the unit of kcal/mol.

Name	ΔG^{P}	G^{NP}	ΔG	ΔG^{Exp}[175]	Error
3-hydroxybenzaldehyde	-9.17	0.39	-8.78	-9.52	-0.74
4-hydroxybenzaldehyde	-9.60	0.19	-9.41	-8.83	0.58
o-cresol	-5.32	-1.04	-6.36	-5.90	0.46
m-cresol	-5.71	-0.86	-6.57	-5.49	1.08
phenol	-5.81	-0.14	-6.95	-6.61	0.34
p-cresol	-5.80	-1.05	-6.85	-6.13	0.72
naphthalen-1-ol	-5.50	-0.75	-6.25	-7.67	-1.42
3,4-dimethylphenol	-5.72	-0.49	-6.21	-6.50	-0.29
2,5-dimethylphenol	-5.34	-0.48	-5.82	-5.91	-0.09
4-tert-butylphenol	-5.55	0.86	-4.69	-5.91	-1.22
2,4-dimethylphenol	-5.55	-1.03	-6.58	-6.01	0.57
3,5-dimethylphenol	-5.69	-0.41	-6.10	-6.27	-0.17
naphthalen-2-ol	-5.85	-0.72	-6.57	-8.11	-1.54
2,3-dimethylphenol	-5.47	-1.13	-6.60	-6.16	0.44
2,6-dimethylphenol	-5.07	-1.07	-6.14	-5.26	0.88
3-ethylphenol	-5.67	-0.37	-6.04	-6.25	-0.21
4-propylphenol	-5.79	-0.05	-5.84	-5.21	0.63
4-ethylphenol	-5.76	-0.48	-6.24	-6.13	0.11
RMSE					0.76

Table 3.7: The partition of the molecules into sub-groups.

Molecule	Group 1	Group 2	Group 3	Group 4	Group 5
Alkane	8	8	8	7	7
Alkene	5	5	5	4	4
Ether	4	4	3	3	3
Alcohol	5	5	5	5	5
Phenol	4	4	4	3	3

3.3.3 Five-fold cross validation

Having verified that the DG based solvation model with the optimized parameters provides very good regression results, we perform a five-fold cross validation to further illustrates the predictive power of the present method for independent data sets. Specifically, the parameters learned from a group of molecules can be employed for the blind prediction of other molecules.

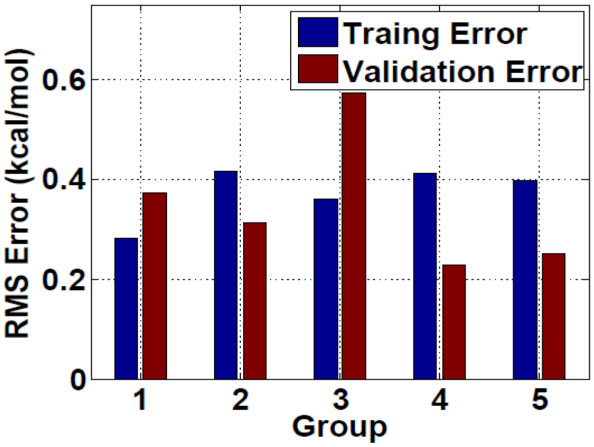


Figure 3.8: The bar plot of the training and validation errors of alkanes.

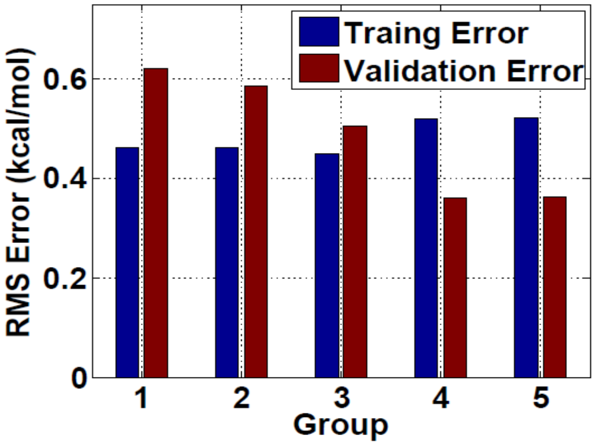


Figure 3.9: The bar plot of the training and validation errors of alkenes.

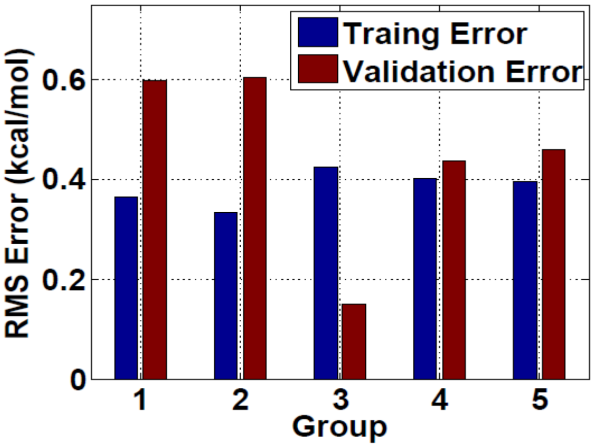


Figure 3.10: The bar plot of the training and validation errors of the ethers.

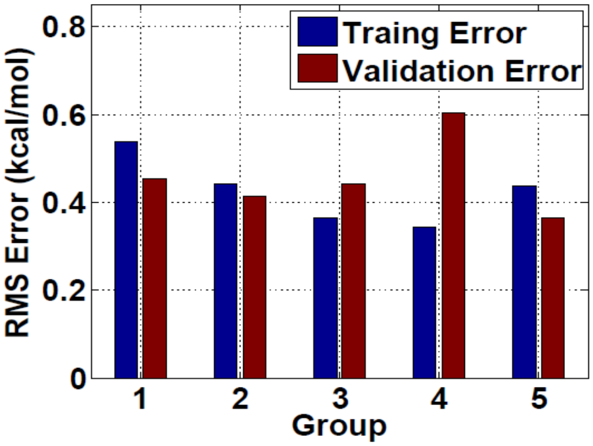


Figure 3.11: The bar plot of the training and validation errors of alcohols.

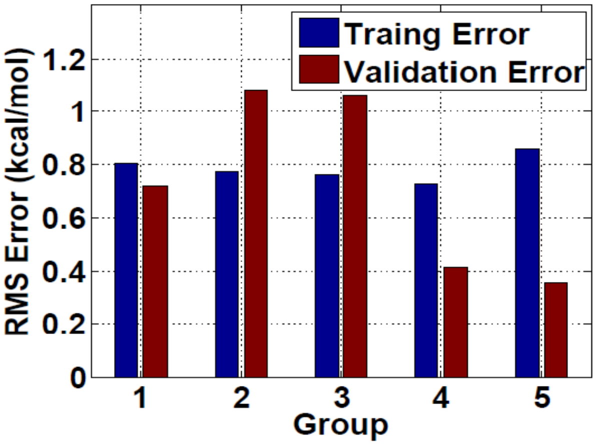


Figure 3.12: The bar plot of the training and validation errors of phenols.

To perform the five-fold cross validation, each type of molecules is subdivided into five sub-groups as uniformly as possible, Table 3.7 lists the number of molecules in each sub-group for each type of molecules. In our parameters optimization, we leave out one sub-group of molecules and use the rest of molecules to establish our DG based solvation model. The optimized parameters are then employed for the blind prediction of solvation free energies of the left out sub-group of molecules.

Figures 3.8, 3.9,3.10, 3.11, and 3.12 demonstrate the cross validation results of the alkane, alkene, ether, alcohol, and phenol molecules, respectively. It is seen that training and validation errors are similar to each other, which verifies the ability of our model in the blind prediction of solvation free energies.

In the real prediction of the solvation free energy for a given molecule of unknown category, we can first assign it to a given group, and then employ the DG based solvation model with the optimal parameters learned for this specific group for a blind prediction.

3.4 Conclusion

Differential geometry (DG) based solvation models have had a considerable success in solvation analysis [269, 43, 44, 45]. Particularly, the DG based nonpolar solvation model was shown to offer some of the most accurate solvation energy predictions of various nonpolar molecules [46]. However, the DG based full solvation model is subject to numerical instability in solving the generalized Laplace-Beltrami (GLB) equation, due to its coupling with the GPB equation. To stabilize the coupled GLB and GPB equations, a strong constraint on the van der Waals interaction was applied in our earlier work [43, 44, 45], which hinders the parameter optimization of our DG based solvation model. In the present work, we resolve this problem by introducing new parameter optimization algorithms, namely perturbation method and convex optimization, for the DG based solvation model. New stability conditions are explicitly imposed to the parameter selection, which guarantees the stability and robustness of solving the GLB equation and leads to constrained optimization of the DG based solvation model. The new optimization algorithms are intensively validated by using a large number of test molecules, including the SAMPL0 test set [185], alkane, alkene, ether, alcohol and phenol types of solutes. Regression results based on our new algorithms are consistent extremely well with experimental data. Additionally, a five-fold cross validation technique is employed to explore the ability of the DG based solvation models for the blind prediction of the solvation free energies for a variety of solute molecules. It is found that the same level of errors is found in the training and validation sets, which confirms our model's predictive power in solvation free energy analysis. The present DG based full solvation model provides a unified framework for analyzing both polar and nonploar molecules.

Nevertheless, the capability of the DG based solvation model for blind solvation free energy prediction for general molecules is still quite limited. The blind solvation free energy prediction by some other models will be further discussed in the following chapters.

Chapter 4

Molecular Conformation in Solvent

4.1 Introduction

In the previous chapters, we briefly presented several solvation models. The molecular conformation modeling is one of the most fundamental issues that we need to resolve, especially for the implicit solvation models. In the differential geometry based solvation model, the molecular conformation is modeled based on the variational principle. However, for general solvation models this modeling is inappropriate. In this chapter we present the models for the molecular conformation modeling in solvent. The solvated molecular is usually characterized as the cavity enclosed by the molecular surface , and among the literature there are mainly three categories of surface definitions:

- van der Waals surface [53], it is the surface defined as the surface created when each atom is represented by a sphere with a radius equal to the van der Waals radius of that atom. The VDW surface for a molecule is the union of all the individual van der Waals spheres.

- Solvent Accessible Surface (SAS) [147, 206], it is defined as the trace of the probe center, when a probe is used to roll along the atoms' van der Waals spheres.

- Solvent Excluded Surface (SES) [51, 52], it is defined as the surface traced by the inward-facing surface of the probe, when a probe is used to roll along the atom spheres.

It is composed of two parts: contact surface, is the part of the van der Waals surface that can be touched by the probe; reentrant surface, is formed by the inward-facing part of the probe when it is in contact with more than one atom.

The SES is generally accepted to be the most accurate molecular conformation model in the implicit solvent model community[209, 94, 259]. Geometrically, SES is relatively smooth compared to the other two popular surface definitions, nevertheless, it still takes some geometric singularities, such as tips and cusps.

Many efforts have been paid in the past decades for developing the analytical, robust, and efficient SES software. The pioneer work is due to Connolly [32, 51, 52], he formulated the mathematical representation of the SES for arbitrary biomolecules in terms of convex surfaces, saddle surfaces, and concave surfaces. State-of-the-art triangulated SES software, MSMS, is developed by Sanner et al. [212], the MSMS software provides very fast SES generation and triangulation. Many works have also been done for the development of approximated SES, for instance, the Smoothed Numerical Surface (SNS) which is used in the Delphi software [209] for the biomolecular electrostatics modeling; GEPOL which is the widely used approximation of SES in the polarizable continuum model [238].

Besides the above three surface definitions, for the computational convenience, many other surface definitions are also introduced. Especially, the Gaussian surface which is introduced to relieve the geometric singularity problem [159]. Many variants of the Gaussian surface have been proposed for different purposes [21].

In this chapter, we will present our work on the development of the Eulerian representation of SES, which is convenient for the Cartesian mesh based numerical methods. Furthermore, it provides a paradigm for surface area, molecular volume, and molecular topological analysis.

This chapter is structured as following. Section 4.2 introduces the algorithm for Eulerian solvent excluded surface (ESES) generation, which contains two parts, SES generation and embedding the SES to the Cartesian mesh. Section 4.3 discusses the area and volume calculation for the ESES. Section 4.4 addresses some topological structure issues on the ESES. The electrostatics of the solvated molecule analysis will be discussed in Section 4.5. This chapter ends up with a conclusion.

4.2 Eulerian Solvent Excluded Surface (ESES)

Our algorithm is designed to develop an analytical Eulerian representation of SES, this amounts to classify the Cartesian grid points with respect to the analytical SES, accurately compute the locations of intersection points between the interface and Cartesian mesh lines, and calculate the associated outer normal directions. There are mainly three steps in this framework. First, an analytical SES is built. Second, all the Cartesian mesh grids are classified as either inside or outside the surface. Third, the intersections points and associated outer normal directions are computed for each edge with one points inside and the other one outside. We will discuss these three steps in detail in the following subsections.

4.2.1 Construction of Solvent Excluded Surface

In the first step, an analytical SES is constructed mainly based on the algorithm proposed by Connolly [32, 51]. In the algorithm, the SES is divided into three types of patches, namely, convex patches, saddle patches, and concave patches. As shown in Fig. 4.1, where three different patches are rendered by different colors. The convex patches (red) are the accessible atomic surface by a spherical probe rolling around; saddle patches (green) are the

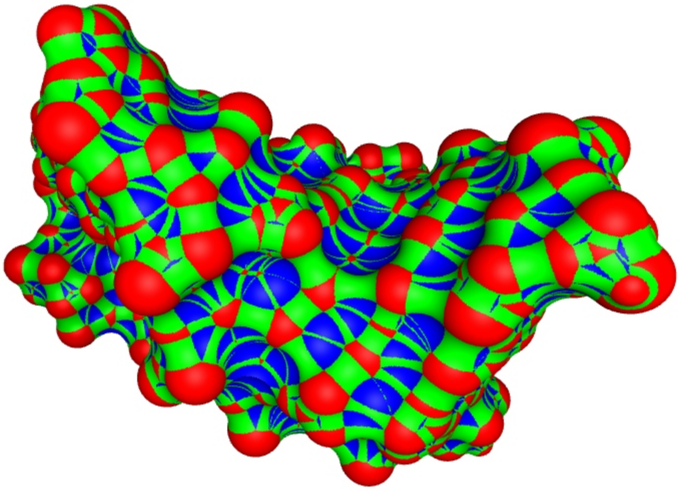


Figure 4.1: Three types of patches for SES: convex patches (red), saddle patches (green) and concave patches (blue).

trace of the probe inward-facing surface touching with two atoms at the same time; concave patches (blue) are the spherical triangle faces of the probe touching three or more atoms simultaneously.

The boundary curves of the convex patches are marked as convex edges, and the boundary curves of the concave patches are concave edges. Corresponding concave edges and convex edges compose the saddle patches. When a sadle/torous patch is free, its boundary convex edges are the complete contact circles. If there are no associated concave edges generated for the non-free torus, this torus is marked as blocked. Connolly [32, 51] provides the algorithm to compute the centers, radii, and boundary planes/edges/points of all the possible patches, where we perform torus construction, probe placement, and saddle face construction processes in order.

Note that when an atom is detected as an interior atom (i.e., completely buried by other atoms), it will not be considered for the torus construction. Additionally, to make the algorithm compatible with the case when the probe touches more than three atoms at the same time, a special treatment of the concave edges is needed after probe placement.

For a pair of the concave edges that are opposite to each other, these two concave edges are eliminated and their corresponding concave patches are tagged as belonging to the same probe sphere. For the concave edges that are the same, only one concave edge is kept and their corresponding concave faces are also tagged as a subset of the same probe sphere.

4.2.2 Classification of Cartesian Grid Points

In the second step, we classify Cartesian grid points in the computational domain as either inside or outside the SES. Apparently, Cartesian grid points that are outside the collection of the augmented atomic spheres can be safely labeled as outside points, where the radius of the augmented atom is expanded by r_p with r_p being the radius of the probe. Thus we only need to pay extra attention to the Cartesian grid points that are included by the augmented atoms. Upon careful observation, a Cartesian grid point can be classified as inside of the molecular surface if it satisfies

$$\text{Inside Atom} \mid [(\text{InsideVS} \mid \text{InsideVT}) \ \& \ !(\text{Inside Saddle Probe}) \ \& \ !(\text{Inside Concave Probe})] \tag{4.2.1}$$

Each component in the expression given by Eq.(4.2.1) will be discussed below in details. Note that the above equation is only a sufficient condition, further ray-tracing techniques is adapted for determining the status of the remaining grid points.

4.2.2.1 Inside Atom

Let an atom centered at $\mathbf{C}_i = (C_{i,x}, C_{i,y}, C_{i,z})$ with radius R_i, for a given Cartesian grid point $\mathbf{P} = (P_x, P_y, P_z)$. **Inside Atom** is true if:

$$|\mathbf{C}_i - \mathbf{P}|^2 - R_i^2 \leq 0. \tag{4.2.2}$$

Otherwise, the grid point $\mathbf{P}$ is outside the atom, i.e., **Inside Atom** is false.

4.2.2.2 InsideVS and Inside Saddle Probe

InsideVS is an abbreviation for inside the visibility sphere of the saddle patch, which is proposed in Krone's work [141]. The center and the radius of the visibility sphere can be computed by Eq. (4.2.3) and Eq. (4.2.4), respectively.

$$\mathbf{C}_{vs} = \frac{|\mathbf{C}_p - \mathbf{C}_i|}{|\mathbf{C}_p - \mathbf{C}_i| + |\mathbf{C}_p - \mathbf{C}_j|}\mathbf{C}_j + \frac{|\mathbf{C}_p - \mathbf{C}_j|}{|\mathbf{C}_p - \mathbf{C}_i| + |\mathbf{C}_p - \mathbf{C}_j|}\mathbf{C}_i, \tag{4.2.3}$$

$$R_{vs} = |\frac{R_i}{|\mathbf{C}_p - \mathbf{C}_i|}(\mathbf{C}_p - \mathbf{C}_i) + \mathbf{C}_i - \mathbf{C}_{vs}|, \tag{4.2.4}$$

where $\mathbf{C}_{vs}$ and R_{vs} stand for the center and radius of the visibility sphere, respectively. $\mathbf{C}_i$, $\mathbf{C}_j$, and $\mathbf{C}_p$ denote the center of i-th, j-th atom, and the probe, respectively. $| * |$ means the magnitude of the vector $*$. R_i is the radius of the i-th atom. As illustrated in Fig. 4.2.

From Fig. 4.2 we can see that **InsideVS** includes some possible interior points introduced by the saddle patches. **Inside Saddle Probe** eliminates the points that are accessible to the probe (the purple area). Note that when the current saddle patch is not free, we only check the part of the visibility sphere that is located within the range of saddle faces.

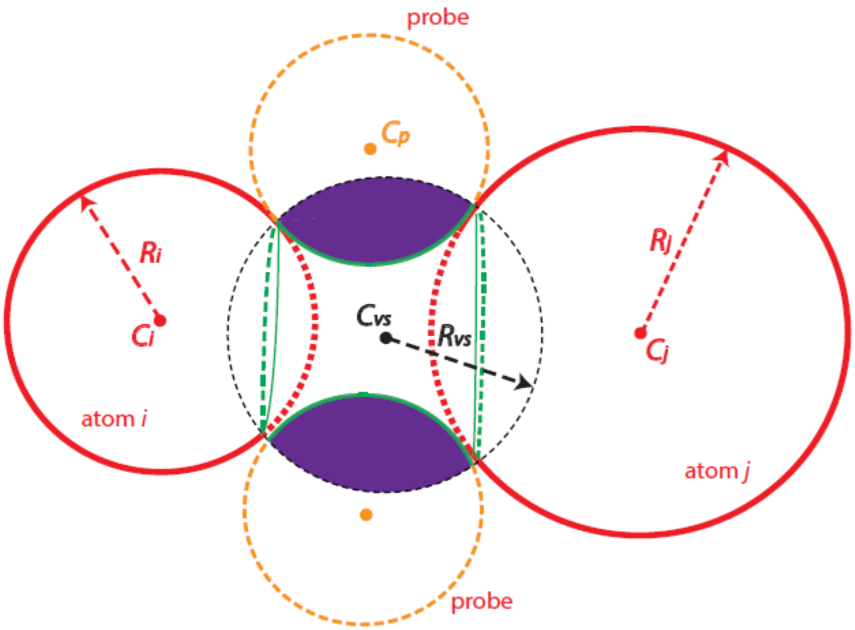


Figure 4.2: Figure for InsideVS. Solid lines are the outline of a simple SES composed of two atoms (red) and one saddle patch (green). Black circle is the outline of the corresponding visibility sphere when the probe touches atom *i* and atom *j* simultaneously. InsideVS is tagged as *true* when the point is detected as inside the visibility sphere.

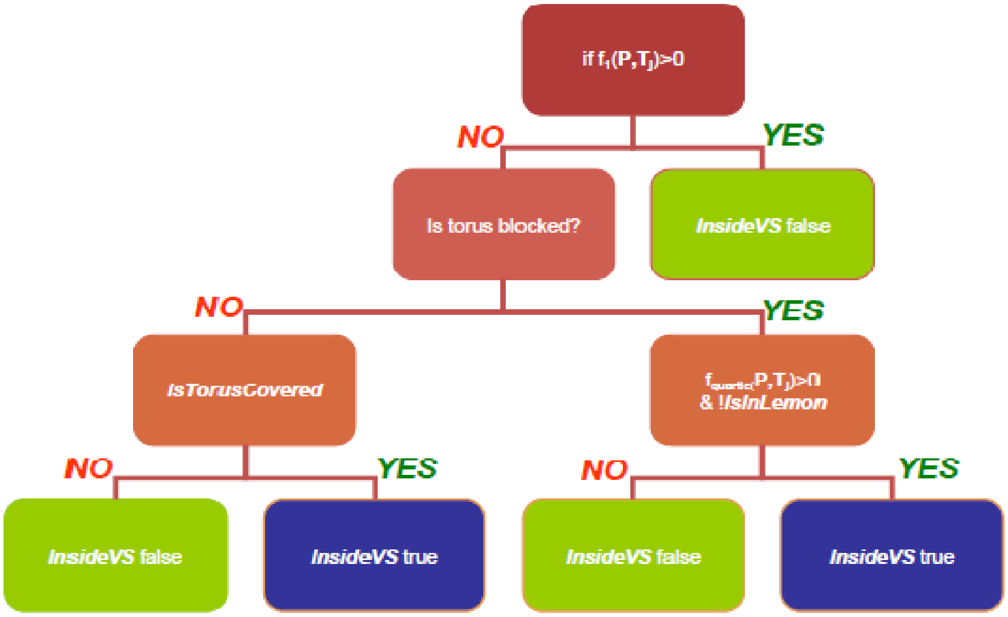


Figure 4.3: Flowchart for InsideVS.

Figures 4.3 and 4.4 depict the flowchart for determining the conditions **InsideVS** and **Inside Saddle Probe**, respectively. The equations involved in the above two flowcharts are

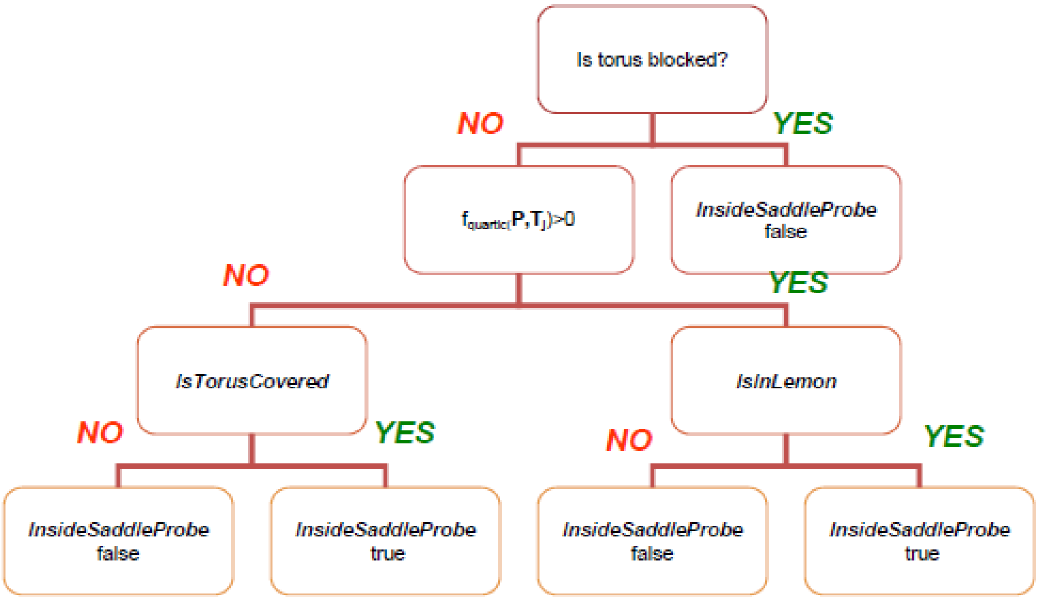


Figure 4.4: Flowchart for InsideTorus.

listed below:

$$f_1(\mathbf{P}, T_j) = |\mathbf{P} - \mathbf{C}_{T_j,VS}|^2 - R^2_{T_j,VS}$$

$$f_2(\mathbf{P}, T_j) = \tilde{\mathbf{P}}_z^2 - r_p^2 + \left(\sqrt{\tilde{\mathbf{P}}_x^2 + \tilde{\mathbf{P}}_y^2} + R_{T_j}\right)^2$$

$$f_3(\mathbf{P}, T_j) = \sqrt{\tilde{\mathbf{P}}_x^2 + \tilde{\mathbf{P}}_y^2} - r_p + R_{T_j}$$

$$f_{\text{quartic}}(\mathbf{P}, T_j) = 4R^2_{T_j}(\tilde{\mathbf{P}}_z^2 - r_p^2) + (r_p^2 + R^2_{T_j} - \tilde{\mathbf{P}}_x^2 - \tilde{\mathbf{P}}_y^2 - \tilde{\mathbf{P}}_z^2)^2$$

$$\mathrm{IsInLemon} = (\mathrm{R_{T_j}} \leq \mathrm{r_p})\ \&\ \mathrm{f_2} \leq 0\ \&\ \mathrm{f_3} \leq 0$$

$$\mathrm{IsTorusCovered} = \mathrm{T_j}\ \text{is free}\ |\ \text{covered by saddle faces}$$

where $\mathbf{C}_{VS}$ and R_{VS} are the center and radius of the visibility sphere. T_j represents the j-th generated torus when the probe touches two atoms at the same time. $\tilde{\mathbf{P}}$ is the location of the point projected on the torus parametrization domain.

4.2.2.3 InsideVS and Inside Concave Probe

InsideVT is short for inside the visibility tetrahedron of the probe, which is composed of the centers of the three touching atoms and the probe as shown in Fig. 4.5. When the Cartesian points is detected as inside the probe sphere with a fixed position, i.e., touching three or more atoms simultaneously, then its **Inside Concave Probe** tag is set to be *true*. We can see that **Inside Concave Probe** helps to exclude the points accessible to the probe in the tetrahedron.

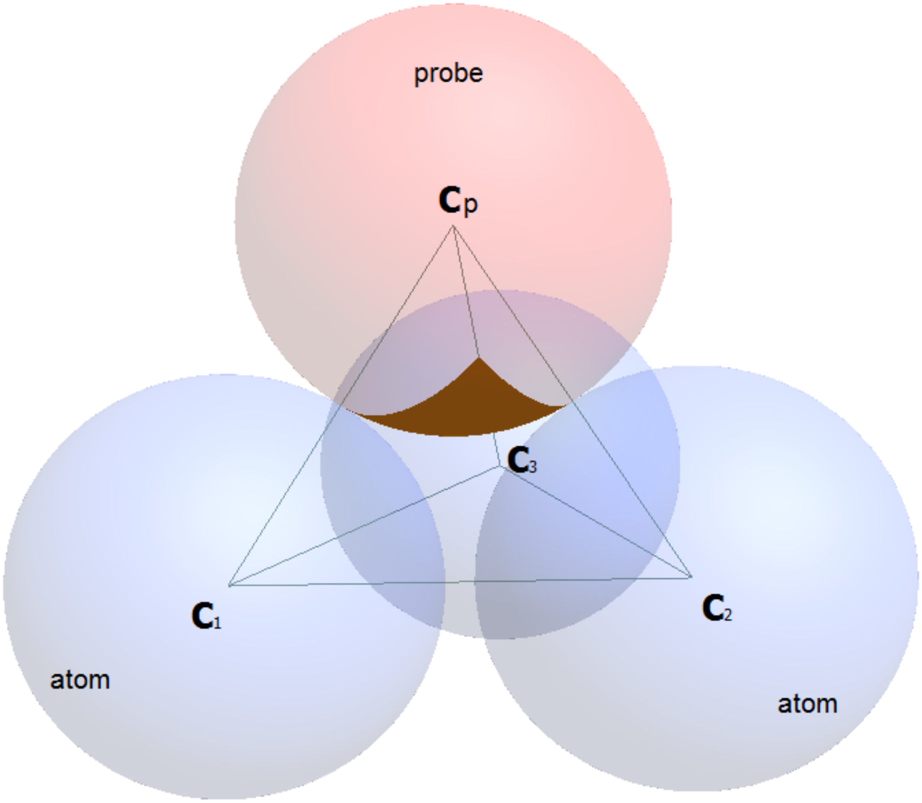


Figure 4.5: Figure for InsideVT. A probe (red) centered at C_p is in contact with three atoms (blue) centered at C_1,C_2 and C_3 respectively. InsideVT is tagged as *true* when the point is detected as inside the tetrahedron $C_pC_1C_2C_3$.

4.2.2.4 Ray Tracing

Since Eq.(4.2.1) only provides a sufficient condition for tagging inside Cartesian grid points, there may exist unlabeled inside Cartesian grid points because of the interior ”tunnel” structure of SES, as depicted in Fig. 4.6. These points can be correctly tagged by counting the number of intersection points with SES. For a Cartesian edge with one Cartesian grid point $\mathbf{P}_1$ labeled and the other point $\mathbf{P}_2$ unlabeled. Point $\mathbf{P}_2$ shares the same Boolean label with $\mathbf{P}_1$ if there are even numbers of points of intersections between SES and the mesh. Otherwise $\mathbf{P}_2$ is tagged with the opposite label of $\mathbf{P}_1$. To compute the correct number of intersection points analytically, a validation for the intersection points is needed:

- Validation for intersections with the *convex patch* (Fig. 4.7)

  1. Not inside of the nearby atoms, by checking the distance to the neighboring atom centers

  2. Not covered by its associated *saddle patches*, by checking the boundary normal directions

- Validation for intersections with the *saddle patch* (Fig. 4.8)

  1. Associated *IsInLemon* is false

  2. Inside of its associated visibility sphere

- Validation for intersections with the *concave patch*

  1. Located in the scope area constrained by the three triangle faces $\mathbf{C}_1\mathbf{C}_2\mathbf{C}_p$, $\mathbf{C}_2\mathbf{C}_3\mathbf{C}_p$, $\mathbf{C}_1\mathbf{C}_p\mathbf{C}_3$ of the tetrahedron $\mathbf{C}_p\mathbf{C}_1\mathbf{C}_2\mathbf{C}_3$, where $\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3$ are the centers of the three touching atoms (Fig. 4.5).

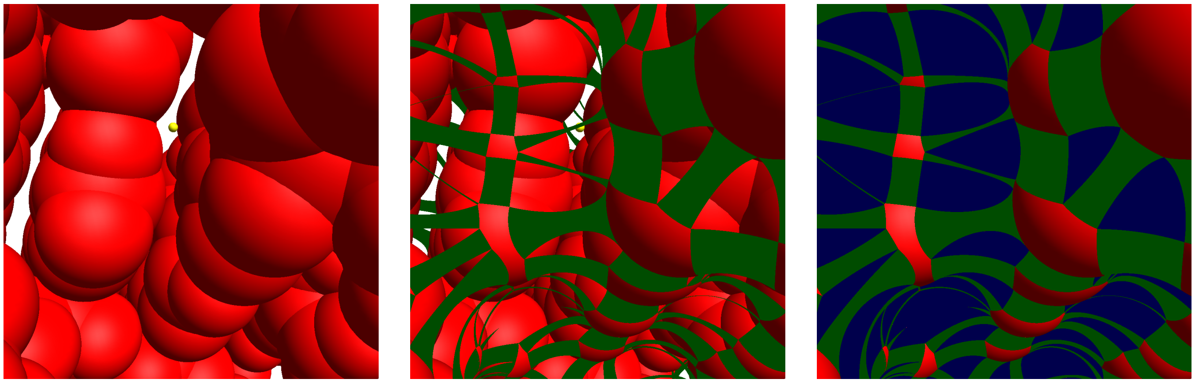


Figure 4.6: Figure for an unlabeled Cartesian grid point. The subfigures from left to right highlight part of the *1A2E* model with different patches respectively, where we can see the Cartesian point(yellow) should be tagged as *inside* the SES. However, it does not satisfy the *sufficient* conditions proposed in Eq. (4.2.1). From left to right are: visualization with atom only, Visualization with atoms and saddle patches only, and Visualization for overall SES, respectively.

2. Outside of the nearby probe spheres with fixed locations

Computing the analytical intersection with convex patches and concave patches is a simple quadratic equation while for saddle patches it is a quartic equation. Here we adopt Jenkins-Traub algorithm [109] to solve the quartic equation for better numerical stability.

4.2.2.5 Computation of the Intersection Coordinates

After labeling the Cartesian grid points, we only need to compute the intersection coordinates and corresponding normals for Cartesian edges with one Cartesian point inside SES and the other outside. Note that the same validation rules discussed above are used for determining the intersection coordinates with nearby patches.

4.2.3 Surface Morphology

In this part, we will present a surface morphology comparison with that generated by the MSMS software [212]. MSMS short for the Micheal Sanner molecular surface, is based on the

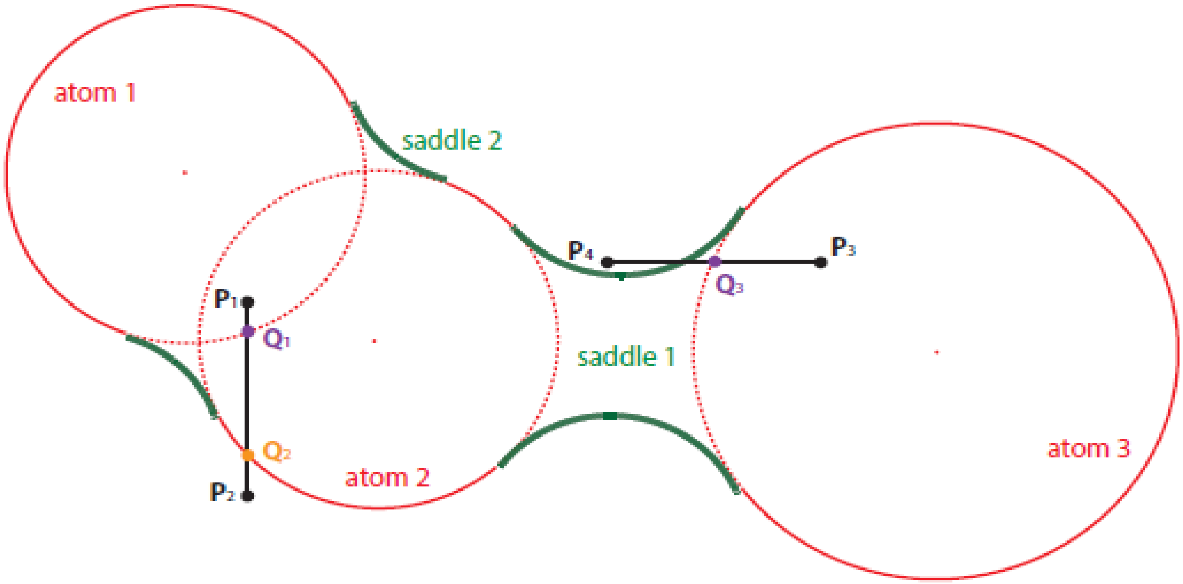


Figure 4.7: Validation for intersection with convex patches. Solid lines illustrate the outline of a simple SES composed of three atoms (red) and two saddle patches (green). P_1P_2 and P_3P_4 are the Cartesian edges. Note here we only mark the possible intersections with convex patches. Edge P_1P_2 seems to have two intersections Q_1 and Q_2 with convex patches, where Q_1 is actually an invalid intersection since it is inside of atom 2. Similarly, the intersection point Q_3 for edge P_3P_4 is not valid, since it is included by saddle patch 2.

so called reduced surface. It is the first program that can handle the geometric singularities. The MSMS algorithm consists the following four major steps for the SES generation and triangulation:

- First, compute the reduced surface a molecule.
- Second, build an analytical representation of the SES from the reduced surface.
- Third, remove all the self-intersecting parts from the analytic SES built above.
- Fourth, produce a triangulation of the reduced surface based on the SES.

Due to the efficiency and robustness of the MSMS software, it has been incorporated into several molecular modeling software suites, e.g., the Visual Molecular Dynamics software (VMD) [118], the UCSF Chimera [196] et al.

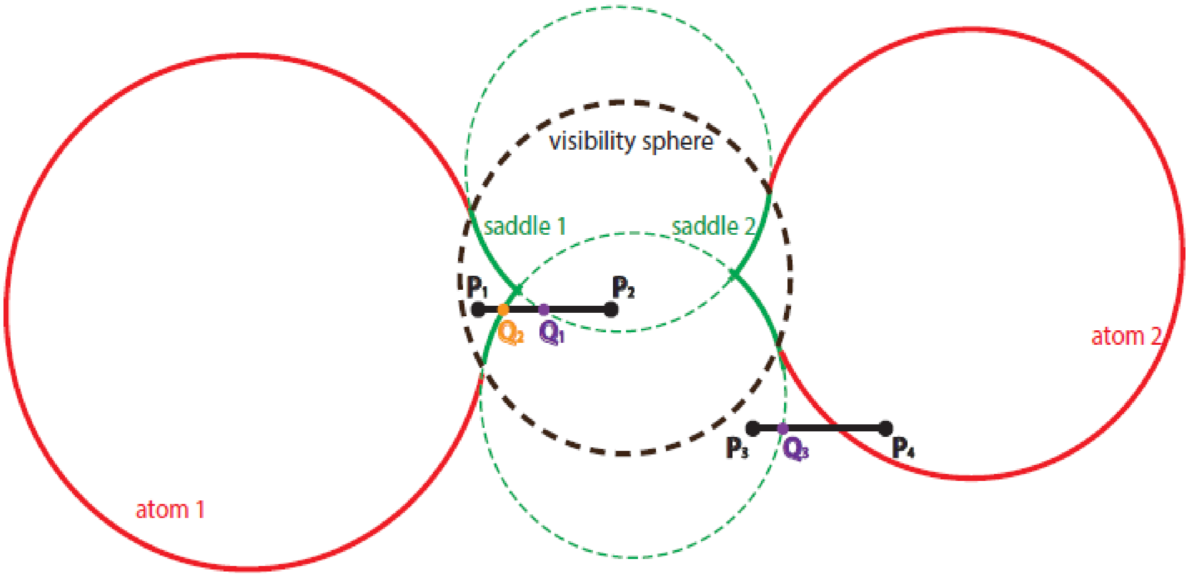


Figure 4.8: Validation for intersection with saddle patches. Solid lines illustrate the outline of a simple SES composed of two atoms (red) and two saddle patches (green), where two separate saddle patches belong to the same torus. The black dashed circle is the outline of the corresponding visibility sphere. Singularities are generated in this case and they are appropriately handled implicitly with our designed conditions. P_1P_2 and P_3P_4 are the Cartesian edges. Note here we only mark the possible intersections with saddle patches. There are two intersections Q_1 and Q_2 for edge P_1P_2 with saddle patches, where Q_1 is invalid since its associated Boolean tag *IsInLemon* is *true*. The intersection point Q_3 for edge P_3P_4 with saddle patches is not valid either, since it is outside of the corresponding visibility sphere.

The MSMS surface generation depends on an additional parameter besides the intrinsic parameters for describing the SES, namely, the surface density, which is the approximated number of triangles on per unit Å^2 area. Usually, the larger the surface density is, the closer of MSMS to the analytical SES, provided the surface can be generated successfully.

Despite the great success of of the MSMS, it should be point out some problems that we encountered during the usage of MSMS. On the one hand, the MSMS surface may change the atomic radius of the molecule automatically. On the other hand, MSMS may fail to generate SES at high density. As depicted in Fig. 4.9, MSMS generates the SES for the molecular (PDB ID: 1dqz) successfully at density 5, 10, and 25, while fails at density 50.

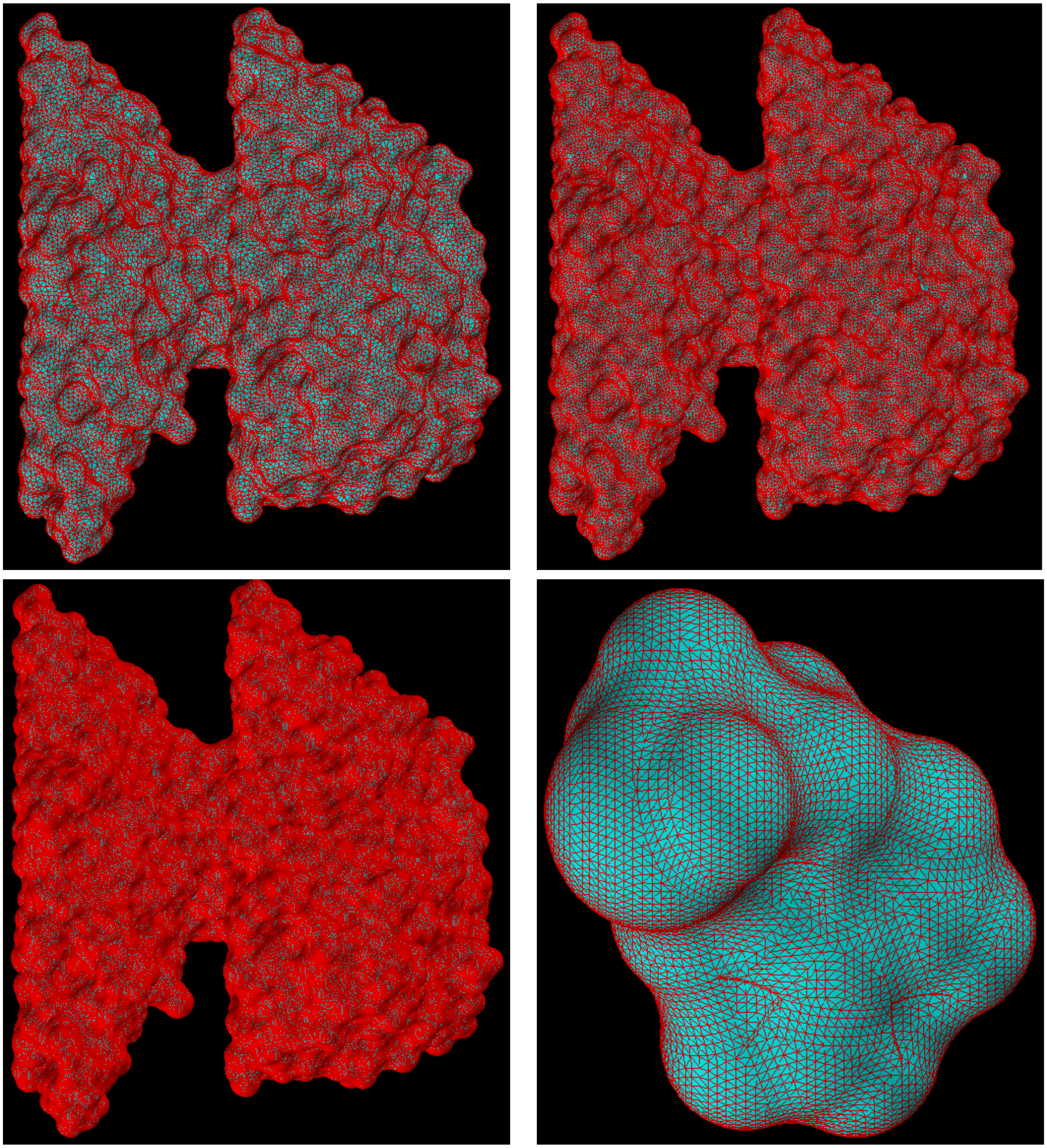


Figure 4.9: The MSMS surfaces for the biomolecule (PDB ID: 1dqz). Charts from top left to bottom right are the MSMS surfaces with the densities 5, 10, 25 and 50, respectively. The MSMS fails to generate correct surface at density 50, while it works well other densities.

For a comparison, we depicts the SES for molecular 1dqz generated by ESES software in Fig. 4.10. It is easy to see that, the surfaces generated by MSMS at density 5, 10, 25 are consistent with that by ESES. With the increasing of the surface density, MSMS converges to the ESES.

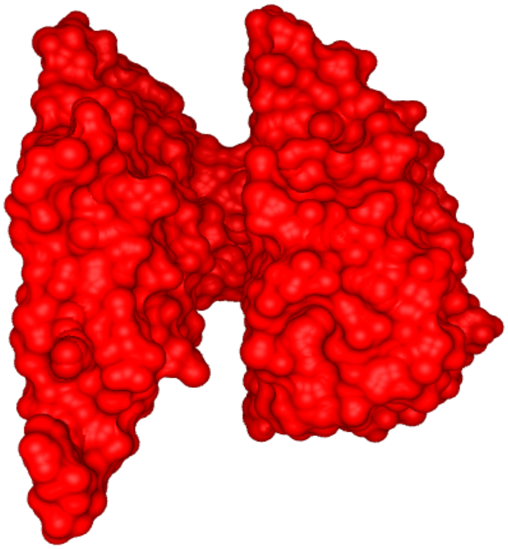


Figure 4.10: The ESES surface for protein 1DQZ.

4.3 ESES: Area, Volume Calculation

The molecular surface area and enclosed volume are widely used in modeling the nonpolar solvation free energy in biophysics, the accurate calculation of the area and volume is of fundamental importance in the solvation and binding analysis [56]. The ESES embeds SES in the Cartesian mesh, which is represented by two sets of grid points J_1 and J_2, represent the grids inside and outside the SES Γ, respectively. Together with the intersection of the Cartesian mesh with the SES and the associated outer normal direction. We denote the irregular grid points set as J_{Irr} where a grid point is said to be irregular if at least one of its six neighbor grids located in the different side of the SES.

The area of the SES Γ can be approximated by the following formula:

$$\text{Area} = \int_{\Gamma} dS \approx \sum_{(i,j,k)\in I} \left(|n_x| + |n_y| + |n_z|\right) h^2, \tag{4.3.1}$$

where h is the grid size, I is the set of irregular grid points that is either inside or on the surface Γ. $|n_x|$ is the magnitude of the first component of the outer normal direction at the (x_0, y_j, z_k), which is the intersection of Γ with the x-mesh line that passes through (x_i, y_j, z_k), $|n_y|$ and $|n_z|$ are defined analogously.

The volume enclosed by the molecular surface Γ is similarly evaluated by:

$$\text{Volume} = \int_{\Omega^m} d\mathbf{r} \approx \frac{1}{2}\left(\sum_{(i,j,k)\in J_1} + \sum_{(i,j,k)\in J_1 \bigcup J_{\text{Irrin}}}\right) h^3, \tag{4.3.2}$$

where J_{Irrin} is the set of irregular grid points that inside the surface Γ.

Smereka originally proposed the similar scheme for evaluating the surface integration and did the convergence analysis in the work [223]. Similar scheme is also used for calculating the dielectric boundary force in the PB based implicit solvent molecular dynamics simulation by Geng et al [89].

Table 4.1 gives the grid refinement analysis of the numerical scheme given by Eqs. (4.3.1) and (4.3.2) on the sphere with radius 2, obviously the numerical scheme for area calculations is roughly of second order convergence, while that for the calculation of surface enclosed volume is approximated of third order convergence.

To test the accuracy of the numerical scheme for surface area and volume calculation to the real biomolecules, we perform the calculation on the Amber PBSA test set, which

Table 4.1: The grid refinement analysis of the area and volume calculation for the sphere with radius 2.

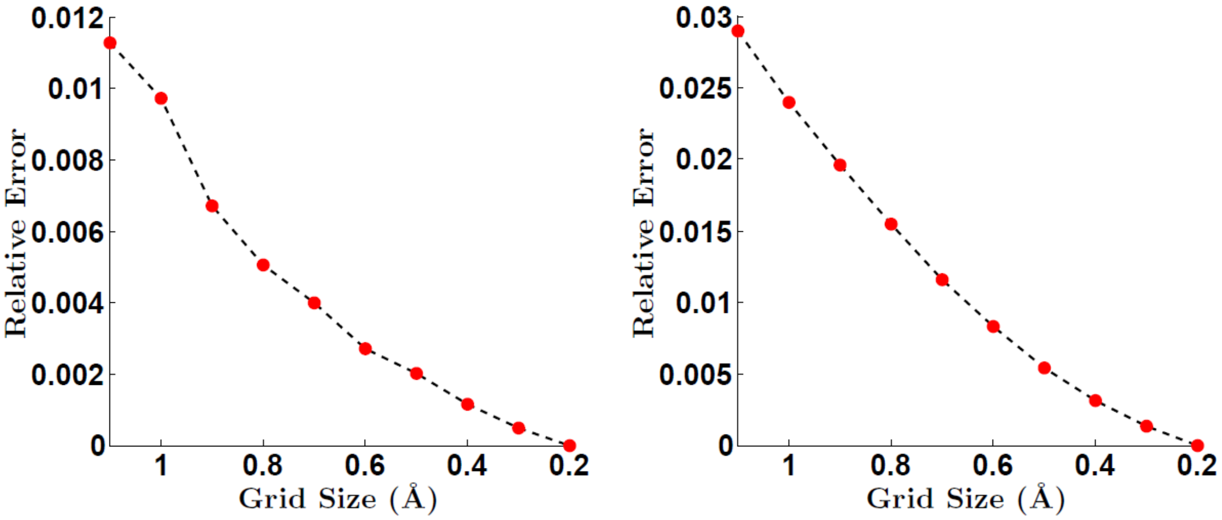
Grid size	Area	Error	Order	Volume	Error	Order
1.0	43.755	6.51		47.00	13.490	
0.5	48.824	1.44	2.18	35.28	1.865	2.85
0.25	49.839	0.43	1.76	33.80	0.287	2.70
0.125	50.193	0.07	2.56	33.57	0.058	2.30
0.0625	50.243	0.02	1.71	33.52	0.007	3.05
Exact Value	50.28			33.52		

contains two parts, the protein and nucleic acid molecules. There are 937 molecules in total in this test set, the number of atoms of which range from several hundreds to approximate ten thousands. The data set is downloaded from the web http://rayl0.bio.uci.edu/rayl/. Figure. 4.11 depicts the relative convergence of the numerical scheme for the surface area and volume calculation, where the relative error for the area calculation is defined as:

$$\text{RelativeError}_h := \frac{|\text{Area}_h - \text{Area}_{0.2}|}{\text{Area}_{0.2}},$$

where RelativeError_h stands for the relative error at the grid size h, Area_h and $\text{Area}_{0.2}$ are the numerical surface area at grid size h and 0.2 Å, respectively. The relative error for the volume calculation is defined in the same way.

To further verify the accuracy of the area and volume calculation in the ESES software. We compare that by both ESES and MSMS software, to ensure the accuracy of the MSMS calculation, we use the density 100 for MSMS surface generation. Due to the robustness problem, only 740 molecular surfaces are generated successfully at density 100, the failure of others are either due to the failure of surface generation or the change of radius in the surface generation. The details about this comparison can be found in [157]. The results depicted in Fig. 4.12 demonstrate the excellent consistency between the area and volume



(a)

(b)

Figure 4.11: Convergence test of the surface area and enclosed volume for the PBSA test set compared to the results obtained at grid size 0.2 Å. (a) Area; (b) Volume.

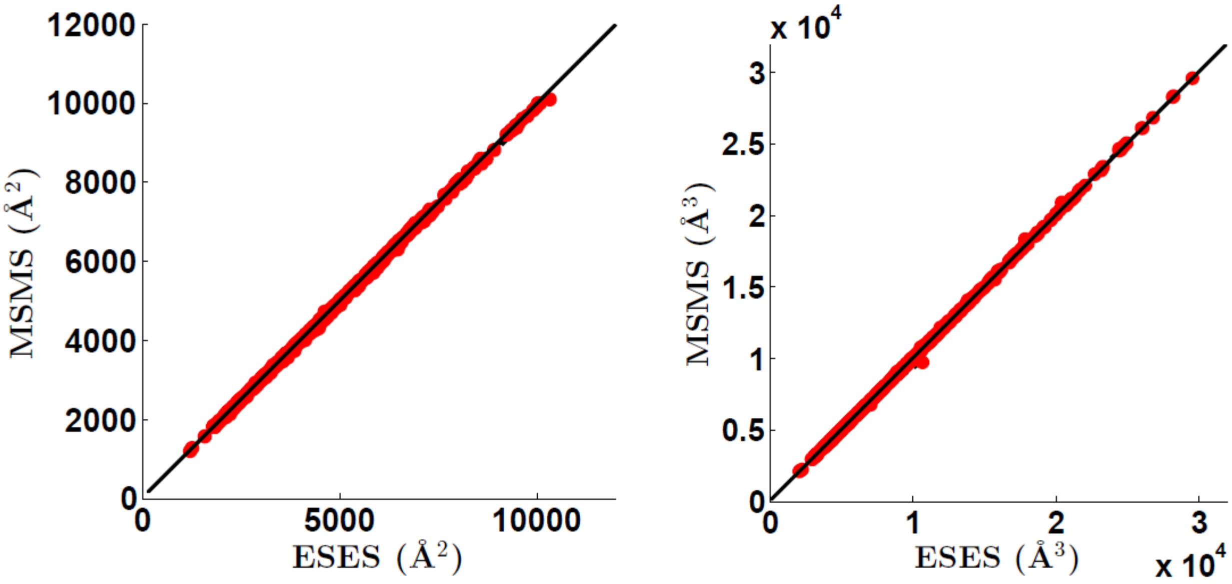
calculation by two software.

4.3.1 Atomic Area

In the implicit solvent modeling, the whole surface area usually employed for a crude model of the area contribution in nonpolar solvation free energy, where the nonpolar energy is modeled by $G_1^{\mathrm{NP}} = \gamma \mathrm{Area}$, with γ being the surface tension. A more accurate area model for modeling the nonpolar solvation effects is based on the atomic surface areas, in which different type of atoms admit different surface tensions. As such, one can model the nonpolar solvation free energy by:

$$G_2^{\mathrm{NP}} = \sum_j \gamma_j \mathrm{Area}_j, \tag{4.3.3}$$

where γ_j and Area_j are atomic surface tension and surface area of j-th type of atoms, respectively. Therefore, in addition to the surface area calculation, the ESES software also provides the atomic surface area calculation. In the ESES formulation, each piece of patch is



(a) (b)

Figure 4.12: The consistence of the area and volume calculation between the MSMS and ESES software packages for part of the PBSA test set that MSMS can generate the surface successful without change of atomic radius at the density 100 (here 740 biomolecules used) for volume calculation and analytic surface area by MSMS software. The ESES results are generated at grid size 0.2 Å. (a) For the area consistence, the correlation is 0.9999, the best fitting line is $y = 0.9934x+14.3307$; (b) For the volume consistence, the correlation is 0.9999, the best fitting line is $y = 1.0040x - 23.9577$.

represented by some intersecting points with the Cartesian mesh line, due to the definition of the SES, the contact surface is contributed from one atom; the toric is generated due to the contacting of probe with two atoms simultaneously; and the concave patch is associated with three atoms. This fact is inherited to the surface area partitioning. Note that in our area formula Eq. (4.3.1), the whole area is the cumulative contribution from each intersection. In the atomic area partitioning, we distribute the area comes from each intersecting point to the associated atom. For a given intersection point with coordinate $\mathbf{r}$ and the associated area is A_0, the partitioning of the area is based on the following criteria:

- If the intersection is associated with only one atom, the whole area A_0 is assigned to the corresponding atom.

- If the intersection is associated with two atoms indexed by i and j, whose centers are $\mathbf{r}_i$ and $\mathbf{r}_j$, the radius are r_i and r_j, respectively. Then we calculated the weighted distance from the intersection $\mathbf{r}$ to two centers, they are given by:

  $$d_i = ||\mathbf{r} - \mathbf{r}_i|| - r_i,$$

  and

  $$d_j = ||\mathbf{r} - \mathbf{r}_j|| - r_j,$$

  respectively, where $|| * ||$ denotes for the Euclidean distance between two points. And the whole area is partitioned to the closer atom under the weighted distance measure.

- If the intersection is associated with three atoms, we distribute the area based on the same manner as that in two atoms case, and the whole area is attributed to the closest atom.

The above partitioning is based on the basic idea of the weighted Voronoi diagram [5].

4.4 ESES: Topological Analysis

Loops and cavities are omnipresent in SESs. Usually these loops or cavities are related to the binding pockets or binding sites. Accurate and efficient algorithms for detecting the loops and cavities together with measuring the size of these topological features play an important role in the practical applications, including computer aided drug design. In this section, we provide an accurate and efficient numerical algorithm based on the homology theory [133, 73] for the loop and cavity detection. Furthermore, we propose a level set method based filtration to generate persistent bar codes which characterize the size of loops and cavities.

Homology group theory, provides an effective theoretical framework for computing the loops and cavities on the manifold. The homology group constructed based on the cubical complex provides a practical methodology for computing the loops and cavities for the SES embedded in the Euclidean space, i.e., ESES, for detail theory about the homology on cubical complex setting, readers are refer to [133]. Furthermore, the persistent homology theory [72] provides a measure to measure the size of loops and cavities on the manifold.

4.4.1 Loops and Cavities Detections

In this part, a brief introduction of the homology theory in cubical complex setting which gives the general framework for computing the loops and cavities in the Eulerian representation of the manifold will be provided, for detail the readers are refer to [133]. In the cubical complex setting, the homology theory is built up from geometric and algebraic building block. In the following, we will first briefly review these concepts and then turn to discuss the application of these tools for loops and cavities detection on ESES.

4.4.1.1 Geometric Building Block

In the Eulerian representation, the cubes are the basic geometric building blocks of the homology theory, first we need the following basic concepts of the cubes:

- An elementary non-degenerate interval is a closed interval $I \subset \mathbb{R}$ of the form $I = [m, m+1]$ (or $I = [m]$ for simplicity) for some integer m. An elementary degenerate interval is a point $I = [m, m]$.
- An elementary cube Q or d cube is a d-product if elementary intervals, i.e.,

  $$Q = I_1 \times I_2 \times \cdots \times I_d \subset \mathbb{R}^d,$$

  where each I_i, $i = 1, 2, \cdots d$ is an elementary interval of non-degenerated or degenerated type, and d is called the embedding number of Q, denoted as $embQ = d$. The dimension of Q, denoted as $dimQ$, is defined to be the number of non-degenerated components in Q, and $\mathcal{K}_k$ denotes the set of all k dimensional elementary cubes. Let $\mathcal{K} \doteq \bigcup_{d=1}^{\infty} \mathcal{K}^d$ be the set of all elementary cubes, and $\mathcal{K}^d$ be the set of all elementary cubes in $\mathbb{R}^d$.
- The set of k-dim cubes with embedding number d is $\mathcal{K}_k^d \doteq \mathcal{K}_k \bigcap \mathcal{K}^d$. Obviously, if $Q \in \mathcal{K}_k^d$ and $P \in \mathcal{K}_{k'}^{d'}$, then $Q \times P \in \mathcal{K}_{k+k'}^{d+d'}$.

A set $X \subset \mathbb{R}^d$ is said to be cubical provided it can be written as a finite union of elementary cubes. For a given cubical set $X \subset \mathbb{R}^d$, we define the cubical set $\mathcal{K}(X)$ and

k-cube set $\mathcal{K}_k(X)$ by:

$$\mathcal{K}(X) \doteq \{Q \in \mathcal{K} | Q \subset X\},$$

$$\mathcal{K}_k(X) \doteq \{Q \in \mathcal{K}(X) | dimQ = k\},$$

the elements of $\mathcal{K}_k(X)$ are called the k-cubes of X.

4.4.1.2 Algebraic Operations

To study the topological properties of the molecular surface in the Eulerian representation, the basic operations on the aforementioned geometric building blocks will be presented in this part. Each elementary k-cube $Q \in \mathcal{K}_k^d$ is associated with an algebraic object $\hat{Q}$ which is called the elementary k-chain of $\mathbb{R}^d$, the set of all elementary k-chains of $\mathbb{R}^d$ is $\hat{\mathbb{K}}_k^d \doteq \{\hat{Q} | Q \in \mathcal{K}_k^d\}$, and the set of all elementary chains of $\mathbb{R}^d$ is: $\hat{\mathcal{K}}^d \doteq \bigcup_{k=0}^{\infty} \hat{\mathcal{K}}_k^d$.

The first algebraic operation to be defined on the cubical complex is the addition operation. First, we define the k-chains as the linear combination of k-chain:

$$c = a_1\hat{Q}_1 + a_2\hat{Q}_2 + \cdots + \hat{a_m}\hat{Q_m},\ a_i \in \mathbb{Z}, i = 1, 2, \cdots, m,$$

the set of all the above k-chains is denotes by C_k^d.

The addition of two k-chains is defined by:

$$\sum a_i\hat{Q}_i + \sum b_i\hat{Q}_i = \sum (a_i + b_i)\hat{Q}_i.$$

It is obviously that under the addition operation, C_k^d is an Abelian group.

Before defining the homology group on the cubical complex, we need to define the boundary operations on the cubical complex. To define the boundary operator, we need the following defined scalar and cubical products.

Definition 4.4.1. ***Scalar product:*** *Let $c_1,\ c_2 \in C_k^d$, where $c_1 = \sum_{i=1}^m a_i \hat{Q}_i$ and $c_2 = \sum_{i=1}^m b_i \hat{Q}_i$. The scalar product of chains c_1 and c_2 is defined as:*

$$< c_1, c_2 > \doteq \sum_{i=1}^{m} a_i b_i.$$

Definition 4.4.2. ***Cubical product:*** *For $\forall$ elementary cubes $P \in \mathcal{K}_k^d$ and $Q \in \mathcal{K}_{k'}^{d'}$, the cubical product between P and Q is defined to be:*

$$\hat{P} * \hat{Q} \doteq \widehat{P \times Q}.$$

Furthermore, for $\forall$ $c_1 \in C_k^d$ and $c_2 \in C_{k'}^{d'}$, the cubical product is:

$$c_1 * c_2 = \sum_{P \in \mathcal{K}_k, Q \in \mathcal{K}_{k'}} < c_1, \hat{P} >< c_2, \hat{Q} > \widehat{P \times Q},$$

and $c_1 * c_2 \in C_{k+k'}^{d+d'}$.

For the cubical product, the following factorization property holds:

Lemma 4.4.3. *For $\forall \hat{Q} \in \hat{\mathcal{K}}^d$ with $d > 1$, there exists unique elementary cubical chains $\hat{I}$ and $\hat{P}$ with $embI = 1$ and $embP = d - 1$, s.t., $\hat{Q} = \hat{I} * \hat{P}$.*

With the previous preparation, the boundary operation can be defined in the following inductive manner.

Definition 4.4.4. ***Boundary operator:*** *For $k \in \mathbb{Z}$, the cubical boundary operator:*

$$\partial_k : C_k^n \to C_{k-1}^n,$$

is a homomorphism of Abelian groups, defined for an elementary chain $\hat{Q} \in \hat{\mathcal{K}}_k^n$ by induction on the embedding number n as follows:

- *For $n = 1$, Q is an elementary interval, i.e., $Q = [m]$ or $Q = [m, m+1]$ for some $m \in \mathbb{Z}$, and one defines:*

$$\partial_k \hat{Q} = \begin{cases} 0, & \textit{if } Q = [m], \\ \widehat{[m+1]} - \hat{[m]}, & \textit{if } Q = [m, m+1]. \end{cases}$$

- *For $n > 1$, let $I_1(Q)$ and $P = I_2(Q) \times \cdots \times I_n(Q)$ so that $\hat{Q} = \hat{I} * \hat{P}$, then one defines:*

$$\partial_k \hat{Q} = \partial_{dim} \hat{I} * \hat{P} + (-1)^{dim} \hat{I} * \partial_{dim} \hat{P}.$$

By linearity, the boundary operator can be extended to chains, i.e., if $c = \sum_{i=1}^{p} a_i \hat{Q}_i$, then $\partial_k c = \sum_{i=1}^{p} a_i \partial_k \hat{Q}_i$.

It is easy to show that the boundary operator satisfies $\partial_k \circ \partial_{k-1} = 0, \forall k > 1$ in the cubical complex setting.

Now for a given manifold X embedded in the Euclidean space $\mathbb{R}^d$ and X is represented as a cubical set, let $\hat{K}_k(X) \doteq \{\hat{Q} | Q \in \mathcal{K}_k(X)\}$ and let $C_k(X)$ be the subgroup of C_k^d generated by the elements of $\hat{K}_k(X)$, which is called the set of k-chains of X. The boundary operator maps $C_k(X)$ to a subset of $C_{k-1}(X)$, thus one can restrict the boundary operator to the cubical set X. The boundary operator for the cubical set X: $\partial_k^X : C_k(X) \to C_{k-1}(X)$ can be defined by restricting $\partial_k : C_k^d \to C_{k-1}^d$ to $C_k(X)$, The cubical chain complex for the cubical set $X \subset \mathbb{R}^d$ is defined as $C(X) \doteq \{C_k(X), \partial_k^X\}_{k \in \mathbb{Z}}$, where $C_k(X)$ are the groups of cubical k-chains generated by $\mathcal{K}_k(X)$ and ∂_k^X is the cubical boundary operator restricted to X.

For a given cubical set X, the corresponding k-chains group $C_k(X)$ is defined, it is straightforward to introduce two subgroups of $C_k(X)$:

- k-cycle group $Z_k(X) \doteq C_k(X) \bigcap ker \partial_k \subset C_k(X)$.
- k-boundary group $B_k(X) \doteq im \partial_{k+1}^X = \partial_{k+1}(C_{k+1}(X)) \subset C_k(X)$.

$\partial_k \circ \partial_{k-1} = 0, \forall k > 1$ implies that $B_k(X) \subset Z_k(X)$, therefore, we can define the following homology group.

Definition 4.4.5. ***Homology group:*** *The k-th homology group of the cubical set X is defined as the quotient group:*

$$H_k(X) \doteq Z_k(X)/B_k(X).$$

The k-th Betti number is defined as the rank of the k-th homology group, $\beta_k = rank H_k$.

Remark 4.4.6. *$H_k(X)$ describes k-dimensional holes of X, e.g., $H_0(X)$ measures connected components, $H_1(X)$ measures loops, and $H_2(X)$ measures voids. In other words. β_0 is the number of connected components, β_1 is the number of loops, β_2 is the number of voids, and so on.*

4.4.1.3 Homology on ESES

In this part, we will discuss the computation of the homology on the manifold enclosed by ESES. Note that during the ESES generation, all the grid points are classified as inside or outside the SES, the inside grids can be used to construct the cubical complex that represents the manifold enclosed by SES.

In this work the cubical homology computation will be carried out by the Perseus persistent homology software [181], which is an efficient persistent homology program that can handle both the simplicial and cubical complex represented manifolds [172]. The SES generated by the present ESES software provides suitable input data for the Perseus software.

We consider a benchmark test example, the C_{60} molecule, to illustrate the ESES performance in loop and cavity generation [251]. Figure 4.13 depicts C_{60} generated at the atomic radius of 0.8 Å for all the Carbon atoms with the probe radius of 0.1 Å. It has 32 rings. Figure 4.14 provides the number of loops generated by ESES at different grid sizes. When the grid size used for the surface generation is coarser than 0.6 Å, the numerical method cannot capture all the small loops, which are about 0.5 Å in diameter, in the C_{60} molecule. However, when the grid size is finer than 0.6 Å, all the loops can be resolved. Note that homology computation reports only 31 loops because one of the 32 loop can be expressed as a linear combination of all other loops.

Figure 4.15 shows the solvent excluded surfaces generated with probe radius 1.4 Å for some molecules from the PBSA test set. Their corresponding numbers of loops and cavities are also presented and calculated at the grid resolution of 0.3Å.

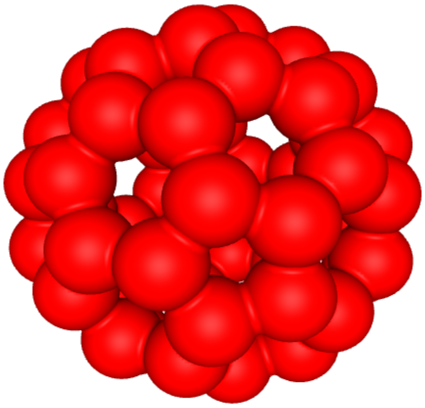


Figure 4.13: The SES for the C_{60} molecule with probe radius 0.1 Å, and the van der Waals radius of the Carbon atom is set to be 0.8 Å.

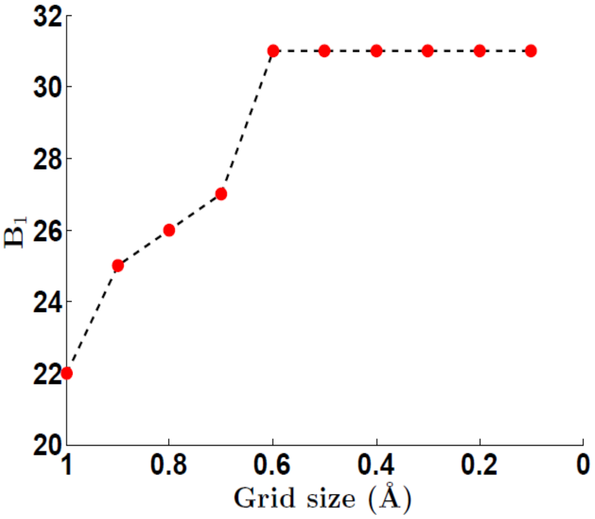


Figure 4.14: The number of loops calculated at different grid sizes for the above C_{60} solvent excluded surface.

4.4.2 Persistence

The method for detecting the loops and cavities of the manifold formed by the molecular surface has been provided in the previous part. Nevertheless, ability to detecting the number of loops and cavities usually not very useful in practice. The size of the loops and cavities also needs an effective way to characterize.

Persistent homology theory provides a way to measure the size of the loops and cavities.

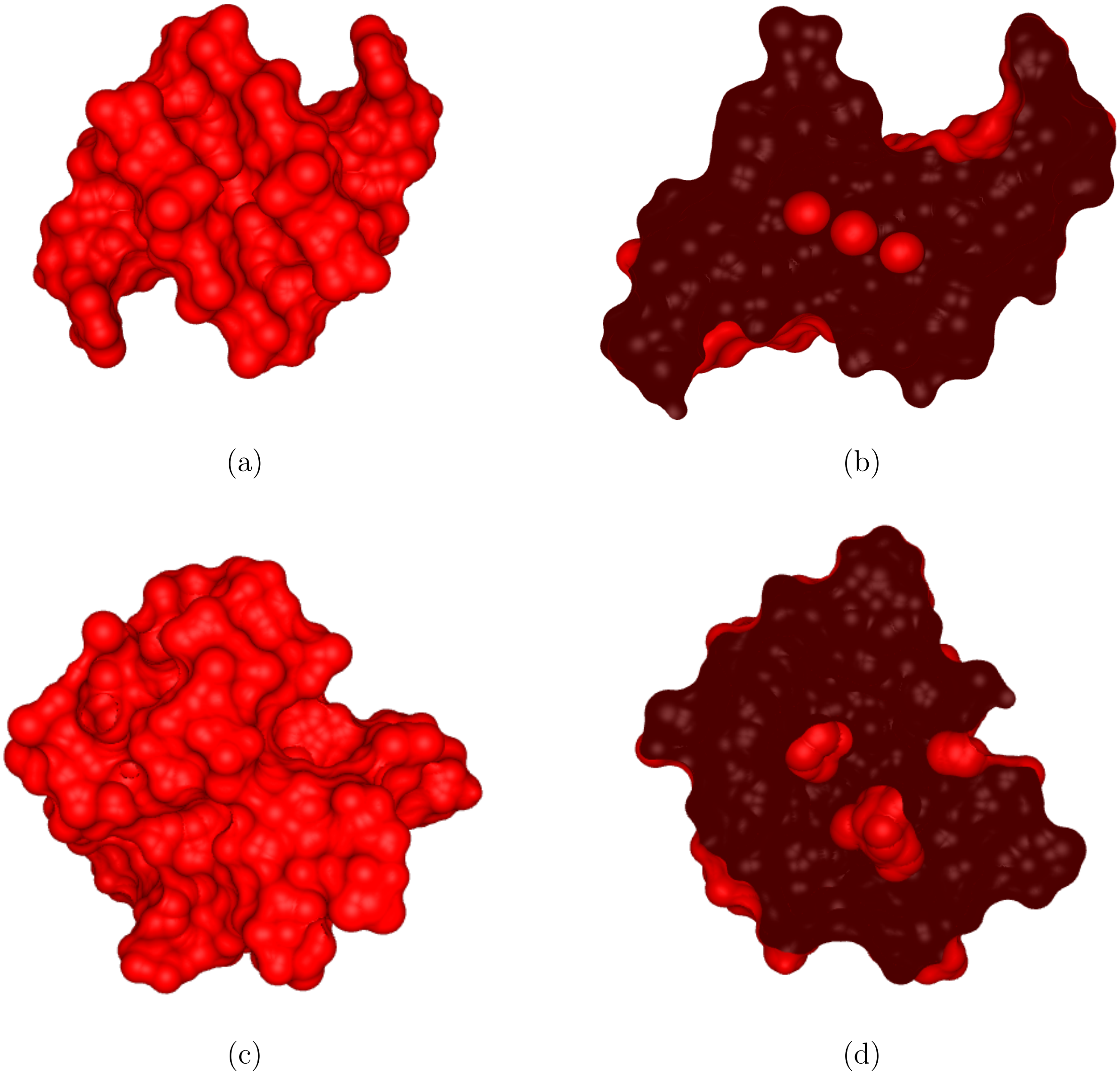


Figure 4.15: The solvent excluded surfaces and their topology of two biomolecules. (a) The ESES result for protein 2AVH with 3 loops and 1 cavity. (b) A cross section of protein 2AVH showing the loops. (c) The ESES result for protein 1af8 with 2 loops and 2 cavities. (d) A cross section of protein 1af8 showing the loops and cavities.

The measure is called persistence in the filtration. We have introduced a time propagation approach for generating the persistent homology filtration in our recent work [251]. In the present work, we replace the geometric flow propagation by constant velocity propagation so as to uniformly measure the sizes of different loops or cavities.

To define the persistent homology, we need a filtration, i.e., a complex K together with nested sequence of sub-complexes $\{K^i\}_{0\leq i\leq n}$, such that:

$$\emptyset = K^0 \subset K^1 \subset \cdots \subset K^n = K,$$

each sub-complex K^i in the filtration has an associated chain group C_k^i, cycle group Z_k^i and boundary group B_k^i, for $\forall i \geq 1$, and thus one has the following definition.

Definition 4.4.7. *The p-persistent of kth homology group K^i is:*

$$H_k^{i,p} = Z_k^i / \left(B_k^{i+p} \bigcap Z_k^i \right),$$

here $H_k^{i,p}$ captures the topological features of the filtrated complex that persists for at least p steps in the filtration.

To measure the size of loops and cavities of the manifold formed by the SES, a slightly modification is needed in the previous defined filtration, the modified filtrated sequence of complexes is defined as:

$$\Omega^{\mathrm{m}} = K_0 \subset K_1 \subset \cdots \subset K_n = K,$$

where K_0, or Ω^{m}, is the manifold formed by the SES. K is the manifold that the loops and cavities of K_0 are filled.

The remaining thing for building the theoretical framework of persistence measuring is filtration construction. The level set method provides a general theoretical framework and and effective numerical method for the surface propagation with constant velocity.

For the molecular surface Γ, we can give it a level set representation $\psi(\mathbf{r})$, such that:

$$\psi(\mathbf{r}) \begin{cases} > 0, & \mathbf{r} \in \Omega^{\mathrm{s}} \text{ outside the molecular surface,} \\ = 0, & \mathbf{r} \in \Gamma \quad \text{on molcular surface,} \\ < 0, & \mathbf{r} \in \Omega^{\mathrm{m}} \text{ inside the molecular surface.} \end{cases} \tag{4.4.1}$$

$\psi(\mathbf{r})$ can be chosen as the signed distant function to the molecular surface Γ.

To propagate the surface with a given velocity along the outer normal direction of the SES in the Eulerian representation, we introduce the time variable t to the level set function, i.e., let $\psi(\mathbf{r}) = \psi(\mathbf{r}, t)$. By taking the derivative with respect to t of the SES level set function $\psi(\mathbf{r}, t) = 0$, one has

$$\frac{\partial \psi}{\partial t} + \nabla\psi \cdot \frac{\partial \mathbf{r}}{\partial t} = 0.$$

Note that $\frac{\partial \mathbf{r}}{\partial t}$ is exactly the surface propagation velocity, denoted as $\vec{v}$. Projecting the velocity on to the outer normal direction of the surface $\frac{\nabla\psi}{|\nabla\psi|}$, one has the projected velocity along the normal direction $v_N \doteq \vec{v} \cdot \frac{\nabla\psi}{|\nabla\psi|}$.

Finally, the level set equation for describing the surface propagation along the outer normal direction can be written as:

$$\frac{\partial \psi}{\partial t} + v_N |\nabla\psi| = 0, \tag{4.4.2}$$

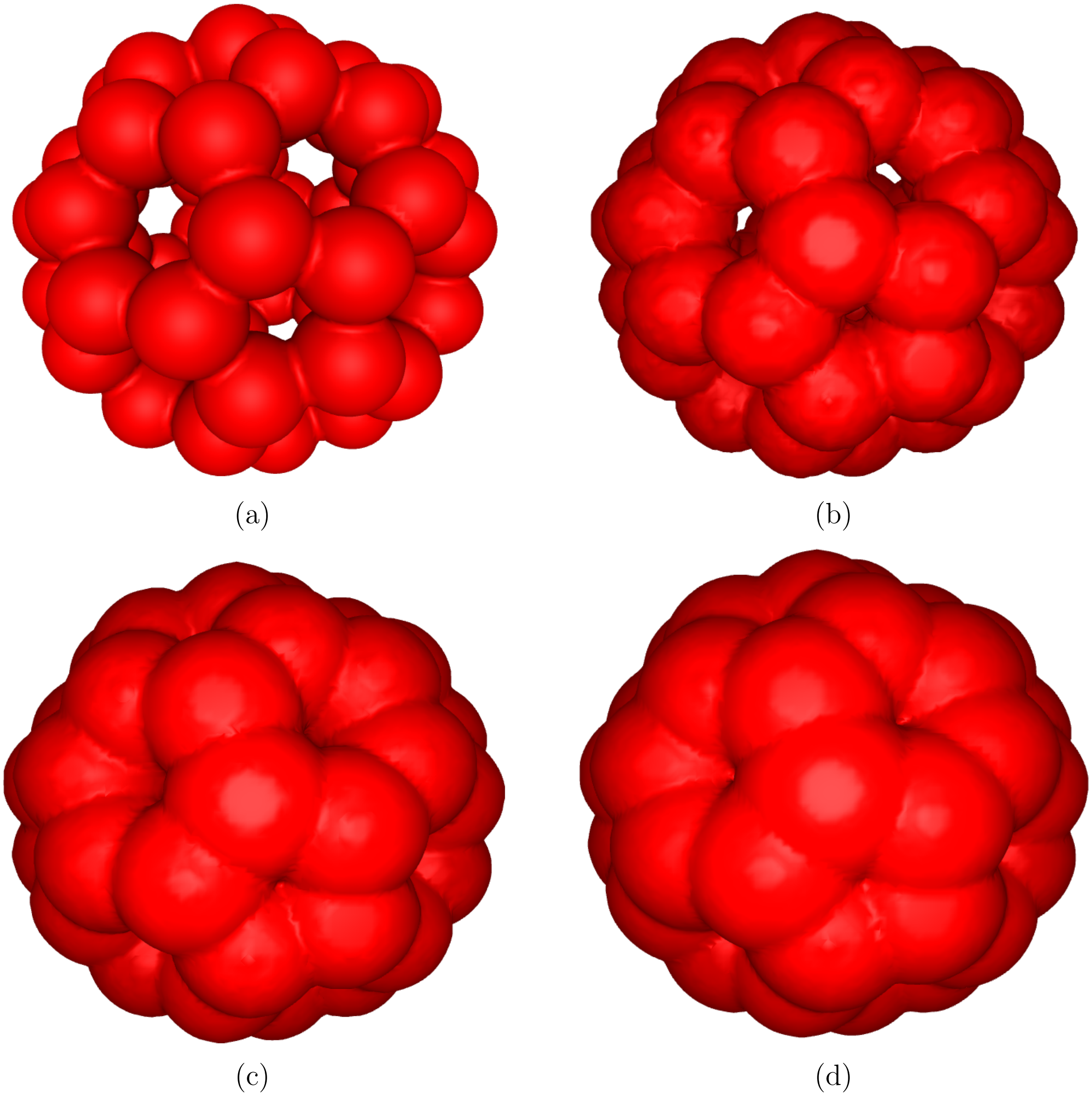
which is a Hamilton-Jacobi equation. The level set Eq. (4.4.2) is solved by a simple upwind scheme with periodic boundary condition. For detail description and numerical implementation of the level set method, readers are referred to [194, 217].

First, we consider the benchmark test example, the SES of C_{60} generated by the same parameters as mentioned before. The loops on the SES can be classified into two categories, the pentagon and hexagon loops, these two types of loops are of different sizes. In the following, we solve the level set equation at grid resolution 0.1 Å in spatial domain discretization with the velocity 0.1 Å per unit time along the outer normal direction. To ensure the CFL condition, the grid resolution 0.002 is employed in the time discretization. Figure 4.16 depicts some frames of the evolution procedure of the C_{60} SES under the driven of the level set equation, different frames show that the SES is propagate along the outer normal direction with a constant velocity, and after some time, both type of loops will be closed. The persistent time of these two types of loops reflects the size of the loops, which actually should be proportional to the size of loops since the surface grows with a constant velocity.

After the above validation on the C_{60} molecular surface, we apply the above level set based persistent homology theory to biomolecules to investigate the size of loops and cavities of the corresponding manifold enclosed by the SESs. All the implementations are carried out at the grid resolution of 0.3 Å in spatial discretization and 0.01 at the temporal discretization which guarantees the stability of the numerical integrator. The surface is propagated with a constant velocity 0.2 Å per unit time along the outer normal direction.

First we study the persistence of the loops and cavities of the protein 1clh, the previous homology theory predicts that the SES generated with probe radius 1.4 Å with the Amber force field has 5 loops and 5 cavities.

Figure 4.17 shows the frames of the growing surfaces at time 0, 25, 50, and 100, respectively. Intuitively, during the surface propagation, the molecule should become fatter, the



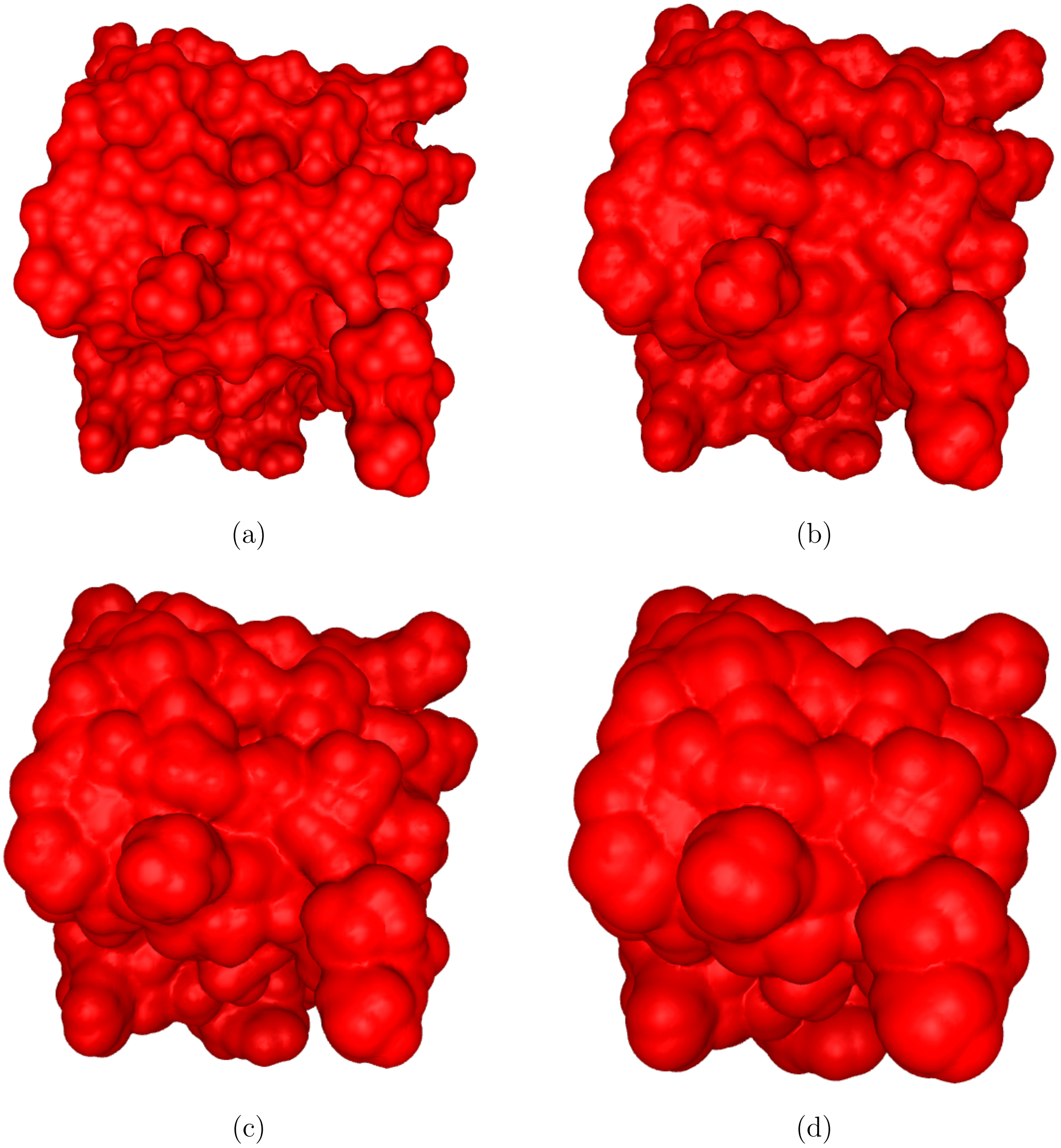
(a)

(b)

(c)

(d)

Figure 4.16: The level set representation of the SES for the C_{60} molecule. (a) the level set representation of the SES; (b), (c), and (d) represent the frame of the evolution at time 20, 40, and 60, respectively.



(a)

(b)

(c)

(d)

Figure 4.17: The level set representation of the SES for protein 1clh. (a) The level set representation of the original SES; (b), (c), and (d) Frames of propagated surfaces at time 25, 50, and 100, respectively.

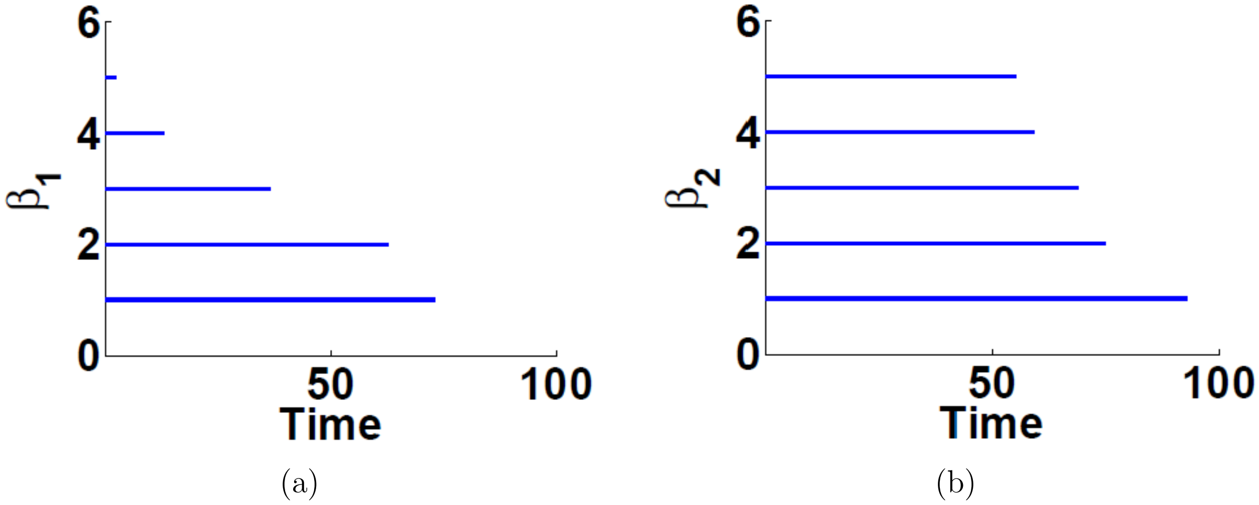


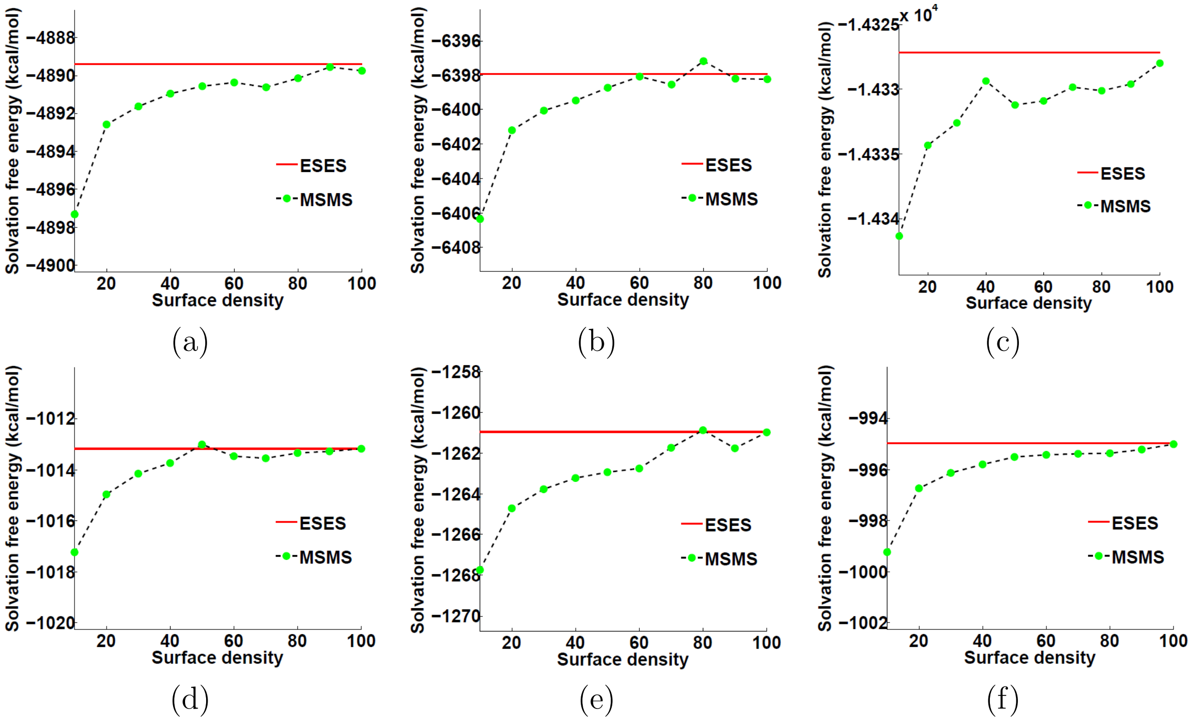
Figure 4.18: The persistence diagrams of the loops (Left chart) and cavities (Right chart) in the SES of protein 1clh, respectively.

numerical results is consistent with this intuition.

The left and right charts of Fig. 4.18 show the persistence of the loops and cavities on the molecular surface of the protein molecule 1clh, respectively. The persistence barcode of the loops shows that there are two quite short lived loops, i.e., two tiny loops on the surface, and one middle sized loops, together with two long persistent loops. The largest loop has persistent length of 73, which corresponds to a diameter of 14.6 Å. The cavity persistence barcode demonstrates that there is no short lived cavity in the SES manifold. Five cavities all have a relatively long persistence. The largest cavity has persistence length around 93, which corresponds to a cavity length of 18.6 Å.

4.5 ESES: Electrostatics Analysis

To further validate the present ESES software, we consider electrostatic solvation free energy calculations using both the ESES and MSMS surfaces. Here the electrostatic solvation free energies are computed based on the PB model introduced in the previous chapter, the



(a) (b) (c)

(d) (e) (f)

Figure 4.19: The convergence of the electrostatics solvation free energies calculated by using MSMS surfaces to those obtained by using the ESES surfaces. (a), (b) and (c) are for nucleic peptide molecules with PDB IDs 1A2E, 1BNA and 1L4J, respectively. (d), (e) and (f) are for protein molecules with PDB IDs 1a93, 1aca and 1b8w, respectively. All the energies are obtained by solving the PB equation with the MIBPB software at grid size 0.5 Å.

numerical method will be presented in the next chapter. For the sake of simplicity, we consider the pure water solvent with the solvent dielectric constant set to be 80 while 1 for the solute.

Obviously, for a given molecule with the same force field assignment and the same PB solver, when the SESs generated by different software packages are consistent with each other, the calculated electrostatics solvation free energies should be the same. In the following, we demonstrate that with the increasing of the density of the MSMS surface, the MSMS surface based electrostatics solvation free energies converge to those based on the ESES surface.

Figure 4.19 shows the convergence of the electrostatics solvation free energies calculated by using the MSMS surfaces to the ESES surfaces. All the calculations are carried out by the highly accurate PB solver MIBPB software [90, 278, 280, 293, 253], in which the PB equation is solved on a Cartesian mesh. It is shown the highly accurate and robust property of the MIBPB solver in calculating the electrostatics solvation free energies [253]. We will discuss this PB solver in the next chapter. The MSMS surface is generated with densities varying from 10 to 100. Since the MSMS surface is given by the Lagrangian representation, i.e., the triangulation of the molecular surface, a Lagrangian to Cartesian transformation is employed to embed the MSMS surface to the Cartesian mesh [292].

4.6 Conclusion

Solvent excluded surface (SES) is the most popular surface definition in computational biophysics and molecular biology for biomolecular modeling and simulation. Existing SES software packages, such as MSMS [212], typically provide SESs in the Lagrangian representation. For applications in implicit solvent models, one needs to convert the triangular SES into the Eulerian representation, i.e., the Cartesian domain. Additionally, quality of MSMS depends on the density selected and the method might not work well at all required densities for certain molecules. Therefore, it is desirable to generate analytical SESs directly on the Cartesian mesh. This work offers a software package, called Eulerian solvent excluded surface (ESES), for the construction of SESs on the Cartesian mesh.

We generate analytical SESs based on Connolly's algorithm [51], which divides a SES into three patches: convex patches, saddle patches and concave patches. The mathematical representation, computing algorithm, and data structures for each individual patch are formulated. We immerse the analytical SES into the Euclidean space $\mathbb{R}^3$ and describe the surface or interface by its intersecting coordinates with the Cartesian mesh lines and associated normal directions at all the intersecting points.

The proposed ESES software is validated by a large number of benchmark tests, including morphological visualization, solvation analysis, surface area and enclosed volume calculation, and topological feature analysis and characterization. We utilize the Amber PBSA test set in our validation. The MSMS software is employed for comparison. In the morphological visualization, it is shown that ESES successfully generate correct morphology while MSMS does not always work for the Amber test set at all densities. ESES also provides second order accurate estimates for biomolecular surface area and enclosed volume. It is found that electrostatic solvation free energies computed using the ESES are in close consistence with those calculated based on MSMS. A special feature of the present software is that it provides atomic surface areas calculation, which can be used for atomic modeling of nonpolar solvation free energies. Finally, we introduce homology theory to accurately detect topological features, namely, loops and cavities on/in SESs. A novel level set based filtration is proposed to measure the sizes of loops and cavities.

The present ESES software will be improved on a few aspects. First, a better method for rendering the surface need to be implemented for the surface visualization. Second, compared to the MSMS software, ESES is analytical but slower. Therefore, acceleration via multi-thread computational techniques will be further investigated to make the ESES software more efficient. Third, a robust triangulation to the ESES is durable for the further work.

Chapter 5

Coarse Grid Poisson Boltzmann Solver

5.1 Introduction

Poisson Boltzmann (PB) model is a multiscale model which models the electrostatics of the solvated molecules especially for the biological system, through sophisticated physical modeling of the focusing part, e.g., solvated biomolecules modeled with atomistic detail, the ions in the solvent is modeled as a density distribution, and the solvent is modeled as a dielectric continuum. The PB model is one of the most important implicit solvent models, especially for the study of the biological systems.

There are mainly two challenges in numerical solution to the PB model: One is the description of the solvated molecular conformation structure; the other one is development of the accurate and convergent numerical solution to the PB equation (PBE). The first issue is addressed in the previous chapter, the second issue will be discussed in this chapter.

With proper force field parametrization of the solute molecule, mathematical challenges in terms of the numerical approximation to the PBE can be summarized as: (i) efficient construction of the molecular surface of the solvated solute molecule; (ii) treatment of the singular charges of the solute molecule, which is represented by the singular δ-function; (iii) treatment of the complex interface geometry in the elliptic interface problem of the PBE;

(iv) accurate evaluation of the electrostatic solvation free energy after resolving the PBE.

There are extensive PB software available, and the numerical method to solve the PB model can be classified into three categories: finite difference method (FDM), finite element method (FEM), and boundary element method (BEM). Here we give a short review of the current existing popular PB software, the critical advantages of each PB software will be briefly discussed.

- AFMPB [162]: Adaptive Fast Multipole Poisson-Boltzmann (AFMPB) solver is a numerical simulation package for solving the linearized Poisson-Boltzmann (LPB) equation which models electrostatic interaction in biomolecule systems. In this package, a boundary integral equation approach is applied to discretize the LPB equation.

- APBS [7]: The adaptive Poisson-Boltzmann Software (APBS), together with the PDB2PQR are software packages designed to help the users analyze the solvation properties of small and macro-molecules such as proteins, nucleic acids, and other complex systems. The multigrid algorithm is employed for the discretization of the PB equation, it is one of the most popular and widely used PB software.

- Delphi [209]: The Delphi software is a very accurate and efficient PB software. In which the solvated molecular conformation is described by the smoothed numerical surface. The induced surface charge method is employed for accurate and efficient PB calculation. It is tested that the grid size influence on the PB calculation is very small.

- Amber PBSA [259]: Amber PBSA is the PB software in the Amber software, which solves both linear and nonlinear forms of PBE. Various algorithms are implemented to solve the linear PBE, such as, conjugate gradient, modified incomplete Cholesky conjugate gradient (ICCG), geometric multigrid, and successive over-relaxation methods

(SOR); and to solve nonlinear PBE, as the inexact Newton method in conjunction with modified ICCG or geometric multigrid, conjugate gradient, SOR, and other complex systems.

- PBEQ-Charmm [127]: The PBEQ module within Charmm MD distribution allows the setting up and the numerical solution of the PBE on a discretized grid for a solute molecule.

- MIBPB [39]: Matched Interface and Boundary Based Poisson-Boltzmann (MIBPB) Solver is a software package for evaluating electrostatic properties of biomolecules via the solution of the PBE, an established two-scale model in biomolecular simulations. It distinguishes itself from other PBE solvers by rigorously enforcing the interface flux continuity condition. This chapter is focused on improving the accuracy and robustness of the previous MIBPB software.

- ZAP [101]: The ZAP software produces PB electrostatic potentials and, from them, biologically interesting properties including solvent transfer energies, binding energies, pKa shifts, solvent forces, electrostatic descriptors, surface potentials and effective dielectric constants. ZAP TK works well for small molecules, proteins and macromolecular ensembles. Unique to ZAP TK is a dielectric function based on atom-centered Gaussians, which avoids the pitfalls of discrete dielectric constants [276].

This chapter is organized as follows: In section 5.2 we will formulate the PB model as an elliptic interface problem, the interface and boundary conditions will be presented. We will present the numerical methods for solving the four mathematical issues listed above in section 5.3, in which both MSMS and ESES surface [157, 212] are utilized for characterizing the solute molecular conformation structure. The Green's function [49, 90] technique is adopted for removing singular source in the PBE, the matched interface and boundary method (MIB) [278, 90] is used for handling the complex interface geometry, and the numerical method for evaluating the reaction field energy will also be discussed. The numerical results for the electrostatics solvation free energy and binding free energy calculation are presented in sections 5.4 and 5.5, respectively. This chapter ends up with a conclusion.

5.2 Poisson Boltzmann model

In this section, we will present a review for the PB model in the variational principle point of view. Consider an open domain $\Omega \in \mathbb{R}^3 = \Omega^{\mathrm{m}} \bigcup \Gamma \bigcup \Omega^{\mathrm{s}}$, where Ω^{m} is the solute domain that enclosed by the surface formed by the biomolecule, and Ω^{s} is the solvent domain. Γ is the molecular surface, for instance, the SES, that separates the solute and solvent domains.

First the charge distribution in the solute and solvent domains are modeled at the following different scales:

- Consider the atomistic modeling of the solute domain, in which the charge distribution $\rho_{\mathrm{m}}(\mathbf{r})$ is given by:
$$\rho_{\mathrm{m}} := \rho_{\mathrm{m}}(\mathbf{r}) = \sum_{i=1}^{N_m} Q_i \delta(\mathbf{r} - \mathbf{r}_i),$$
where Q_i is the partial charge of the i-th atom, and N_m is the number of charged atoms.

- The solvent domain is modeled as a dielectric continuum medium, in which the charge distribution is modeled by the Boltzmann distribution, mathematically formulated as:
$$\rho_{\mathrm{s}} := \rho_{\mathrm{s}}(\mathbf{r}) = \sum_{j=1}^{N} q_j c_j e^{q_j \phi / k_B T},$$

where N is the number of ionic species, c_j's are the bulk concentration of each ionic species, and q_j's are charges of each ionic species, k_B is the Boltzmann constant, T is the absolute temperature.

Furthermore, the permittivity $\epsilon(\mathbf{r})$ are both assumed to be constants in solute and solvent domain:

$$\epsilon(\mathbf{r}) = \begin{cases} \epsilon_{\rm s}, & \text{if } \mathbf{r} \in \Omega^{\rm s} \\ \epsilon_{\rm m}, & \text{if } \mathbf{r} \in \Omega^{\rm m}, \end{cases}$$

in the PB model, $\epsilon_{\rm s}$ and $\epsilon_{\rm m}$ are usually set to be 80 and 1, respectively. This specially selected constants will be adapted for all the numerical results in this chapter.

The electrostatic solvation free energy of the whole solvation system can be expressed as:

$$G^{\rm elec} = \int_\Omega \left(\chi_{\rm m}\rho_{\rm m}\phi - \chi_{\rm m}\frac{\epsilon_{\rm m}}{2}|\nabla\phi|^2 - \chi_{\rm s}\frac{\epsilon_{\rm s}}{2}|\nabla\phi|^2 + \chi_{\rm s}k_BT\sum_{j=1}^{N}[c_j(e^{-q_j\phi/k_BT}-1)] \right) d\mathbf{r}. \tag{5.2.1}$$

where we have introduced the characteristic function $\chi_{\rm m}$ and $\chi_{\rm s}$ for the solute and solvent domain, respectively.

Taking the variational derivative of $G^{\rm elec}$ with respect to ϕ, i.e., set $\frac{\delta G^{\rm elec}}{\delta\phi} = 0$, yields the following PB equation:

$$-\nabla \cdot (\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})) = \rho_m + \rho_s. \tag{5.2.2}$$

The Eq. (5.2.2) derived from the variational principle is consistent with the one introduced in chapter 2.

In the following context of this chapter, for the sake of simplicity, we only consider the simple case that ionic strength in the solvent is 0, i.e., pure water solvent. The ionic solvent case can be treated in the same manner.

The PBE is well posed by subjected to the following interface and boundary conditions constraint:

- The far field boundary condition:

$$\phi(\infty) = 0.$$

  In practical computation, the following Debye-Hückel boundary condition are enforced:

$$\phi(\mathbf{r}) = \sum_{i=1}^{N_m} \frac{Q_i}{4\pi\epsilon_{\mathrm{s}}|\mathbf{r} - \mathbf{r}_i|} \tag{5.2.3}$$

- Across the interface Γ, the continuity of the electrostatics potential and flux are enforced:

  – Continuity of electrostatics potential:

$$[\phi] = \phi_{\mathrm{s}}(\mathbf{r}) - \phi_{\mathrm{m}}(\mathbf{r}) = 0. \tag{5.2.4}$$

  – Continuity of electrostatics flux:

$$[\epsilon(\mathbf{r})\phi_{\mathbf{n}}] = \epsilon_{\mathrm{s}}\nabla\phi_{\mathrm{s}}(\mathbf{r}) \cdot \mathbf{n} - \epsilon_{\mathrm{m}}\nabla\phi_{\mathrm{m}}(\mathbf{r}) \cdot \mathbf{n} = 0, \tag{5.2.5}$$

  where $\mathbf{n} = (n_x, n_y, n_z)$ is the outer normal direction of the interface Γ which pointing from the solute domain to the solvent domain.

5.3 Numerical Method

In this section, we will present the numerical method for solving the PBE, the numerical method is highly accurate, and make coarse grid PB solver without loss of accuracy possible. The numerical method mainly contains three parts: The construction of the solvent solute boundary, we choose SES in this context. Treatment of the singular charges that represented by the δ-function for describing the solute charge distribution, which will be treated by the Green's function technique. Treatment of the complex solvent solute boundary geometry, we adopt the Matched Interface and Boundary method (MIB) for treating the versatile geometry. Furthermore, we will present the new electrostatic solvation free energy calculation scheme.

5.3.1 Solvent Solute Boundary

In the PB model, the solvent and solute domains are separated by the molecular surface and each part is modeled as a dielectric medium with a given dielectric constant, the SES is the most widely used surface definition, detail description has already been presented in the previous chapter. In our numerical coarse grid PB solver, we implements both ESES and MSMS, ESES is suit for extremely accurate calculation, while MSMS is objected for fast calculation with a slightly accuracy reduction.

5.3.1.1 MSMS Surface

The MSMS surface is the reduced surface of the original SES, which is the first software that can handle the geometric singularities that arise from the self-intersecting surfaces. For detail about the MSMS surface, the readers are referred to Sanner et al's work [212].

In our interface method based PB solver, we have to transform the Lagrangian represented surface, i.e., the triangulation of the SES, into the Eulerian representation. This transformation needs to embed the triangulated surface into a bounding box in the three dimensional Euclidean space $\mathbb{R}^3$. The embedding contains the following three major steps:

- Classification inside and outside grid points.
- Find the intersections of the triangulated reduced SES with the Cartesian mesh line.
- Find the out normal direction at the intersection points.

The classification of the grid points can be done via a ray tracing technique, which is based on the discrete Jordan curve lemma. The intersections and out normal vectors computation can be simply done via some basic Euclidean geometric knowledge. The detailed steps are presented in the work [292]. Compare to the embedding of the analytical SES into the Euclidean space, this embedding is much simpler, since MSMS is represented by the simplicial complexes, all the equations associated to the transformation are linear.

5.3.1.2 ESES surface

ESES surface, as presented in the previous chapter, is a direct Eulerian represented SES designed for the finite difference type of method based PB solvers. During the ESES generation and further processing, no numerical error is introduced for the grid points classification, intersections detection, and outer normal vectors calculation. All these three instruments are accurate without any numerical error, this provides the foundation for the highly accurate PB solver.

5.3.2 Solving the Poisson Boltzmann equation

The numerical method for solving the PB model contains two parts, treatment of the singular charges and treatment of the complex geometry of the interface. Classical approach for handling this singular charges is the direct projection of these charges to the Cartesian grid points. The deficient of this approach is that when the coarse grid is applied, the solute charge may be projected to the grid points in the solvent domain. This unreasonable projection may leads to unacceptable error when the large difference of the dielectric constants exists between solvent and solute domains. Another treatment of these singular charges is based on the Green's function of the singular charges, and in terms of accuracy, Green's function formalism is superior to the classical direct projection method [90].

To handle the complex geometry of the interface in the PB model, or a more general class of elliptic interface problems. We need to note the fact in Lemma 5.3.1, which helps to the development of the numerical scheme for the elliptic interface problems.

Lemma 5.3.1. *If the interface Γ is locally Lipschitz, and a given function ϕ defined in both Ω^{m} and Ω^{s} are continuous and the first order partial derivatives are well defined. At any point of the interface, the derivatives of function are continuous along the tangential directions, provided the function is continuous across the interface.*

Proof. (Sketch of the proof.) Let us denote the function ϕ to be ϕ_{m} and ϕ_{s} in Ω^{m} and Ω^{s}, respectively. And let $f \doteq \phi_{\mathrm{m}} - \phi_{\mathrm{s}}$, it is obviously that $f|_{\Gamma} = 0$.

For $\forall (x_0, y_0, z_0) \in \Gamma$, there is a tangential plane Π defined at (x_0, y_0, z_0) to the interface Γ, $\forall \tau \in \Pi$ we have:

$$\frac{\partial \phi_{\mathrm{m}}}{\partial \tau} - \frac{\partial \phi_{\mathrm{s}}}{\partial \tau} = \frac{\partial f}{\partial \tau} = \nabla f \cdot \tau.$$

Further, note that $f|_{\Gamma} = 0$, which means $\Gamma \subset \{(x, y, z)|f(x, y, z) = 0\}$, i.e., the interface

Γ can be represented by the implicit function $f(x, y, z) = 0$.

Obviously, $\nabla f = \mathbf{n}$, which is the direction normal to $\mathbf{n}$ at (x_0, y_0, z_0). Hence

$$\frac{\partial \phi_{\mathrm{m}}}{\partial \tau} - \frac{\partial \phi_{\mathrm{s}}}{\partial \tau} = \nabla f \cdot \tau = \mathbf{n} \cdot \tau = 0.$$

By the arbitrary of τ, the conclusion obtained. □

Based on the above lemma, we can introduce two more interface conditions on the tangential directions, thus a group of four interface conditions will be available for designing numerical scheme for solving PBE. This is tested to be enough for designing second order convergent numerical scheme for discretizing the PBE. One state-of-the-art accurate scheme MIB method utilized the above fact without proof[278, 295]. Many other prestigious schemes also exist for solving the elliptic interface problem. The global second convergence of the elliptic interface numerical scheme is proved in the work [17], which provides a general framework for proving the convergence of the general finite difference based elliptic interface scheme, Mayo's method [170, 168] and Immersed Interface Method (IIM) scheme [115, 149] is utilized for the illustrations.

5.3.3 Electrostatic Solvation Free Energy Calculation

One major application of the PB model is for evaluating the electrostatic solvation free energy, which also provides a good criteria for measuring the accuracy of the PB solver. In this section, we will provide the numerical method for calculating the electrostatics solvation free energies.

5.3.4 Reaction field potential representation of the solvation free energy

In the conventional implicit solvent theory, the reaction field potential is defined as the difference between the electrostatic potential in solvent and in vacuum, that is,

$$\phi_{\text{rec}}(\mathbf{r}) = \phi_{\text{dielec}}(\mathbf{r}) - \phi_{\text{vac}}(\mathbf{r}), \tag{5.3.1}$$

where $\phi_{\text{rec}}, \phi_{\text{dielec}}$ and ϕ_{vac} are the reaction field potential, electrostatic potential in solvent and vacuum, respectively. The vacuum electrostatic potential is calculated through setting the solvent dielectric constant the same as that in the solute domain.

The electrostatics solvation free energy, or reaction field energy, is defined by:

$$\Delta G_{\text{RF}} = \frac{1}{2}\sum_{i=1}^{N_m} Q(\mathbf{r}_i)\phi_{\text{rec}}(\mathbf{r}_i), \tag{5.3.2}$$

where $Q(\mathbf{r}_i)$ are the charge at the position $\mathbf{r}_i$.

5.3.5 Approximate the atomic center reaction field potential

According to the equation for evaluating the electrostatics solvation free energy, Eq. (5.3.2), the electrostatics potential at the atomic centers are needed. Whereas, according to our numerical scheme, only the electrostatic potential on the grid of the Cartesian mesh grids is obtained, we need to approximate the electrostatic potential at the atomic centers by that at the grid points. In this work, the trilinear interpolation scheme will be utilized for interpolating the electrostatics potential at the atomic centers.

For $\forall \mathbf{r}_i \doteq (x_0, y_0, z_0)$, suppose the closest grid to $\mathbf{r}_i$ inside the solute domain is indexed by (i, j, k). The following 27 grids will be utilized for the further interpolation:

$$\{(i+m, j+n, k+p)|m=-1,0,1; n=-1,0,1; p=-1,0,1\}. \tag{5.3.3}$$

The trilinear interpolation scheme is given in the following steps:

- Interpolate the values at the following 9 points:

  $$\{(x_0, j+n, k+p)|n=-1,0,1; p=-1,0,1\},$$

  by the above 27 grids $\{(i+m, j+n, k+p)|m=-1,0,1; n=-1,0,1; p=-1,0,1\}$, due to the utilization of the uniform Cartesian mesh, we have:

  $$\phi_{\text{rec}}|_{(x_0,j+n,k+p)} = w_{x,-1}\phi_{\text{rec}}|_{(i-1,j+n,k+p)} + w_{x,0}\phi_{\text{rec}}|_{(i,j+n,k+p)} + w_{x,1}\phi_{\text{rec}}|_{(i+1,j+n,k+p)},$$

  for $n=-1,0,1; p=-1,0,1$, where $w_{x,-1}, w_{x,0}$ and $w_{x,1}$ are the Lagrangian interpolation coefficients, it is easy to be obtained via Fornberg's method [77].

- Interpolate the values at the following 3 points:

  $$\{(x_0, y_0, k+p)|p=-1,0,1\},$$

  by the nine points $\{(x_0, j+n, k+p)|n=-1,0,1; p=-1,0,1\}$, similar to the above scheme, we have:

  $$\phi_{\text{rec}}|_{(x_0,y_0,k+p)} = w_{y,-1}\phi_{\text{rec}}|_{(x_0,j-1,k+p)} + w_{y,0}\phi_{\text{rec}}|_{(x_0,j,k+p)} + w_{y,1}\phi_{\text{rec}}|_{(x_0,j+1,k+p)},$$

for $p = -1, 0, 1$, here $w_{y,-1}, w_{y,0}$ and $w_{y,1}$ are the interpolation coefficients.

- Interpolate the value at the atom center (x_0, y_0, z_0) by the 3 points $\{(x_0, y_0, k+p)|p = -1, 0, 1\}$, which is given by:

$$\phi_{\text{rec}}|_{(x_0,y_0,z_0)} = w_{z,-1}\phi_{\text{rec}}|_{(x_0,y_0,k-1)} + w_{z,0}\phi_{\text{rec}}|_{(x_0,y_0,k)} + w_{z,1}\phi_{\text{rec}}|_{(x_0,y_0,k+1)},$$

  $w_{z,-1}, w_{z,0}$ and $w_{z,1}$ are the interpolation coefficients.

In sum, the approximation of the reaction field potential at the atomic center $\mathbf{r}_i$ is given by:

$$\phi_{\text{rec}}|_{(x_0,y_0,z_0)} = \sum_{m=-1}^{1}\sum_{n=-1}^{1}\sum_{p=-1}^{1} w_{x,m}w_{y,n}w_{z,p}\phi_{\text{rec}}|_{(i+m,j+n,k+p)}. \tag{5.3.4}$$

5.3.5.1 The extension of the reaction field potential

In the Eq.(5.3.4), for the atomic centers that close to the boundary when the coarse grid employed, it is possible that some of this grids may in the solvent domain, if we directly use the reaction field potential computed from the previous PB solver, the accuracy in evaluating the solvation free energy will be reduced. Therefore, we need to extend the reaction field potential in the solute domain to the outside grids that referred in interpolating the reaction field potential at the atomic centers.

For a given grid (i_1, j_1, k_1) belongs to the above 27 grids, the schemes that used for extension of the reaction field potential at (i_1, j_1, k_1), $\phi_{\text{rec}}(i_1, j_1, k_1)$, are listed in the following according to its priority, the scheme employed should have as top priority as possible.

- **Top priority:** Use the sum of fictitious value and the extended solution of the boundary value problem (specifically for the Green's function treatment of the singular charge case) at the grid point (i_1, j_1, k_1) as the extended reaction field potential at (i_1, j_1, k_1).

- **Middle priority:** Choose three consecutive grid points next to (i_1, j_1, k_1) along a given direction in the solute domain to extrapolate the reaction field potential at (i_1, j_1, k_1).

- **Low priority:** Select two inside grids neighbor to (i_1, j_1, k_1) , say, (i_2, j_2, k_2) and (i_3, j_3, k_3) and approximate the reaction field potential at (i_1, j_1, k_1) by:

$$\phi_{\text{rec}}(i_1, j_1, k_1) = \phi_{\text{rec}}(i_2, j_2, k_2) + \phi_{\text{rec}}(i_3, j_3, k_3) - \phi_{\text{rec}}(i, j, k).$$

5.3.6 Electrostatic Binding Free Energy Calculation

Another important application of the PB model is the calculation of the electrostatic binding free energy, which is a crucial part of the binding energy between different molecules, it is of great importance for the computer aided drug discovery.

Consider the binding of two molecules A and B, where the complex after the binding of two molecules is denoted by C, by a simple thermodynamic cycle analysis, the electrostatics binding free energy of this process is given by:

$$\Delta\Delta G_{\text{el}} = (\Delta G_{\text{RF}})_C - (\Delta G_{\text{RF}})_A - (\Delta G_{\text{RF}})_B + (\Delta\Delta G_{\text{el}})_{\text{coulomb}}, \tag{5.3.5}$$

where $(\Delta G_{\text{RF}})_C$ is the electrostatic solvation free energy of the bounded complex, $(\Delta G_{\text{RF}})_A$ and $(\Delta G_{\text{RF}})_B$ are the electrostatic solvation free energies of the unbounded components, and $(\Delta\Delta G_{\text{el}})_{\text{coulomb}}$ is the electrostatics binding free energy of the two components in vacuum,

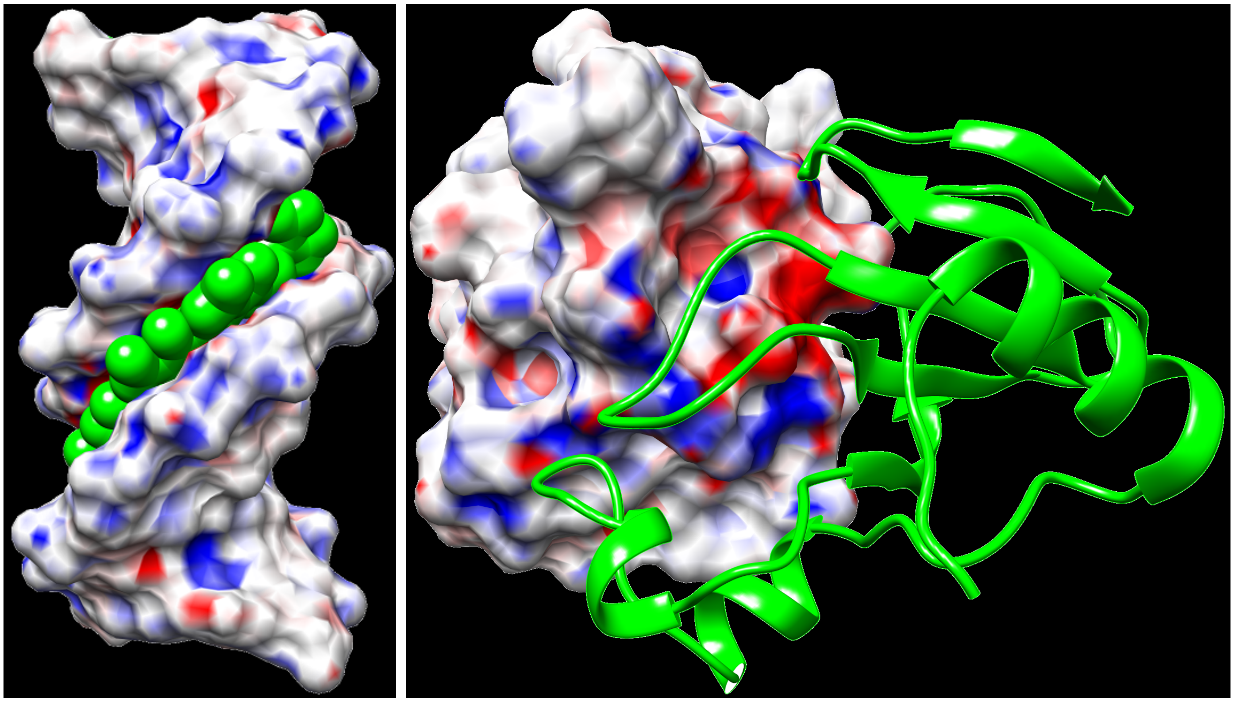


Figure 5.1: Left chart: binding of DNA-drug complex (PDB ID: 121D). Right chart: binding of barnbase-barstar complex (PDB ID: 1b3s).

which is computed by the formula:

$$(\Delta\Delta G_{\mathrm{el}})_{\mathrm{coulomb}} = \sum_{i,j} \frac{q_i q_j}{\epsilon_{\mathrm{m}} r_{ij}}, \tag{5.3.6}$$

where the summation is taken over all those pairs of atoms which have one member in each component of the complex, q_i, q_j are the corresponding charges of the given interacted pair atoms, and r_{ij} is the distance between this pairs, ϵ_{m} is the dielectric constant of the solute domain as defined before.

Figure 5.1 illustrates two binding examples, the left chart is binding of the DNA-drug complex, the right chart is binding of the barnbase-barstar complex.

5.4 Numerical Results-Solvation Free Energy

In this section, a large amount of numerical results that used to verify the grid spacing independent properties of the current PB solvers will be presented, the consistent with different PB solvers will also be addressed. The robustness of the PB solver is directly verified by the huge amount of the test examples. Our previous results shown that with the increasing of the surface triangulate density [157], the solvation free energy on MSMS surface will converge to our ESES analytical solvent excluded surface, also the ESES is much more robust than the MSMS surface according to the test on the Amber test set.

5.4.1 Analytical Tests

In this subsection, the PB solver will be tested by the benchmark tests with exact solution, says, the dielectric sphere with a single center distributed charge and multiple charges, the analytical solution to the previous one is given by the dielectric Born theory, while Kirkwood theory [136] gives the analytical solution to the later case.

5.4.1.1 Single center-distributed dielectric sphere

First, consider the dielectric sphere with single central distributed unit charge with different radius. Born theory states that, for a dielectric sphere with radius R and center charge q placed in a solvent with dielectric constant ϵ and 1 for the dielectric sphere itself, the electrostatic solvation free energy is

$$\Delta G_{\mathrm{RF}} = -\frac{q^2}{2R}\left(1 - \frac{1}{\epsilon}\right). \tag{5.4.1}$$

Table 5.1 shows the electrostatic solvation free energy for the dielectric sphere with a central distributed unit positive charge, calculated from grid size 0.1 to 1.1 Å and the exact value. Different radius of the dielectric spheres are tested, where the radius are the typical atomic radii in the Amber force field.

Table 5.1: Electrostatic solvation free energies of the dielectric spheres with the central located unit positive charge equipped with different radius.

$\frac{\text{GridSize}}{\text{Atomradius}}$	1.1	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	Exact
1.1	-143.63	-116.22	-148.63	-145.9	-146.03	-148.84	-149.00	-148.98	-149.01	-149.04	-149.05	-149.05
1.3	-125.75	-118.24	-126.03	-123.07	-125.75	-125.99	-126.03	-126.07	-126.10	-126.11	-126.12	-126.12
1.359	-120.36	-117.25	-120.61	-119.68	-120.21	-120.54	-120.59	-120.61	-120.63	-120.64	-120.65	-120.65
1.4	-116.88	-103.73	-117.10	-116.89	-116.97	-117.03	-117.07	-117.08	-117.10	-117.11	-117.11	-117.11
1.5	-109.17	-103.43	-109.18	-109.17	-109.22	-109.25	-109.28	-109.29	-109.29	-109.30	-109.31	-109.13
1.55	-105.68	-101.74	-105.61	-105.68	-105.70	-105.72	-105.74	-105.76	-105.77	105.78	-105.78	-105.78
1.7	-94.44	-95.80	-96.34	-96.40	-96.39	-96.40	-96.42	-96.43	-96.44	-96.44	-96.45	-96.45
1.8	-90.96	-90.96	-91.02	-91.01	-91.04	-91.05	-91.07	-91.07	-91.08	-91.09	-91.09	-91.09
1.85	-88.52	-88.52	-88.57	-88.57	-88.58	-88.59	-88.61	-88.62	-88.62	-88.62	-88.63	-88.63
2.0	-81.86	-81.95	-81.90	-81.92	-81.95	-81.96	-81.98	-81.97	-81.98	-81.98	-81.98	-81.98

According to the above results, the electrostatic solvation free energies calculated by the current PB solver converge very fast to the exact value. Furthermore, it should be pointed out that the proposed method can provide very reliable calculation even with very low grid resolution. For instance, when the dielectric sphere with radius 1.1 Å, the calculated reaction field energy is -143.63 kcal/mol at grid spacing 1.1 Å(that is, only one grid inside the dielectric sphere), the exact value of which is -149.05 kcal/mol.

5.4.1.2 Multiple charge dielectric sphere

To further test the accuracy of the PB solver, especially the Green's function treatment of the singular charges, in this part, we further test on the dielectric sphere, while now there are multiple charges distributed in the dielectric sphere. The analytical solution to these cases is due to Kirkwood's work [136].

We consider the following five different distributions of point charges but their radii are all set to be 2Å.

- Case 1. Two positive unit charges symmetrically placed at $(1, 0, 0)$ and $(-1, 0, 0)$.

- Case 2. Two positive unit charges symmetrically placed at $(1, 0, 0)$ and $(-1, 0, 0)$, and two negative unit charges symmetrically placed at $(0, 1, 0)$ and $(0, -1, 0)$.

- Case 3. Two positive unit charges symmetrically placed at $(1.2, 0, 0)$ and $(-1.2, 0, 0)$, and two negative unit charges symmetrically placed at $(0, 1.2, 0)$ and $(0, -1.2, 0)$.

- Case 4. Six Positive unit charges placed at (0.4, 0.0, 0.0), (0.0, 0.8, 0.0), (0.0, 0.0, 1.2), (0.0, 0.0, -0.4), (-0.8, 0.0, 0.0) and (0.0, -1.2, 0.0).

- Case 5. Six Positive unit charges placed at $(0.2, 0.2, 0.2), (0.5, 0.5, 0.5), (0.8, 0.8, 0.8), (-0.2, 0.2, -0.2), (0.5, -0.5, 0.5)$ and $(-0.8, -0.8, -0.8)$.

The first three cases were employed by Li to compare Delphi and Amber PB performance [3]. The last two cases were used in our earlier work in 2007 [90], which also contains many test examples similar to the first three cases. These benchmark tests are more difficult than the Born model. We highly recommend PB methodology developers to consider these test cases before they try to make any bold statement about the PB model and/or their methods.

Table 5.2 lists electrostatic solvation free energies of the MIBPB method for the above multiple charge tests on a set of mesh sizes range from 1.1 to 0.1Å. For Case 1, all MIBPB relative errors are less than 1%. For Case 2, MIBPB is smaller than 5% on all grid sizes. Case 3 is relatively difficult because charges are located more close to the interface. For the last two cases, errors are bound by 1.5% on all grid sizes.

5.4.2 Robust on different SES

In this part, we employ 25 biomolecules and data set used in Li's work [3] as the test set for comparing the performance of the Delphi [209] and Amber PBSA [259] PB software,

Table 5.2: Electrostatic solvation free energies (kcal/mol) of Kirkwood dielectric sphere with multiple charges calculated by MIBPB, where RE is the relative error compare to the analytical electrostatic solvation free energy.

	Case 1		Case 2		Case 3		Case 4		Case 5	
Grid	$\Delta G_{\rm RF}$	RE	$\Delta G_{\rm RF}$	RE	$\Delta G_{\rm RF}$	RE	$\Delta G_{\rm RF}$	RE	$\Delta G_{\rm RF}$	RE
1.1	-351.12	0.39%	-63.61	1.27%	-135.41	0.00%	-2974.59	0.49%	-3079.70	1.43%
1.0	-352.54	0.80%	-65.66	4.53%	-152.32	12.45%	-2952.83	1.22%	-3138.85	0.47%
0.9	-348.84	0.25%	-60.70	3.36%	-131.88	2.59%	-2993.72	0.15%	-3099.03	1.80%
0.8	-351.20	0.42%	-64.13	2.10%	-137.33	1.42%	-2996.27	0.23%	-3110.38	0.44%
0.7	-350.56	0.24%	-63.24	0.68%	-135.17	0.16%	-2995.02	0.19%	-3103.84	0.65%
0.6	-350.68	0.27%	-63.77	1.52%	-137.34	1.43%	-2994.73	0.18%	-3118.62	0.18%
0.5	-350.39	0.19%	-63.50	1.09%	-136.64	0.92%	-2986.62	0.09%	-3124.43	0.00%
0.4	-350.02	0.08%	-63.10	0.46%	-136.27	0.64%	-2991.70	0.08%	-3119.22	0.15%
0.3	-349.81	0.02%	-61.95	1.36%	-136.20	0.59%	-2991.01	0.06%	-3120.24	0.13%
0.2	-349.72	0.00%	-62.90	0.14%	-135.79	0.29%	-2989.89	0.02%	-3122.89	0.04%
0.1	-349.64	0.02%	-62.81	0.00%	-135.40	0.00%	-2989.54	0.00%	-3123.90	0.01%
Exact	-349.73		-62.81		-135.40		-2989.30		-3124.30	

the data was originally downloaded from the protein data bank. All the HETATM records in PDB files are removed by using MMTSB toolset, and the AMBER99SB force field was employed for the PDB file parametrization. The structures were further optimized by the AMBER software with the same protocol used by the work [3]. The PDB IDs for this set of proteins are: 1ajj, 1ptq, 1vjw, 1bor, 1fxd, 1sh1, 1hpt, 1fca, 1bpi, 1r69, 1bbl, 1vii, 2erl, 451c, 2pde, 1cbn, 1frd, 1uxc, 1mbg, 1neq, 1a2s, 1svr, 1o7b, 1a63, and 1a7m.

Table 5.3 shows the electrostatic solvation free energies (kcal/mol) calculated by using the MIBPB method with SESs generated by the MSMS software at density 10. Similarly, Table 5.4 shows the the electrostatic solvation free energies (kcal/mol) calculated by the MIBPB software with SESs generated by the ESES software. First, there are some minor discrepancies (less than 1%) between the electrostatic solvation free energies computed by two SESs. MSMS accuracy depends on its triangular mesh density and its results converge to those ESES at a very high triangular mesh density, say 100 triangles per Å^2. Additionally, for a given surface, MIBPB is able to deliver very consistent results. The results listed in the

Table 5.3: Electrostatic solvation free energies (kcal/mol) calculated via the MIBPB software with MSMS surface at different grid sizes.

					Grid size (Å)					
ID	1.1	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
1ajj	-986.26	-984.17	-986.38	-986.89	-985.82	-986.68	-986.64	-986.85	-987.09	-987.15
1ptq	-723.95	-722.99	-724.50	-725.34	-724.54	-725.51	-725.18	-725.49	-725.64	-725.68
1vjw	-1120.64	-1117.29	-1119.03	-1119.37	-1120.61	-1121.60	-1121.98	-1121.52	-1121.95	-1122.12
1bor	-773.51	-774.38	-777.01	-776.91	-778.85	-778.04	-778.33	-778.65	-778.66	-778.7
1fxd	-2424.97	-2433.53	-2432.13	-2433.24	-2434.06	-2435.02	-2434.64	-2435.11	-2435.14	-2435.38
1sh1	-568.12	-570.69	-570.74	-571.80	-573.34	-573.98	-573.48	-573.88	-574.05	-574.32
1hpt	-662.19	-666.16	-664.79	-666.10	-667.18	-666.89	-667.57	-668.17	-668.25	-668.45
1fca	-1175.78	-1181.58	-1181.77	-1180.31	-1181.93	-1181.71	-1181.72	-1182.04	-1182.10	-1182.11
1bpi	-1150.37	-1148.62	-1149.68	-1150.61	-1151.08	-1151.53	-1152.48	-1152.38	-1152.63	-1152.82
1r69	-945.34	-938.93	-940.63	-940.05	-942.28	-942.08	-941.88	-941.98	-942.15	-942.23
1bbl	-846.48	-846.80	-847.00	-846.90	-846.64	-847.33	-847.45	-847.08	-847.14	-847.10
1vii	-683.20	-677.26	-679.27	-680.61	-679.94	-680.31	-681.19	-681.28	-681.40	-681.55
2erl	-919.91	-915.41	-918.49	-918.92	-919.00	-919.68	-919.84	-920.21	-920.33	-920.57
451c	-848.53	-843.01	-849.20	-847.92	-847.92	-847.63	-847.85	-848.21	-848.03	-848.07
2pde	-798.26	-799.19	-798.87	-799.86	-800.03	-798.81	-799.39	-799.44	-799.22	-799.26
1cbn	-298.67	-301.97	-303.85	-303.68	-304.11	-304.46	-304.41	-304.79	-304.81	-304.92
1frd	-2561.35	-2556.63	-2557.71	-2558.70	-2557.06	-2559.33	-2560.09	-2560.74	-2561.27	-2561.72
1uxc	-917.39	-914.88	-918.19	-919.73	-920.18	-921.01	-921.32	-921.66	-921.75	-922.02
1mbg	-1286.82	-1281.16	-1283.11	-1285.37	-1285.74	-1285.61	-1285.67	-1286.08	-1286.23	-1286.45
1neq	-1474.17	-1471.80	-1470.82	-1472.88	-1473.94	-1474.13	-1474.74	-1475.49	-1475.62	-1475.63
1a2s	-1846.33	-1842.61	-1849.40	-1847.77	-1848.00	-1847.38	-1848.25	-1848.48	-1848.86	-1849.09
1svr	-1319.62	-1320.81	-1320.06	-1323.10	-1322.45	-1323.89	-1324.12	-1324.76	-1324.68	-1324.99
1o7b	-1738.56	-1747.28	-1743.36	-1743.60	-1743.93	-1745.48	-1746.05	-1746.18	-1746.77	-1747.12
1a63	-1940.89	-1940.05	-1936.96	-1939.81	-1941.67	-1942.27	-1942.61	-1943.36	-1944.14	-1944.55
1a7m	-2028.72	-2033.44	-2032.74	-2029.40	-2031.87	-2033.64	-2033.31	-2033.66	-2034.26	-2034.32

tables 5.3 and 5.4 confirm the grid-size independence property of the MIBPB software, also the numerical solution to the PB does not heavily depend on the solvent excluded surface generation.

Figure 5.4 depicts the relative errors of the electrostatic solvation free energies calculation by MIBPB software on surfaces generated by both ESES and MSMS. These results demonstrate that MIBPB software on both surfaces are of the same level the accuracy. Their relative errors are remarkably less than 0.4% when the mesh size is refined from 1.1Å to 0.2Å. Therefore, if one's goal is 1% relative error, one can just use a mesh size as coarse as 1.1Å in MIBPB based solvation analysis.

Table 5.4: Electrostatic solvation free energies (kcal/mol) calculated via the MIBPB with ESES surface at different grid sizes.

	Grid size (Å)									
ID	1.1	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
1ajj	-996.11	-995.63	-993.17	-993.03	-991.34	-988.16	-987.20	-986.99	-987.12	-987.03
1ptq	-720.68	-719.62	-720.68	-720.76	-721.48	-721.63	-721.71	-722.04	-721.95	-721.97
1vjw	-1122.67	-1123.35	-1124.63	-1127.13	-1124.10	-1125.55	-1125.15	-1125.20	-1125.06	-1125.22
1bor	-771.50	-773.85	-774.17	-772.50	-774.31	-773.74	-774.04	-774.37	-774.44	-774.45
1fxd	-2420.46	-2425.07	-2423.90	-2426.91	-2427.53	-2427.33	-2428.19	-2428.53	-2428.50	-2429.05
1sh1	-564.58	-564.16	-569.12	-569.23	-569.74	-569.50	-570.31	-570.59	-571.01	-571.20
1hpt	-661.72	-656.56	-661.03	-661.16	-663.54	-663.66	-664.00	-663.96	-664.10	-664.15
1fca	-1190.83	-1193.85	-1191.26	-1190.80	-1191.67	-1187.04	-1188.83	-1188.04	-1187.68	-1187.26
1bpi	-1139.46	-1145.12	-1147.15	-1145.35	-1146.96	-1147.57	-1147.99	-1148.14	-1148.28	-1148.36
1r69	-932.98	-937.44	-934.49	-936.89	-937.27	-937.32	-937.10	-937.44	-937.59	-937.66
1bbl	-838.68	-839.89	-841.26	-842.17	-843.02	-842.79	-842.00	-842.80	-842.73	-842.65
1vii	-672.12	-674.33	-674.94	-677.13	-676.71	-677.78	-677.69	-678.05	-678.42	-678.64
2erl	-914.31	-919.35	-918.70	-918.65	-917.98	-917.68	-918.68	-918.60	-918.79	-918.87
451c	-839.89	-838.39	-842.24	-841.94	-842.56	-842.44	-842.56	-843.26	-843.78	-843.96
2pde	-809.06	-814.63	-812.91	-810.62	-814.13	-813.32	-812.27	-813.04	-813.59	-813.42
1cbn	-300.65	-299.73	-301.79	-302.88	-301.94	-302.24	-302.37	-302.50	-302.60	-302.67
1frd	-2546.19	-2549.42	-2551.50	-2553.10	-2554.68	-2555.89	-2556.43	-2557.30	-2557.69	-2558.10
1uxc	-909.77	-908.98	-912.19	-915.23	-914.40	-916.34	-916.68	-916.93	-917.18	-917.44
1mbg	-1277.10	-1278.07	-1277.80	-1278.14	-1280.52	-1281.24	-1281.42	-1281.74	-1281.86	-1282.06
1neq	-1459.44	-1464.56	-1467.13	-1467.57	-1468.99	-1469.33	-1469.55	-1469.64	-1470.11	-1470.12
1a2s	-1834.76	-1838.39	-1841.63	-1842.01	-1841.72	-1842.06	-1842.67	-1843.19	-1843.19	-1843.61
1svr	-1309.47	-1310.67	-1314.35	-1316.24	-1317.25	-1317.10	-1317.75	-1317.89	-1318.42	-1318.37
1o7b	-1734.47	-1726.36	-1739.98	-1741.12	-1739.71	-1740.59	-1742.22	-1742.57	-1743.25	-1743.39
1a63	-1937.29	-1942.42	-1940.78	-1943.18	-1941.26	-1942.13	-1941.64	-1942.12	-1942.27	-1942.67
1a7m	-2038.64	-2032.48	-2034.48	-2034.91	-2038.38	-2040.39	-2037.91	-2038.62	-2038.73	-2038.26

5.4.3 Comparison between Different PB Solvers

In this part we will compare the performance of our PB solver with the Amber PBSA and Delphi PB solvers, the comparisons reflects the consistency of the three solvers, and the grid independent essence of the MIBPB solver. Figure 5.3 depicts the electrostatic solvation free energies for the 25 protein molecules calculated by the three different PB solvers at the grid size from 0.2 to 1.1 Å. The results demonstrates that not only the grid almost independent properties of the MIBPB solver in computing the electrostatic solvation free energy, but also the consistence of the three PB solver. In general, the solvation free energies calculated by the Amber PBSA converge decrease to that of the MIBPB, whereas that by Delphi converges

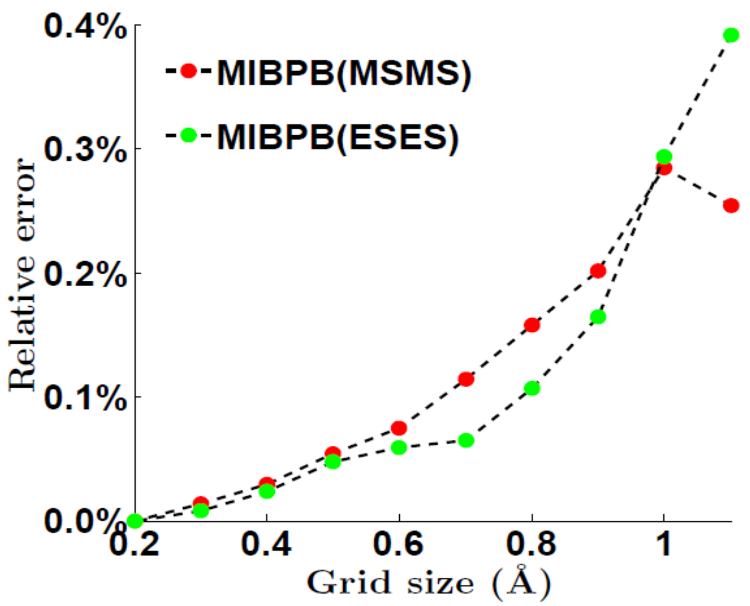


Figure 5.2: The relative errors of the electrostatic solvation free energies compared to the results calculated at 0.2Å computed by the MIBPB method on surfaces generated by ESES and MSMS averaged over 25 proteins.

increase to that of the MIBPB. MIBPB software are grid size independent for both MSMS and solvent excluded surfaces.

To measure the accuracy of the PB solver, since there is no analytical solution for the electrostatics solvation free energy on the biomolecules, the grid dependence of the energies are examined. Errors are estimated either with respect to the analytical solution if it is available or the result at the finest grid if there is no analytical result.

$$\text{Relative error} \doteq \frac{|\Delta G_h - \Delta G_{\text{finest grid}}|}{|\Delta G_{\text{finest grid}}|}, \tag{5.4.2}$$

where ΔG_h and $\Delta G_{\text{finest grid}}$ are the electrostatics solvation free energies calculated at the grid size h and the finest grid used.

Figure 5.4 depicts the relative deviation of the three PB solvers, where MIBPB utilized two different surfaces, at each grid compare to the results at the grid size 0.2 Å, obvious the

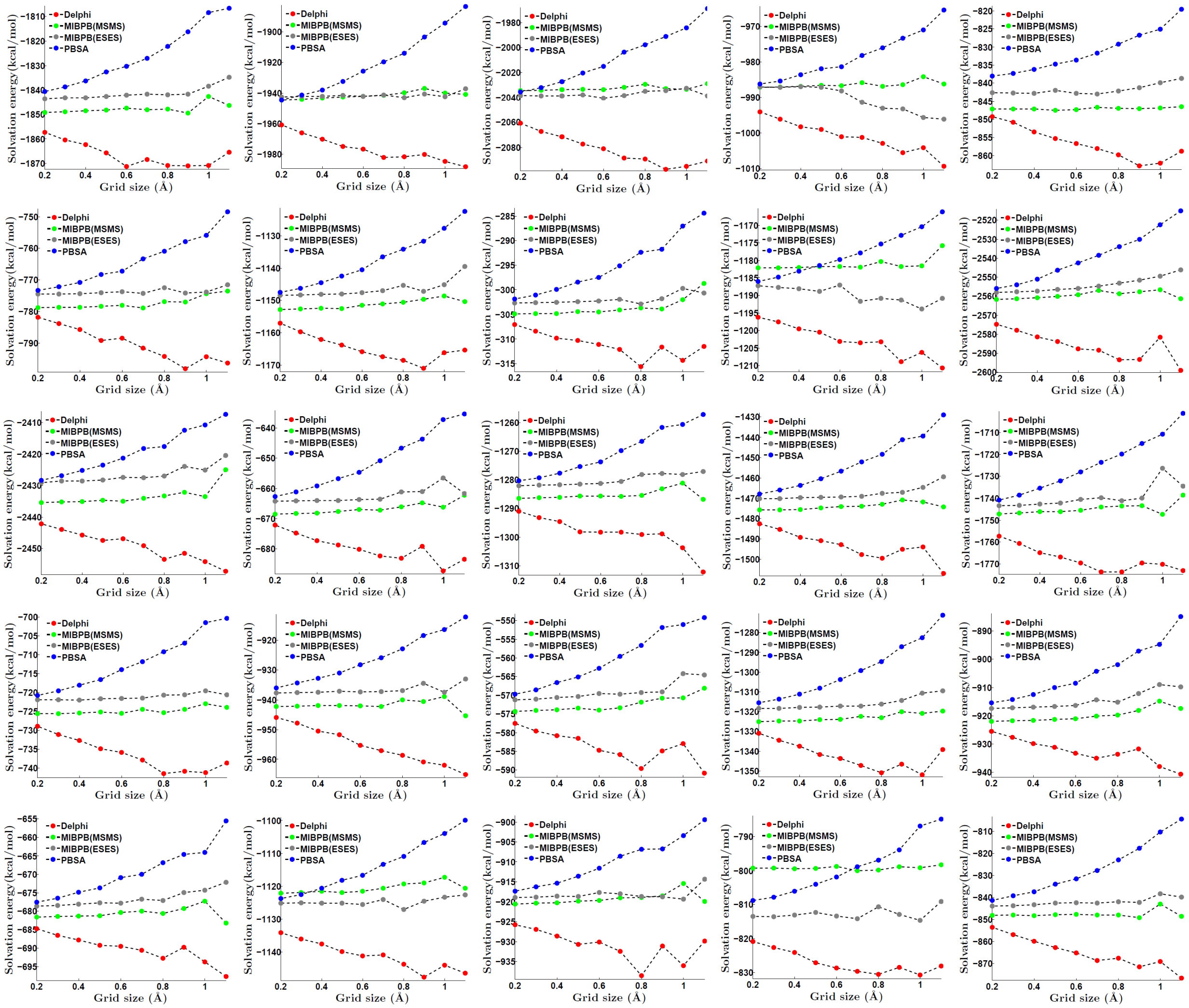


Figure 5.3: Performance of different PB solvers, and the MIBPB solver on both MSMS and solvent excluded surfaces. From left to right, up to bottom, the 25 sub-figures represents the molecule with PDB ID: 1a2s, 1a63, 1a7m, 1ajj, 1bbl, 1bor, 1bpi, 1cbn, 1fca, 1frd, 1fxd, 1hpt, 1mbg, 1neq, 1o7b, 1ptq, 1r69, 1sh1, 1svr, 1uxc, 1vii, 1vjw, 2erl, 2pde, 451c, respectively.

MIBPB solvers are much stabler than both Amber and Delphi, and the grid size independent properties of MIBPB solver is surface independent.

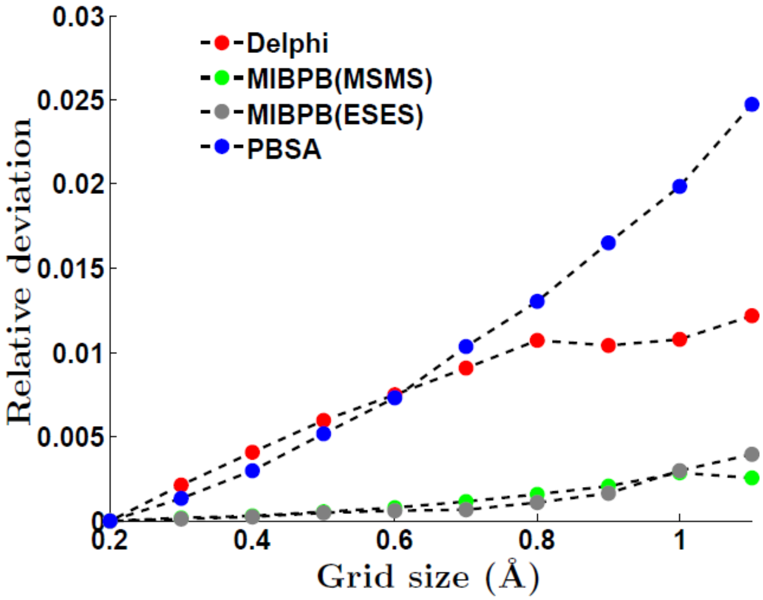


Figure 5.4: The relative deviation of the electrostatics solvation free energies by different PB solvers and MIBPB on different surfaces compare to the results at the smallest grid.

However, there is no free lunch. The stable of our PB solver in computing the electrostatics solvation free energy is build on more CPU cost at a given grid spacing. At a given grid size, both AMBER PBSA and Delphi are several times faster than our PB solvers. Different PB solvers have their own pons and cons. Our PB solver makes the coarse grid calculating possible which also leads to the speeds up the PB solver through the coarse grid electrostatic calculation.

Figure 5.5 illustrates the average cpu cost of the different PB solvers and the MIBPB solver on both MSMS and solvent excluded surfaces. All the test are carried out on the same Intel 14 computing node on the high performance computing center of the Michigan state university.

The test on the above 25 biomolecules shows that the electrostatics solvation free energies

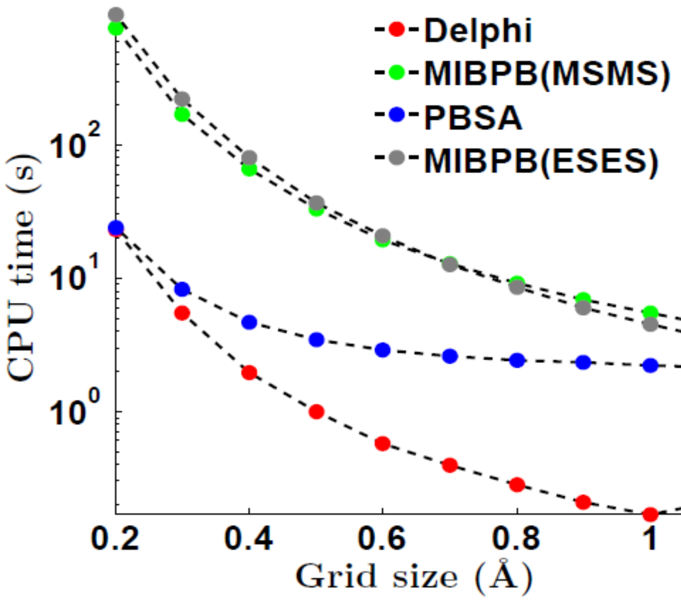


Figure 5.5: The average cpu cost of the 25 biomolecules of the different PB solvers and MIBPB on different surfaces.

calculated by three different PB solvers are consistent with each other, MIBPB solver is more stabler than both Amber and Delphi the results by MIBPB solver are almost grid size independent for both MSMS and solvent excluded surfaces. At the same grid size, both Amber and Delphi are several times faster than our PB solver.

5.4.4 Convergence Test on PBSA Test Set

To further test the grid size independent property of the current PB solver, in this part we adopt our PB solver to a much larger test test, the Amber PBSA test set, as already used as the benchmark test for the ESES surface generation in the previous chapter.

Figure 5.6 shows the relative deviation of the electrostatics solvation free energy with respect to that at grid size 0.3 Å for the 937 biomolecules. The deviation converges to 0 monotonically with the refinement of the grid size, and the error level is consistent with the above 25 test set. Also note that even at the grid size 1.1 Å the relative error is less than

0.4%, which further confirms that the grid size independent properties of the current PB solver.

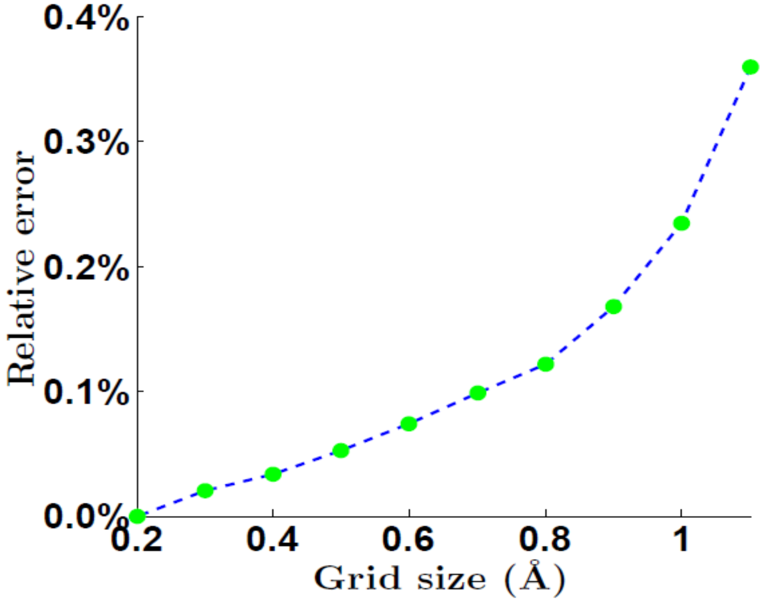


Figure 5.6: The relative deviation of the electrostatics solvation free energy with respect to that at grid size 0.3 Å for the 937 biomolecules.

5.5 Numerical Results-Binding Free Energy

After the study of the accuracy of the electrostatic solvation free energy, in this part, we turn to study the electrostatic binding free energy. The accurate electrostatic binding free energy calculation essentially depends on the accuracy of the electrostatic solvation free energy calculation. However, the accurate electrostatic binding free energy calculation is much more challenging, here the accuracy are measured in two ways, first is the qualitative ranking the molecules based on the energy, second is the quantitative result of the energy value. Due to the fact that the magnitude of electrostatic binding free energy itself is usually much less than that of solvation free energy, the numerical error may lead to unacceptable error, which may yields the different ranking results on the molecules provided different grid spacing used for the numerical solution of the PB equation [107, 24]. Recently, it has been shown that the widely used grid spacing of 0.5 Å produces unacceptable errors in $\Delta\Delta G_{\mathrm{el}}$ [107].

5.5.1 Data Sets

In the present work, we adopt three sets of biomolecular complexes employed in the literature [107] for solvation free energy and binding free energy estimations. Specifically, the first set, Data Set 1, is a collection of DNA-minor groove drug complexes having a narrow range of $\Delta\Delta G_{\mathrm{el}}$. The Protein Data Bank (PDB) IDs (PDBIDs) for this set are as follows: 102d, 109d, 121d, 127d, 129d, 166d, 195d, 1d30, 1d63, 1d64, 1d86, 1dne, 1eel, 1fmq, 1fms, 1jtl, 1lex, 1prp, 227d, 261d, 164d, 289d, 298d, 2dbe, 302d, 311d, 328d, and 360d. The second set, Data Set 2, includes various wild-type and mutant barnase-barstar complexes. Its PDBIDs are as follows: 1b27, 1b2s, 1b2u, 1b3s, 2aza4, 1x1w, 1x1y, 1x1u, and 1x1x. In the last set, Data Set 3, we investigate RNA-peptide complexes with following PDBIDs: 1a1t, 1a4t, 1biv, 1exy, 1g70, 1hji, 1i9f, 1mnb, 1nyb, 1qfq, 1ull, 1zbn, 2a9x, and 484d. The detail of the structural prepossessing can be found in Ref. Harris et al.

5.5.2 Results and Discussion

As described above, we consider three sets of binding complexes, namely, drug-DNA, barnase-barstar and RNA-peptide systems. In the rest of this section, we explore the influence of grid spacing in PBE solvation and binding free energy estimations using our MIBPB solver.

Motivated by well-converged estimations of electrostatic solvation free energies at very coarse grid spacings as previously discussed, we are interested in predicting the binding free energies for all RNA-drug, barnase-barstar, and RNA-peptide complexes using our MIBPB package.

We correlate the binding free energy calculated at the finest grid spacing, h=0.2 Å, and ones estimated at coarser mesh sizes, h=0.3 Å, $\cdots$, 1.1 Å. Figure 5.7 illustrates these relationships with the regression lines whose parameters are revealed in Table 5.5. Indeed, the PB binding energy estimation behaves the same as the PB solvation calculation in our MIBPB technique. Specifically, R^2 is always 1 at the fine mesh, h = 0.3 Å. Moreover, these values are still satisfactory at relatively coarser mesh sizes. For example, at the grid spacing of h = 1.1, the R^2 and slope of the regression line for DNA-drug, barnase-barstar, and RNApeptide complexes are, respectively, (0.9747,1.0081), (0.9965,0.9974), and (0.9999, 0.9974). In contrast, the R-squared values reported in Ref. [107] computed between 0.3Åand 1.0 Åare unacceptable for SESs, and usually less than 0.62. Our statistical measures strongly support the reliable binding energy prediction of our solver at coarse grid sizes.

The trend of binding free energy at different grid spacings can be seen clearly in Figs. 5.8-5.10 which plots $\Delta\Delta G_{\rm el}$ against grid size varying between 0.2 Å and 1.1 Å for DNA-drug, barnase-barstar, and RNA-peptide complexes, respectively. Based on these figures, our solver can rank the binding free energy for DNA-drug complexes at grid spacing of 0.6 Å barnase-barstar complexes at grid spacing of 0.8 Å and RNA-peptide complexes at significantly coarse grid spacing of 1.1 Å. Therefore, we can draw a conclusion that the common use of grid size being 0.5 Å is still adequate for predicting the binding energy free without producing a misleading result.

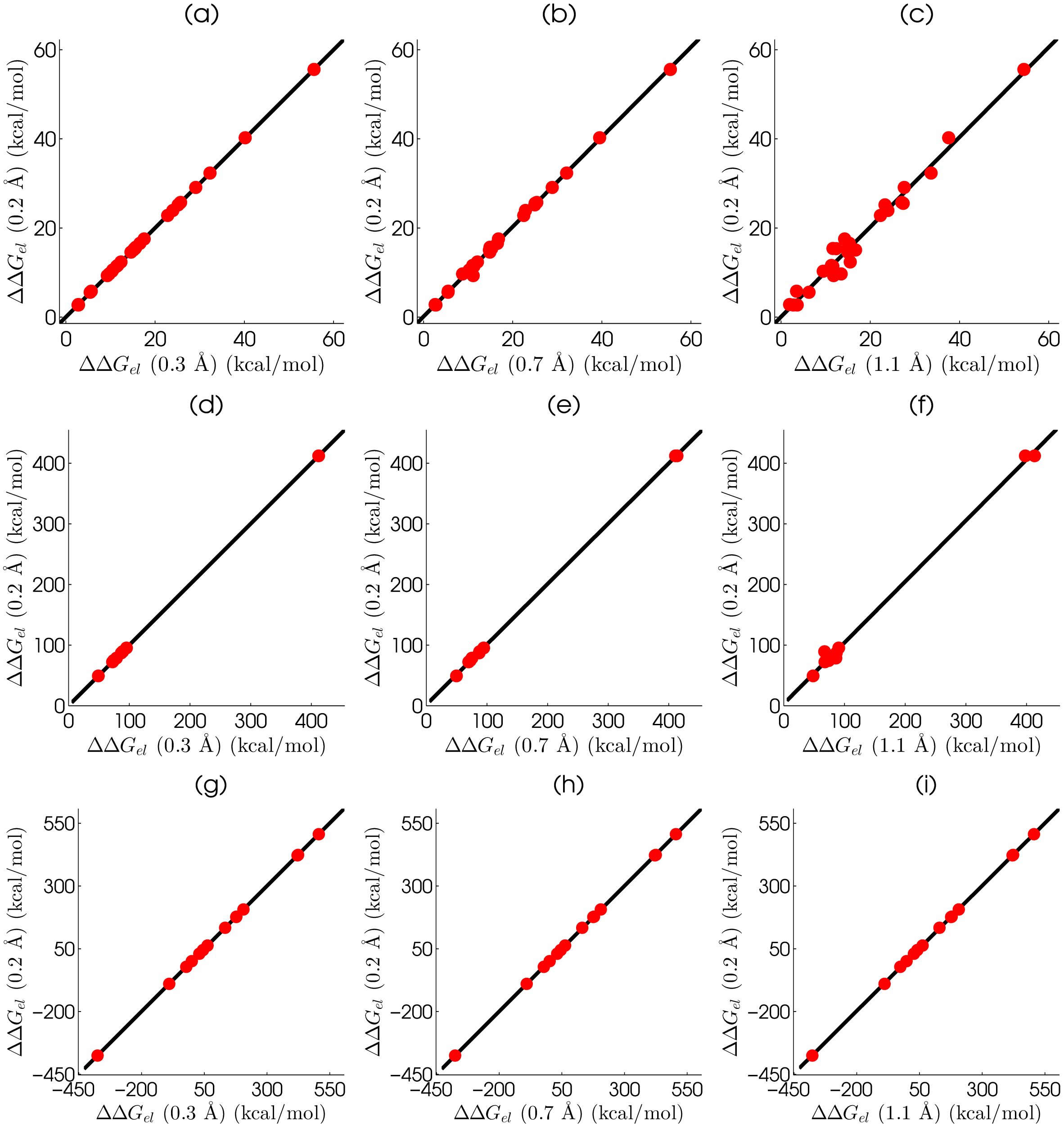


Figure 5.7: Electrostatic binding free energy, for all complexes with different grid sizes plotted against the one computed with a finest grid size of h = 0.2 Å(a) DNA-drug with pair (0.2 Å, 0.3 Å); (b) DNA-drug with pair (0.2 Å, 0.7 Å); (c) DNA-drug with pair (0.2 Å, 1.1 Å); (d) Barnase-barstar with pair (0.2 Å, 0.3 Å); (e) Barnase-barstar with pair (0.2 Å, 0.7 Å); (f) Barnase-barstar with pair (0.2 Å, 1.1 Å); (g) RNA-peptide with pair (0.2 Å, 0.3 Å); (h) RNApeptide with pair (0.2 Å, 0.7 Å); (i) RNA-peptide with pair (0.2 Å, 1.1 Å).

Table 5.5: R^2 values and best fitting lines of electrostatic binding free energies with different grid sizes.

	Grid sizes (pair)	R^2	Best fitting line
DNA-drug	(0.2, 0.3)	1.0000	y=0.9993x+0.0194
	(0.2, 0.4)	0.9999	y=0.9987x+0.0273
	(0.2, 0.5)	0.9998	y=1.0028x+0.0164
	(0.2, 0.6)	0.9991	y=1.0047x+0.2256
	(0.2, 0.7)	0.9982	y=1.0074x+0.1394
	(0.2, 0.8)	0.9966	y=1.0110x+0.1484
	(0.2, 0.9)	0.9906	y=0.9655x+1.2385
	(0.2, 1.0)	0.9875	y=0.9827x+0.5894
	(0.2, 1.1)	0.9747	y=1.0081x+0.0709
Barnase-barstar	(0.2, 0.3)	0.9999	y=0.9974x+0.2035
	(0.2, 0.4)	0.9999	y=0.9997x-0.0492
	(0.2, 0.5)	0.9995	y=1.0318x-2.7755
	(0.2, 0.6)	0.9946	y=0.9878x+1.5525
	(0.2, 0.7)	0.9932	y=1.0090x+0.1819
	(0.2, 0.8)	0.9883	y=0.9976x+3.7333
	(0.2, 0.9)	0.9493	y=0.9382x+5.3970
	(0.2, 1.0)	0.9384	y=1.0912x-3.8377
	(0.2, 1.1)	0.8002	y=0.9974x+18.2837
RNA-peptide	(0.2, 0.3)	1.0000	y=0.9997x-0.0655
	(0.2, 0.4)	1.0000	y=1.0001x-0.1106
	(0.2, 0.5)	1.0000	y=1.0012x-0.2755
	(0.2, 0.6)	1.0000	y=0.9999x+0.2021
	(0.2, 0.7)	0.9999	y=1.0037x-0.3756
	(0.2, 0.8)	1.0000	y=1.0004x+0.6673
	(0.2, 0.9)	0.9999	y=0.9927x+1.9755
	(0.2, 1.0)	0.9997	y=0.9923x+2.8775
	(0.2, 1.1)	0.9998	y=0.9937x+1.7992

5.6 Conclusion

Poisson-Boltzmann (PB) theory is an established model for biomolecular electrostatic analysis and has been widely used in electrostatic solvation and binding energy estimation. In this chapter, we present a grid size almost independent PB solver, i.e., MIBPB software, it makes the accurate and coarse grid PB software possible. The main theme of the MIBPB solver

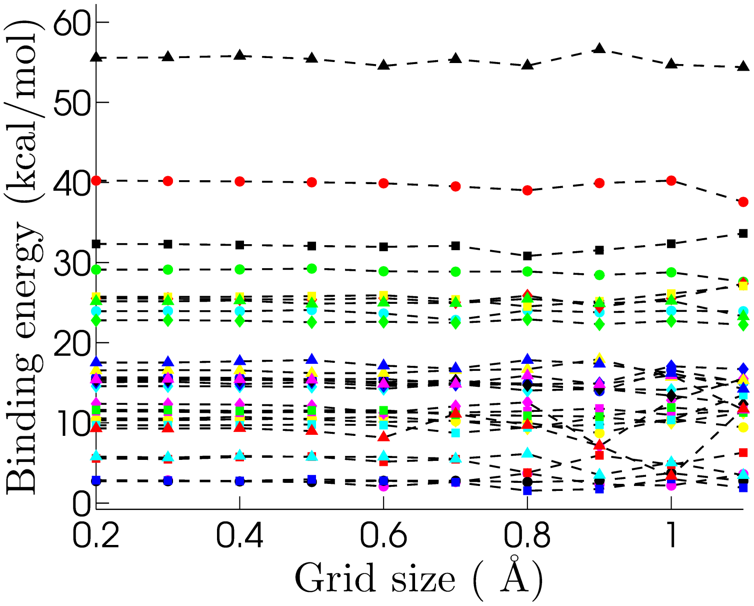


Figure 5.8: Binding electrostatic energy for DNA-drug complexes with grid sizes from 0.2 Å to 1.1 Å. The markers and PDBIDs are as follows yellow circle : 102d, magenta circle : 109d, cyan circle : 121d, green circle : 127d, red circle : 129d, blue circle : 166d, black circle : 195d, yellow diamond : 1d30, magenta diamond : 1d63, cyan diamond : 1d64, green diamond : 1d86, red diamond : 1dne, blue diamond : 1eel, black diamond : 1fmq, yellow square : 1fms, magenta square : 1jtl, cyan square : 1lex, green square : 1prp, red square : 227d, blue square : 261d, black square : 264d, yellow triangle: 289d, magenta triangle: 298d, cyan triangle: 2dbe, green triangle: 302d, red triangle: 311d, blue triangle: 328d, black triangle: 360d.

is rigorously treating four details that referred in the PB equation. First, the molecular surface definition, SES, is analytically implemented in the Cartesian grid for the purpose of the discretization of the PB equation, which is different from the molecular surface used in most of the current existing PB solvers, most of which utilized the approximated SES instead of the exact SES. Second, the singular charges are treated by the Green's function instead of conventional method that project the singular charges to the closest eight grid points, the projection methods usually yields the charges be projected to the grid in the solvent domain when the coarse grid is employed, the Green's function treatment makes the coarse grid treating of the singular charges possible. Third, the interface conditions arise in the PB equation are treated rigorously through the interface conditions matching, in the literature,

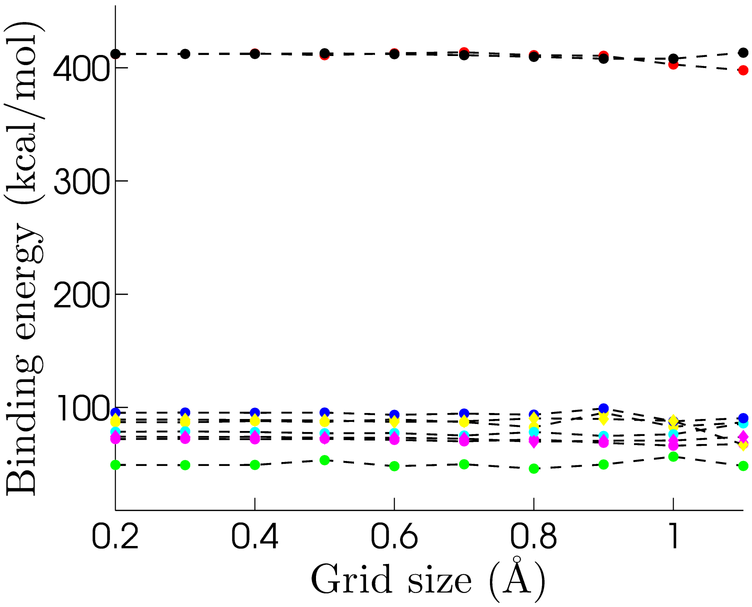


Figure 5.9: Binding electrostatic energy for barnase-barstar complexes with grid sizes from 0.2 Å to 1.1 Å. The markers and PDBIDs are as follows yellow circle : 1b27, magenta circle : 1b2s, cyan circle : 1b2u, green circle : 1b3s, red circle : 1x1u, blue circle : 1x1w, black circle : 1x1x, yellow diamond: 1x1y, magenta diamond: 2za4.

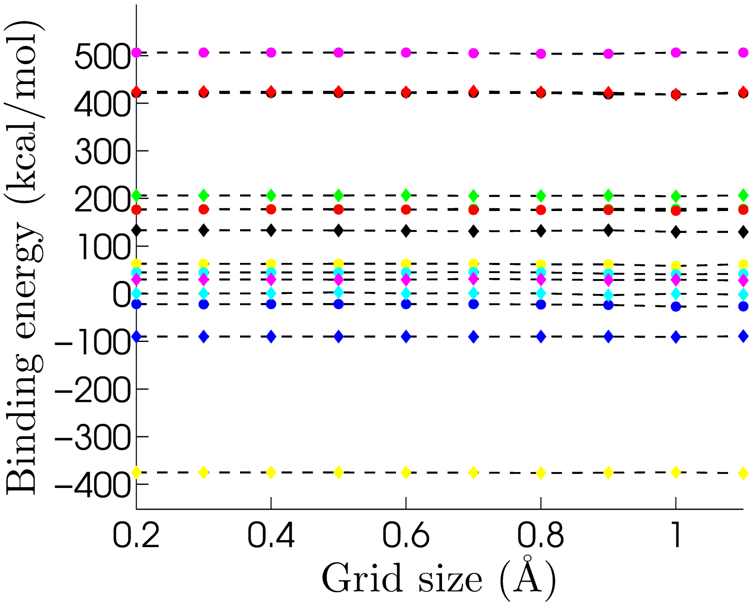


Figure 5.10: Binding electrostatic energy for RNA-peptide complexes with grid sizes from 0.2 Å to 1.1 Å. The markers and PDBIDs are as follows yellow circle : 1a1t, magenta circle : 1a4t, cyan circle : 1biv, green circle : 1exy, red circle : 1g70, blue circle : 1hji, black circle : 1i9f, yellow diamond : 1mnb, magenta diamond : 1nyb, cyan diamond : 1qfq, red diamond : 1ull, green diamond : 1zbn, blue diamond : 2a9x, black diamond : 484d.

there are some other methods that can treat these conditions with simple geometry while our method can handle geometry with arbitrary complex geometry. Fourth, the reaction field potential extension utilized for the evaluation of the reaction field energy, this posterior treatment avoids the accuracy reduction due to the usage of the reaction field potential in solvent.

Due to the mathematical rigorously treatment and the utilization of second order convergence scheme, the MIBPB solver converge very fast. Which leads to the grid size almost independent property of the PB solver, this property was verified by a large amount of test cases includes both analytic tests and the real biomolecule tests, all the tests demonstrate that the current PB solver is grid size independent. Through the test of the stability with respect to different surfaces, the reduced molecular surface (MSMS surface) and the analytical Eulerian solvent exclusive surface (ESES), the current PB solver's grid independent property is shown to be independent from the molecular surface used. Furthermore, the present paper solves two problems proposed in Robert C. Harris et als paper [107] namely, qualitative ranking the biomolecular complexes by their electrostatic binding free energies at different grid spacings, and the convergence calculation of the binding free energies on the solvent excluded surface. Highly accurate and highly stable calculation of the binding free energy plays a critical role in computational chemistry, also has lots of important pharmaceutical applications. With a better estimation of the binding free energies will helps a lot on the drug design and some other industries. The current MIBPB solvers perform extremely accurate and stable for both the electrostatic solvation and binding energies calculation. The current results show that the widely used grid size 0.5 Å can give accurate enough results for both qualitatively ranking the biomolecular complexes and the quantitatively evaluating the electrostatic binding free energies, actually when the grid size is less than 0.7 Å the results is already suitable for the ranking of the complexes, and the binding free energies calculated at the coarser grid is highly consistent and correlated with that at finer grid, the binding energies also converge to the benchmark results. In sum, this work developed a grid size independent PB solver, which makes the coarse grid PB solver becomes possible and indirectly speed up the PB solver.

Chapter 6

Hybrid Physical and Statistical Models for Solvation Free Energy Prediction

6.1 Introduction

Solvation in which separated solvent and solute molecules are combined to form a solution, is an elementary physical process in nature that provides a foundation for more complicated chemical and biological processes, such as ion channel permeations, protein ligand binding, electron transfer, signal transduction, DNA specification, transcription, post-transcription modification, gene expression, protein synthesis, etc. Therefore, the understanding of solvation is a prerequisite for the quantitative study of other more complex chemical and biological processes and is of paramount importance to chemistry, physics and biology [228, 60, 204, 228]. The most basic and reliable experimental observation of solvation is the solvation free energy, which measures the free energy change in the solvation process. As a result, one of the most important tasks in the solvation modeling and computation is the prediction of solvation free energies, which has recently drawn much attention in computational biophysics and chemistry [60, 140, 204, 228, 62, 266].

A large variety of solvation models have been developed in the past few decades for solvation analysis and solvation free energy prediction. In general, these models can be classified into either knowledge-based models or physical-based models. Knowledge-based models usually utilize a large amount of available data of solvation free energies to train statistical models for the solvation free energy prediction. One example of this type of model makes use of solvent accessible surface areas, solvent excluded surface areas or calibrated atomic surface areas for solvation free energy predictions [260]. Knowledge-based models can provide relatively accurate predictions provided that enough experimental data are available. On the other hand, physical-based solvation models are based on fundamental laws of physics. These models can be further classified into three major classes. One of them is explicit solvent models in which both solvent and solute molecules are described at the atomic or electronic level. This class of models are accurate in general, but usually computationally expensive [214, 156, 138]. Another class of models are integral equation based solvation theories [110, 203, 200], in which the solute molecule is still modeled at the atomic level, while the solvent is modeled by statistical mechanics, such as liquid density functional theory, Ornstein-Zernike equation with hypernetted-chain, Percus-Yevick equation or other specified closure. Integral equation based solvation models reduce the number of degrees of freedom dramatically compared to the explicit solvent models. A major feature of integral equation approaches is that these methods are able to provide a good approximation of solvent microstructures near the solvent-solute interface [91]. The other class of the physical based solvation models is implicit solvent models [192, 56, 211, 199, 83, 6, 90, 270, 43], in which the solute molecule can be modeled at either the atomic or the quantum mechanical level, while the solvent is described as a dielectric continuum.

Compared to other methods, implicit solvent models have advantages of involving a small number of degrees of freedom, highly accurate modeling of strong and long-range electrostatic interactions that often dominate solvation phenomena, and convenient treatments of solvent-solute electronic polarization effects [56]. In implicit solvent models, the solvation free energy is typically divided into polar and nonpolar components. The polar solvation free energy is due to electrostatic interactions and can be modeled by a number of approaches, including Born dielectric sphere model, Kirkwood model [136], Generalized Born (GB) model [124, 99, 95, 75, 64, 11], Polarizable Continuum Model (PCM) [121, 239, 71] and Poisson-Boltzmann (PB) theory [195, 218, 208, 112]. Among these approaches, Born dielectric sphere model treats the solute molecule as the dielectric sphere with a centrally located atomic charge is the simplest one. The Kirkwood model gives analytical expressions for electrostatic potentials in a spherical solute-solvent system with multiple charges inside the solute molecule [136]. The Generalized Born (GB) model [124, 99, 95, 75, 64, 11] is able to deal with arbitrarily shaped molecules and accounts for the impact of each point charge by its effective distance (i.e., the Born radius) from the solvent-solute interface [234, 57]. The GB model is relatively fast, while depends on other methods, such as the PB theory, for its parametrization. The PCM describes solute electrons by using quantum mechanics so that solvation induced polarizations can be accounted through an iterative procedure. A few different versions of PCMs have been reported, including dielectric PCM, integral equation formalism version PCM, and the variational PCM models [121, 239, 71]. PCM is relatively computationally expensive due to its quantum mechanical charge calculation while its accuracy is bounded by the PB type of treatment of electrostatic potential as well as the quality of solvent-solute interfaces. The PB theory can be derived from more fundamental Maxwell's equations [195, 218, 208, 112, 90]. It is more accurate than the GB model and computationally more efficient than the PCM. The PB can be easily coupled with quantum mechanics for a more accurate description of the solute charge density [262, 263, 45]. It has become one of the most popular solvation models due to its relatively low computational cost and high modeling accuracy for biomolecules.

The nonpolar solvation component has been modeled by a number of terms. A popular approach, based on the Scaled-Particle Theory (SPT) for nonpolar solutes in aqueous solutions [226, 197], is to use a solvent-accessible surface area (SASA) term [230, 167]. It was shown that a Solvent-Accessible Volume (SAV) term is relevant in large length scale regimes [163, 114]. Recent studies indicate that SASA based solvation models may not describe van der Waals interactions near solvent-solute interfacial region [84, 83, 50, 246]. A combination of surface area, surface enclosed volume, and van der Waals potential has been shown to provide accurate nonpolar solvation predictions [46, 248].

The polar and nonpolar components are typically decoupled in traditional implicit solvent models. Recently, new efforts have been made to couple polar and nonpolar components [70, 270, 43]. Differential geometry based solvation models make use of the differential geometry theory of surfaces to dynamically couple polar and nonploar models by surface evolution, which is driven by free energy optimization [270, 45, 43, 44]. Coupled with an optimization procedure, this model was shown to deliver some of the best nonpolar solvent free energy prediction [46]. By applying constrained optimization to nonpolar parameter selection, this model provides state-of-the-art solvation free energy fitting and cross validation results for a large number of solute molecules [248].

It is important to validate solvation models by experimental data. SAMPLx blind solvation prediction project, aimed at testing protein ligand binding models, also provides benchmark tests of the performance of the solvation models. Many different solvation approaches have been tested by the SAMPLx challenging molecules [183, 177, 106, 205, 176, 173, 207, 174, 81, 105, 86, 87]. Implicit solvent models are among the most competitive approaches in SAMPLx tests.

The objective of the present work is to develop a protocol that hybrids physical and knowledge (or statistical) models for the *blind* solvation free energy prediction. In our approach, physical models are used for modeling the polar solvation free energy. Specifically, we utilize PB theory for computing electrostatic solvation free energies. Both a point charge based PB model and a KSDFT based polarizable PB model are developed in the present work. For the KSDFT based PB model, an iterative process is developed to take care of the solvent polarization and solute response [45]. In our statistical models, the nonpolar solvation free energy is predicted based on the similarity analysis of solute molecules. Essentially, we assume that similar molecules admit the same set of nonpolar parameters. In this work, we examine two statistical models. One of them is based on a functional group scoring, and the other is based on a nearest neighbor approach. Both statistical models utilize machine learning methods to optimize parameters. A database contains the solvation free energies in aqueous solvent for 668 solute molecules is collected from the literature [31, 175] to validate the proposed solvation protocol. It is found that the present approach provides some of the best blind predictions of solvation free energies.

The rest of this chapter is organized as follows. In section 6.2, we present the polar solvation models, in which both the classical Poisson model and the polarizable Poisson model will be discussed. Section 6.3 presents two nonpolar solvation models, namely, functional group scoring and nearest neighbor approaches. An algorithm for unified blind solvation free energy prediction is developed in Section 6.3.3. The leave-one-out test of 668 molecules is employed to validate the proposed protocol. Finally, the accuracy and robustness of the proposed protocol is demonstrated by five SAMPL test sets.

6.2 Polar Solvation Models

In this section, we will derive the polarizable Poisson (Boltzmann) model based on the variational principle. We will start from reviewing the Poisson (Boltzmann) model from the energetic variational approach point of view, then we will introduce the energy functional for the polarizable Poisson (Boltzmann) model. There are two main advantages of the polarizable Poisson (Boltzmann) model:

- *Ab Initio* calculation of the solute molecule charge distribution, instead of the force field parametrization.
- Solvent solute polarization effects are incorporated into the solvation model.

6.2.1 Free Energy Functional

Consider the solution domain Ω formed by the solute and solvent molecules, which take the domains Ω^{m} and Ω^{s}, respectively. Consider the following multiscale descriptions of the solution system.

- Continuum description of the solvent domain, which models the solvent as dielectric continuum; and the atomic modeling of the solute domain, where the solute molecule is modeled as atoms equipped with partial charges at the atomic centers. The separation of the solute and solvent domain is described by the molecular surface of the solute molecules.
- Continuum description of the solvent domain, while the solute domain is modeled at the electronic level. The molecular surface of the solute molecule is employed as the separation of the two domains.

In the following we will formulate the polar free energy functional of the two multiscale models.

- For the first multiscale model, that is, continuum modeling of solvent domain and atomic modeling of the solute domain. The polar free energy functional can be written as:

$$G_1^{\mathrm{p}} = \int_\Omega \left\{ \chi \left[\rho_{\mathrm{total}}\phi - \frac{1}{2}\epsilon_{\mathrm{m}}|\nabla\phi|^2 \right] + (1-\chi)\left[-\frac{1}{2}\epsilon_{\mathrm{s}}|\nabla\phi|^2 \right] \right\} d\mathbf{r}, \tag{6.2.1}$$

  where χ and $1-\chi$ are the characteristic functions of the solute and solvent domains, respectively. $\rho_{\mathrm{total}} = \sum_{i=1}^{N_m} q_i \delta(\mathbf{r}-\mathbf{r}_i)$ is the partial charges inside the solute molecule, with q_i be the amount of partial charge located at the ith atom center $\mathbf{r}_i$, N_m be the number of atoms of the solute molecules, $\delta(\mathbf{r}-\mathbf{r}_i)$ is the delta function at the position $\mathbf{r}_i$. ϵ_{m} and ϵ_{s} are the dielectric constants of the solute and solvent domains, respectively. $\phi \doteq \phi(\mathbf{r})$ is the electrostatic potential field in the solution.

- For the second multiscale model, in which electronic modeling of the solute molecule is used to replace the atomic modeling in the first model, more specifically, we consider the Kohn-Sham Density Functional Theory (KSDFT) modeling of the electronic structure. The polar free energy functional now can be formulated as:

$$\begin{aligned} G_2^{\mathrm{p}} = \int_\Omega \chi \left[\hat{\rho}_{\mathrm{total}}\phi - \frac{1}{2}\epsilon_{\mathrm{m}}|\nabla\phi|^2 \right] + \quad (6.2.2) \\ (1-\chi)\left[-\frac{1}{2}\epsilon_{\mathrm{s}}|\nabla\phi|^2 \right] + \chi \left[\sum_i \frac{\hbar^2}{2m}|\nabla\psi_i|^2 + E_{XC}[n] \right] d\mathbf{r}, \end{aligned}$$

  where $\hat{\rho}_{\mathrm{total}} = qn(\mathbf{r}) - qn_n(\mathbf{r})$ is the charge distribution in the solute domain, with q be the unit charge of an electron, $n(\mathbf{r})$ be the electron density, $n_n(\mathbf{r}) = \sum_I Z_I \delta(\mathbf{r}-\mathbf{R}_I)$ is the nucleus density, in which Z_I and $\mathbf{R}_I$ are the atomic number and position vector of nucleus I, respectively. $\{\psi_i\}$ are the Kohn Sham orbitals. $m \doteq m(\mathbf{r})$ is the position dependent electron mass, $\hbar = \frac{h}{2\pi}$ with h be the planck constant. The meanings of the same notations are inherited from the previous energy functional G_1^{p}. $E_{XC}[n]$ is the exchange-correlation energy which is a function of the electron density that used to describe the many particles interactions in KSDFT scenario.

Remark 6.2.1. *The Kohn Sham orbital* $\{\psi_i\}$ *subject to the following orthonormal constraint*

$$< \psi_i|\psi_j >= \int \psi_i^*(\mathbf{r})\psi_j(\mathbf{r})d\mathbf{r} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{6.2.3}$$

where $\psi_i^*(\mathbf{r})$ *is the conjugate of* $\psi_i(\mathbf{r})$

Remark 6.2.2. *The following relation between Kohn Sham orbital and electron density holds:*

$$n(\mathbf{r}) = \sum_i |\psi_i|^2. \tag{6.2.4}$$

Remark 6.2.3. $\int_\Omega \chi \left[\hat{\rho}_{\text{total}}\phi\right] d\mathbf{r}$ *describes the electron-electron, electron-nucleus two body Columbic interactions. To avoid double counting, we neglect the two body Columbic interaction energy in the above energy functional that describes the KSDFT energy functional.*

So far, we have already formulated the polar energy functional for both multiscale models of the electrostatics interaction in the solute-solvent system. In the following parts, we will derive the governing equations of the two models by the variational principle.

6.2.2 Poisson model

In this subsection, we will derive the governing equation for the model with atomic modeling of the solute molecule and continuum modeling of the solvent model. Euler-Lagrange equation indicates that by taking $\frac{\delta G_1^{\mathrm{p}}}{\delta \phi} = 0$ yields the following Poisson equation:

$$\chi \left[\rho_{\mathrm{total}} + \nabla \cdot (\epsilon_{\mathrm{m}} \nabla \phi)\right] + (1-\chi)\left[\nabla \cdot (\epsilon_{\mathrm{s}} \nabla \phi)\right] = 0, \tag{6.2.5}$$

which can be written as:

$$-\nabla \cdot (\epsilon(\mathbf{r})\phi((r))) = \rho_{\mathrm{total}}, \tag{6.2.6}$$

where the dielectric function $\epsilon(\mathbf{r})$ has the following form:

$$\epsilon(\mathbf{r}) = \begin{cases} \epsilon_{\mathrm{m}}, & \mathbf{r} \in \Omega^{\mathrm{m}} \\ \epsilon_{\mathrm{s}}, & \mathbf{r} \in \Omega^{\mathrm{s}} \end{cases} \tag{6.2.7}$$

The above Poisson equation describes the electrostatics potential distribution in the solute-solvent system, in general, the solute and solvent domains take different dielectric constants, i.e., $\epsilon_{\mathrm{s}} \neq \epsilon_{\mathrm{m}}$.

6.2.3 Polarizable Poisson model

A more accurate physical description of the solute molecule is to model the solute molecule at the electronic level, in which the charge distribution of the solute molecule is calculated in an *Ab Initio* approach. In the water solvent environment, especially for the polar molecules, the interaction between solvent and solute molecules cannot be neglected, which can change the charge distribution of the solute molecule dramatically compare to that at the vacuum environment. The second model presented above can be used to include the solvent effects to the solute molecule charge distribution. In this part, we will derive the governing equations for the second model.

When taking the variational derivative of the polar free energy functional G_2^{p} with respect to the electrostatics potential $\phi(\mathbf{r})$, i.e., $\frac{\delta G_2^{\mathrm{p}}}{\delta \phi} = 0$, gives the following Poisson equation for describing electrostatics potential:

$$-\nabla \cdot (\epsilon(\mathbf{r}) \nabla \phi(\mathbf{r})) = \hat{\rho}_{\mathrm{total}}. \tag{6.2.8}$$

Due to the *Ab Initio* calculation of the charge distribution, the equation for describing the electronic structure is required, to obtain this, we take the variational derivative of the polar free energy functional G_2^{p} with respective to the Kohn Sham orbital subject to the orthonormal constraints of the orbital, gives

$$\frac{\delta \left[\Delta G_2^{\mathrm{p}} + \sum_{i,j} (\delta_{ij} - < \psi_i | \psi_j >) \right]}{\delta \psi_i} = 0,$$

which leads to

$$\chi \left[2q\phi \right] \psi - \chi \left[2 \frac{\hbar^2}{2m} \nabla \cdot (\nabla \psi_i) \right] + \chi \left[\frac{\delta E_{\mathrm{XC}}[n]}{\delta \psi_i} \right] - \chi \left[2 \sum_j \psi_j \lambda_{ji} \right] = 0,$$

which can be written as

$$\left[-\frac{\hbar^2}{2m} \Delta + q\phi + V_{\mathrm{XC}}[n] \right] \psi_i = \sum_j \psi_j \lambda_{ji}, \tag{6.2.9}$$

where $V_{\mathrm{XC}} \doteq \frac{\delta E_{\mathrm{XC}}[n]}{2\delta \psi_i}$.

Since

$$\lambda_{ij} =< \psi_i|\Lambda|\psi_j >=< \psi_j|-\frac{\hbar^2}{2m}\Delta+q\phi+V_{\mathrm{XC}}[n]|\psi_i >=< \psi_i|-\frac{\hbar^2}{2m}\Delta+q\phi+V_{\mathrm{XC}}[n]|\psi_j >^*= \lambda^*_{ji},$$

that is, $\Lambda \doteq \left(\lambda_{ij}\right)$ is Hermitian, hence there exist a unitary matrix U, such that

$$U^*\Lambda U = diag(E_1, E_2, \cdots),$$

consider the change of basis $\{\hat{\psi}_i\} = \{\psi_i\}U$, then Eq.(6.2.9) becomes

$$\left(-\frac{\hbar^2}{2m}\Delta + q\phi + V_{\mathrm{XC}}[n]\right)\sum_j \hat{\psi}_j\left(U^*\right)_{ji} = \sum_{j,k}\hat{\psi}_k\left(U^*\right)_{kj}\lambda_{ji},$$

multiply both sides by U_{il} and contract over index i, yields

$$\left(-\frac{\hbar^2}{2m}\Delta + q\phi + V_{\mathrm{XC}}[n]\right)\sum_{i,j} \hat{\psi}_j\left(U^*\right)_{ji}U_{il} = \sum_{i,j,k}\hat{\psi}_k\left(U^*\right)_{kj}\lambda_{ji}U_{il},$$

which can be simplified to

$$\left(-\frac{\hbar^2}{2m}\Delta + q\phi + V_{\mathrm{XC}}[n]\right)\hat{\psi}_l = E_l\hat{\psi}_l.$$

Note the unitary transformation does not change the energy of basis, for the sake of simplicity, we do not distinguish between two basis set, then we ends up with the following generalized Kohn Sham equation for the description of the Kohn Sham orbital:

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + q\phi + V_{XC}[n]\right)\psi_i = E_i\psi_i, \tag{6.2.10}$$

where $V_{XC}[n] \doteq \frac{\delta E_{XC}[n]}{2\psi_i \delta \psi_i} = \frac{dE_{XC}[n]}{dn}$.

The above generalized Kohn Sham equation can be further written as:

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + q\phi_{\text{RF}} + U_{\text{eff}}^0\right)\psi_i = E_i\psi_i, \tag{6.2.11}$$

where $\phi_{\text{RF}} = \phi - \phi_0$ is the reaction field potential of the solute in the solvent environment, in which ϕ and ϕ_0 are the electrostatic potential generated by the solute in the solvent and solute environments, respectively. And U_{eff}^0 is the effective Kohn Sham potential in the vacuum environment.

By denoting $U_{\text{eff}} \doteq q\phi_{\text{RF}} + U_{\text{eff}}^0$ as the effective potential of the generalized Kohn Sham equation, the generalized Kohn Sham equation Eq.(6.2.11) can be simplified as:

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + U_{\text{eff}}\right)\psi_i = E_i\psi_i. \tag{6.2.12}$$

Remark 6.2.4. *Compared to the KSDFT in vacuum, in generalized KSDFT, the reaction field energy in the solute-solvent system is added into the the Hamiltonian of KSDFT.*

6.2.4 Boundary and interface conditions

To make the aforementioned two models well-posed, both the boundary and interface conditions need to be specified.

In terms of the boundary condition, generally the following far field boundary condition is enforced for the equation of electrostatics potential:

$$\phi(\infty) = 0. \tag{6.2.13}$$

However, in the practical numerical implementation of the Poisson equation, the finite size solute-solvent domain Ω is required, in this case, the following Deybe-Hückel boundary conditions are enforced, respectively, for the electrostatics equation in two models.

- For the model with atomic modeling of solute molecule, we use the following boundary condition:
$$\phi(\mathbf{r}) = \sum_{i=1}^{N_m} \frac{q_i}{4\pi\epsilon_{\mathrm{s}}|\mathbf{r}-\mathbf{r}_i|},\ \forall \mathbf{r} \in \partial\Omega. \tag{6.2.14}$$

- For the model with quantum mechanics modeling of solute molecule, we use the following boundary condition:
$$\phi(\mathbf{r}) = \int_{\Omega} \frac{\hat{\rho}_{\mathrm{total}}(\tilde{\mathbf{r}})}{4\pi\epsilon_{\mathrm{s}}|\mathbf{r}-\tilde{\mathbf{r}}|} d\tilde{\mathbf{r}}\ \ \forall \mathbf{r} \in \partial\Omega. \tag{6.2.15}$$

Due to the fact that the solute and solvent domains admit different dielectric constants, there is an interface Γ, which is described by the molecular surface of the solute molecule, between two domains. Across this interface, the following conditions on the continuity of the electrostatics potential and electrostatics flux are enforced.

$$[\phi]|_{\Gamma} = \phi_{\mathrm{s}}(\mathbf{r}) - \phi_{\mathrm{m}}(\mathbf{r}) = 0,\ \forall \mathbf{r} \in \Gamma, \tag{6.2.16}$$

$$[\epsilon\phi_{\mathbf{n}}] = \epsilon_{\mathrm{s}}\phi_{\mathrm{s}}(\mathbf{r}) \cdot \mathbf{n} - \epsilon_{\mathrm{m}}\phi_{\mathrm{m}}(\mathbf{r}) \cdot \mathbf{n} = 0,\ \forall \mathbf{r} \in \Gamma, \tag{6.2.17}$$

where $[*]$ denotes the jump of the quantity $*$ across the interface Γ. ϕ_{s} and ϕ_{m} denote for the electrostatic potential limits at the interface point pointing from the solvent and solute domain, respectively. $\mathbf{n}$ is the outer normal direction on the interface pointing from solute domain to solvent domain.

6.2.5 Numerical Method

Mathematically, the polarizable Poisson model is formed by the following elliptic interface problem

$$\begin{cases} -\nabla \cdot (\epsilon(\mathbf{r})\phi(\mathbf{r})) = \tilde{\rho}_{\text{total}}, & \forall \mathbf{r} \in \Omega \\ \phi_{\text{s}}(\mathbf{r}) - \phi_{\text{m}}(\mathbf{r}) = 0, & \forall \mathbf{r} \in \Gamma \\ \epsilon_{\text{s}}\nabla\phi_{\text{s}}(\mathbf{r}) \cdot \mathbf{n} - \epsilon_{\text{m}}\nabla\phi_{\text{m}}(\mathbf{r}) \cdot \mathbf{n} = 0, & \forall \mathbf{r} \in \Gamma \\ \phi(\mathbf{r}) = \int_\Omega \frac{\hat{\rho}_{\text{total}}}{4\pi\epsilon_{\text{s}}|\mathbf{r}-\tilde{\mathbf{r}}|} d\tilde{\mathbf{r}}, & \forall \mathbf{r} \in \partial\Omega. \end{cases} \tag{6.2.18}$$

coupled with the following generalized KSDFT equation

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + U_{\text{eff}}\right)\psi_i = E_i\psi_i. \tag{6.2.19}$$

The elliptic interface problem is solved similar to the above without the Green's function treatment of the singular charge procedure, instead the charge is locally projected to the nearest eight grid points in the discretized computational solute solvent domain. The generalized KSDFT is solved by the SIESTA software with the reaction field energy added into the Kohn Sham effective potential. Due to the coupling of the Poisson model and the KSDFT, in which the electron density calculated by KSDFT is used for the charge source in the Poisson model, in turn, the electrostatic potential calculated from the Poisson model is further used by generalized KSDFT for updating the electronic structure. The coupled models need to be solved in an self-consistent approach, the pseudo-code for the self-consistent algorithm is demonstrated in Algorithm **??**. During the communication of the two solvers, the total charge conservation scheme is utilized, the scheme contains two steps:

- Project the charge density located on each mesh grid of SIESTA solver to the nearest 8 grids of the Poisson solver.

- Assemble the projected charges on the mesh grids of the Poisson solver.

Pseudopotential is used to eliminate the complicated effects of core electrons in SIESTA solver [224], for all calculations, the default double-ζ plus single polarization (DZP) basis are used. The MeshCutOff is set as 125 Rydberg and local density approximation (LDA) is used to approximate the exchange correlation potential. The solution method is set to be "diagon".

6.2.6 Reaction Field Energy Calculation

In this part, we will formulate the polar free energy, which is also called the reaction field energy referred in the previous parts. Due to the different numerical methods in two polar solvation models, the polar solvation free energy is calculated by different formulas.

- In the Poisson model, the reaction field potential can be written as $\phi_{\mathrm{RF}}(\mathbf{r}) = \phi^0(\mathbf{r}) + \tilde{\phi}(\mathbf{r})$, in turn, the polar free energy can be written as:

$$\Delta G_1^{\mathrm{p}} = \sum_{i=1}^{N_m} q(\mathbf{r}_i)\phi_{\mathrm{RF}}(\mathbf{r}_i). \tag{6.2.20}$$

- In the polarizable Poisson model, the reaction field potential can be written as $\phi_{\mathrm{RF}}(\mathbf{r}) = \phi^{\mathrm{inhomo}}(\mathbf{r}) - \phi^{\mathrm{homo}}(\mathbf{r})$, where $\phi^{\mathrm{inhomo}}(\mathbf{r})$ is solved from the Poisson model with the following dielectric function:

$$\epsilon(\mathbf{r}) = \begin{cases} \epsilon_{\mathrm{s}}, & \mathbf{r} \in \Omega^{\mathrm{s}} \\ \\ \epsilon_{\mathrm{m}}, & \mathbf{r} \in \Omega^{\mathrm{m}} \end{cases}$$

and the $\phi^{\mathrm{homo}}(\mathbf{r})$ is solved with the homogeneous permittivity function $\epsilon(\mathbf{r}) = \epsilon_{\mathrm{s}}, \ \forall \mathbf{r} \in \Omega$.

In the charge density formalism, the polar free energy can be written as:

$$\Delta G_2^{\mathrm{p}} = \int_{\Omega^{\mathrm{m}}} q(\mathbf{r})\phi_{\mathrm{RF}}(\mathbf{r})d\mathbf{r}. \tag{6.2.21}$$

For the sake of abbreviation, without ambiguity, we do not distinguish between ΔG_1^{p} and ΔG_2^{p}, and denote them unified as ΔG^{p}.

6.3 Nonpolar solvation models

In our earlier work, the nonpolar solvation free energy was modeled by [270, 45, 43, 44]

$$G^{\mathrm{NP}} = \gamma A + pV + +\rho_0 \int_{\Omega_s} U^{\mathrm{vdW}} \mathrm{d}\mathbf{r}, \tag{6.3.1}$$

where γ and A are, respectively, surface tension and area of the solute molecule. Additionally, p and V are, respectively surface enclosed volume and hydrodynamic pressure difference. Finally, ρ_0 is the solvent bulk density, and U^{vdW} is the van der Waals interaction potential, i.e., the Lennard-Jones potential. The integration is over solvent domain Ω_s. This nonpolar solvation free energy model was shown to offer excellent predictions of experimental data for various nonpolar molecules [46]. In our recent work, we have demonstrated superb results for a large number of polar and nonpolar molecules when the nonpolar model in Eq. (6.3.1) is combined with a PB model for the polar solvation component [248].

The Lennard Jones potential offers a physical model for dealing with vdW interactions near the solvent-solute interface. However, a drawback of this nonpolar term is that the probe radius in the Lennard Jones potential is nonlinear and cannot be optimized together with other nonpolar parameters. Mathematically, for a small probe radius, the vdW term for a given atom is proportional to the atomic surface area. Therefore, atomic surface area approach used by Kollman and co-workers [260] should have a similar modeling effect as that of the Lennard Jones potential. Based on this observation, we propose the following nonpolar solvation energy model

$$\Delta G^{\rm NP} = \sum_{\alpha=1}^{N} \gamma^{\alpha}{\rm Area}^{\alpha} + p{\rm Vol} = \sum_{\alpha=1}^{N}\sum_{i\in\alpha} \gamma^{\alpha}{\rm Area}_{i}^{\alpha} + p{\rm Vol}, \tag{6.3.2}$$

where N is the total number of atomic types in a given solute molecule. Here, γ^{α} and ${\rm Area}^{\alpha}$ are the surface tension and surface area of the αth type of atoms, respectively, and ${\rm Area}_{i}^{\alpha}$ is the surface area of the ith atom in the α type of atoms. In this nonpolar solvation free energy model, the parameters γ_{α} and p need to be learned from a training set. Both the Lennard Jones and the atomic-surface-area based nonpolar solvation models are validated in the present work.

6.3.1 Modeling of nonpolar solvation free energy-scoring functional groups

6.3.1.1 Functional group modeling

In our nonpolar solvation model, we assume that each functional group of molecules admits the same set of optimal parameters. Similar approach has been successfully used in the literature [220], including ours [46, 248]. In this work, we further incorporate machine learning type of methods for the nonpolar solvation free energy prediction. To this end, the whole data set, which contains 668 molecules, is partitioned into training sets and testing sets. In the leave one out test case, each step we only leave one molecule out as the test set, all the other molecules are regarded as the training set. In a specific leave-SAMPLx-out study, all the SAMPLx molecules are left as the testing set, while the remaining molecules are treated as the training set. Furthermore, the training set in this scenario is also divided into two parts: i) the mono-functional group molecules, in which molecules of a specific functional group is used to train a set of parameters for all similar molecules; and ii) the poly-functional groups molecules in which scoring weights are used to weight the contributions of parameters of various involved functional groups.

Table 6.1 lists twenty functional groups used in the training set. We denote $\{T_1, T_2, \cdots, T_n\}$ a given set of n molecules with a specific functional group from the above 20 groups. For a molecule indexed j, the solvation free energy calculated from the polar solvation free energy model, atomic surface areas and molecular volume are listed as:

$$\left\{\Delta G_j^{\mathrm{p}}, \mathrm{Area}_j^1, \mathrm{Area}_j^2, \cdots, \mathrm{Area}_j^N, \mathrm{Vol}\right\}, \tag{6.3.3}$$

Table 6.1: Functional groups and corresponding number of molecules used in the classification.

Group	Number	Group	Number
alkynyl	8	alkenyl	38
aldehyde group	11	nitrile group	5
carboxyl	7	ester group	34
ketone	23	amino	35
nitro	9	alcoholic hydroxy	33
phenolic hydroxyl	16	ether	22
alkane	38	aromatics	33
nitrogen heterocyclic	19	chlorinated hydrocarbon	53
Nitrate	5	amid	7
thiol	4	thioether	5

where $j = 1, 2, \cdots, n$.

In our approach, the solvation free energy is modeled as the sum of the polar and nonpolar parts, thus for the jth molecule, the corresponding modeled solvation free energy can be expressed as:

$$\Delta G_j = \Delta G_j^{\rm P} + \sum_{\alpha=1}^{N} \gamma^\alpha {\rm Area}^\alpha + p{\rm Vol} \doteq G_j(\mathbf{P}), \tag{6.3.4}$$

where $\mathbf{P} \doteq \{\gamma^1, \gamma^2, \cdots, \gamma^N, p\}$ is the set of parameters to be optimized.

For a given set of molecules with the same mono-functional group, we denoted $\mathbf{\Delta G}(\mathbf{P}) \doteq \{G_1(\mathbf{P}), G_2(\mathbf{P}), \cdots, G_n(\mathbf{P})\}$, and the associated experimental solvation free energy is denoted as $\Delta G^{\rm Exp} \doteq \left\{\Delta G_1^{\rm Exp}, \Delta G_2^{\rm Exp}, \cdots, \Delta G_n^{\rm Exp}\right\}$. Then the optimal parameter set can be obtained by solving the following Tikhonov regularized least square problem (also known as ridge regression) which has a closed form solution:

$$\text{argmin}_{\mathbf{P}} \left\{||\Delta G(\mathbf{P}) - \Delta G^{\rm Exp}||_2 + \lambda ||\mathbf{P}||_2\right\}, \tag{6.3.5}$$

where $|| * ||_2$ is the L_2 norm of the quantity $*$. Here λ is the regularization parameter chosen to be a large number, such as 100, in the present work to ensure the dominance of the first term and avoid over-fitting through controlling the magnitude of $||\mathbf{P}||_2$..

6.3.1.2 Scoring the functional groups

Most molecules in the SAMPL blind test sets involve poly-functional groups. In this case, we further employ the poly-functional group molecules in the training set for training the optimal relative scoring weights between different functional groups. According to the relative scoring weights, the scoring weights between all the functional groups can be obtained through a simple normalization procedure. We denote $\left\{\tilde{T}_1, \tilde{T}_2, \cdots, \tilde{T}_{n'}\right\}$ a given set of poly-functional group molecules that has the same functional groups in the training set. The associated optimized parameter sets are $\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_m$, where m is the number of functional groups in this set, each $\mathbf{P}_i$, $i = 1, 2, \cdots, m$ is learned through solving the regularized optimization problem given by Eq. (6.3.5). For the jth molecule in this poly-functional group set, we model its solvation free energy as:

$$\Delta\bar{G}_j(\boldsymbol{\omega}) = \sum_{i=1}^{m} \omega_i \Delta G_j(\mathbf{P}_i), \tag{6.3.6}$$

where $||\boldsymbol{\omega}||_1 \doteq \sum_{i=1}^{m} \omega_i = 1$, with ω_i being the scoring weight of ith functional group.

The relative scoring weights for the m functional groups associated to this set of poly-functional groups molecules can be learned via solving the following constraint optimization problem,

$$\text{argmin}_{\boldsymbol{\omega}} ||\Delta\tilde{G}(\boldsymbol{\omega}) - \Delta\tilde{G}^{\text{Exp}}||_2, \tag{6.3.7}$$

subject to

$$||\boldsymbol{\omega}||_1 = 1, \tag{6.3.8}$$

and

$$\omega_i \geq 0,\ \forall i = 1, 2, \cdots, m, \tag{6.3.9}$$

where $\Delta\tilde{G}(\boldsymbol{\omega})$ and $\Delta\tilde{G}^{\mathrm{Exp}}$ represent, respectively, the predicted and experimental solvation free energies for this group of poly-functional group molecules. Since both the optimization object given by Eq.(6.3.7) and the constraint conditions in Eqs. (6.3.8-6.3.9) are convex with respect to the scoring weights $\boldsymbol{\omega}$. The above constrained optimization can be easily solved via a convex optimization solver in the CVX software package [103, 102].

In the rest of this section, we provide an example to illustrate the procedure of functional group based approach for solvation free energy prediction. As shown in Fig. 6.1, the target molecule 2-Chlorosyringaldehyde (Pubchem ID: 53479) contains four different functional groups. We need to find scoring weights for these functional groups. Note that in the above molecular searching scheme, pairwisely, the relative weights for each two functional groups can be determined by solving the constrained optimization problem in Eqs. (6.3.7)-(6.3.8) for molecules in the two corresponding functional groups. According to the pairwise relative weights, the functional group scoring in the target molecule can be achieved.

6.3.2 Modeling of nonpolar solvation free energy - nearest neighbor approach

The above functional group based model is tested to be able to provide very accurate solvation free energy prediction provided that for each test molecule, parameters for all of the involved mono-functional groups can be determined, and a set of molecules with the same ploy-functional groups can be found in the training set as well. However, for some complex molecules that contain functional groups beyond those listed in Table 6.1, this method fails.

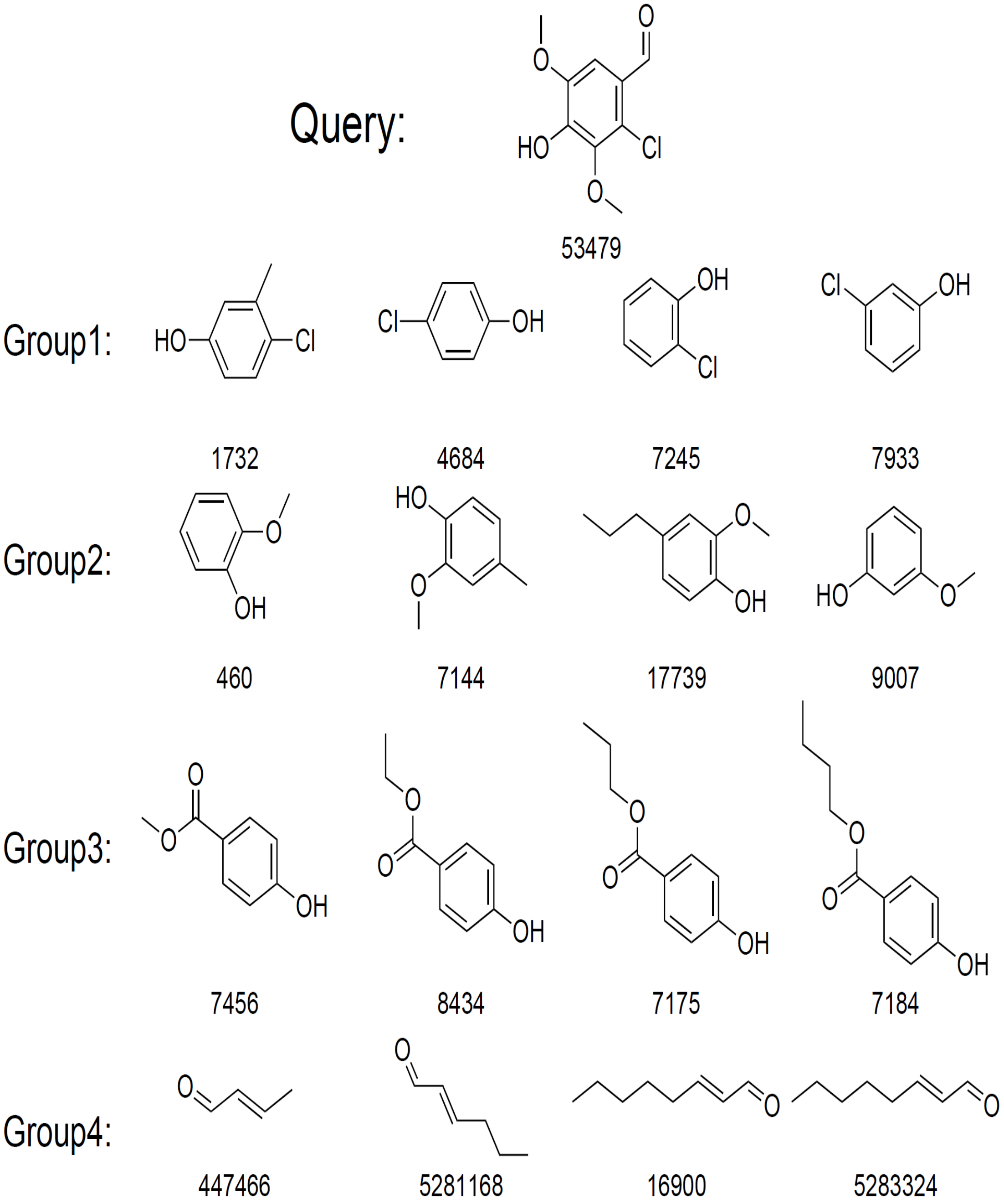


Figure 6.1: An illustration of the functional group scoring method for the prediction of the solvation free energy for molecule 2-Chlorosyringaldehyde (Pubchem ID: 53479), which contains four functional groups: aldehyde group, phenolic hydroxyl, ether and chlorinated hydrocarbon. We first compute relative weights between phenolic hydroxyl and chlorinated hydrocarbon; phenolic hydroxyl and ether; phenolic hydroxyl and ester group; and ester group and aldehyde group. Then relative weights are combined to generate the full set of weights $\omega_1, \omega_2, \omega_3$ and ω_4 for solvation free energy prediction.

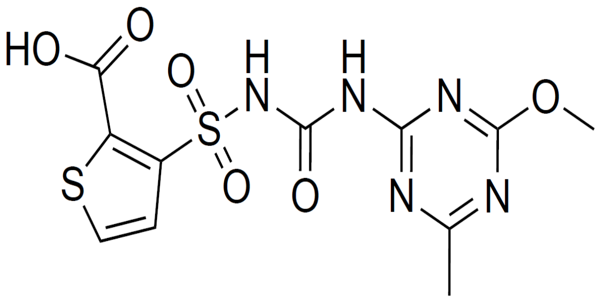


Figure 6.2: Thifensulfuron (Pubchem ID: 91729).

For instance, considering a SAMPL1 test molecule, thifensulfuron, as shown in Fig. 6.2, it has a very complex structure and contains functional groups cannot be found in Table 6.1. We therefore propose the following ranking algorithm for nearest neighbor searching.

6.3.2.1 Atomic feature based molecular ranking

To deal with general and complex molecules, we propose another approach to rank molecules, and develop a nearest neighbor approach for nonpolar solvation free energy prediction. Our molecular ranking is based on atomic features. Our basic assumption is that if two solute molecules have similar atomic features, their chemical properties are also similar. Therefore, these molecules should share the same parameter set for the nonpolar solvation free energy model. This approach is further described below.

Our method is based on the ansatz that molecules chemical properties are mainly determined by atomic features, including structural features and atomic electrostatic features. Among them, atomic structural features include atomic type, atom hybridization state and bonding information. Atomic electrostatic features include atomic charge, atomic dipole, atomic quadrupole and atomic electrostatic solvation free energy. Atomic features are selected based the criteria that a given atomic feature is retained if it can effectively discriminate the previously mentioned mono-functional groups listed in Table 6.1. Table 7.1 lists all the atomic features that are used for molecule ranking.

Table 6.2: Atomic features used for ranking molecules.

Feature name
Number of atoms
Number of heavy atoms
Number of hydrogen atoms
Number of single bonds
Number of double bonds
Number of triple bonds
Number of aromatic bonds
Number of each type of atoms
Number of sp^1 carbon atoms
Number of sp^2 carbon atoms
Number of sp^3 carbon atoms
Number of sp^1 nitrogen atoms
Number of sp^2 nitrogen atoms
Number of sp^3 nitrogen atoms
Number of sp^2 oxygen atoms
Number of sp^3 oxygen atoms
Number of sp^2 sulfur atoms
Number of sp^3 sulfur atoms
Maximum of atomic reaction field energy
Minimum of atomic reaction field energy
Maximum reaction field energy of each type of atoms
Minimum reaction field energy of each type of atoms
Maximum of atomic reaction field energy
Average reaction field energy of each type of atoms
Total absolute charge
Total charge of each type of atoms
Total absolute charge of each type of atoms
Maximum charge of each type of atoms
Minimum charge of each type of atoms
The variation of each type of atomic charges
Maximum of atomic dipole
Maximum of each atomic dipole
Minimum of each atomic dipole
Variation of atomic dipole
Maximum quadrupole
Maximum quadrupole of each type of atoms
Variation of atomic quadrupole
Variation of each type of atom's quadrupole

We also construct atomic features by using the statistics variables, i.e., average, maximum, minimum, and variation of quantities in Table 7.1. Finally, we rule out redundant features, where redundancy is due to the fact that some atomic features are 100 % correlated with each other. In this case, only one of these highly correlated features is retained.

Atomic features are calculated by following methods.

- Molecular structural information is parsed by the Open Babel software [169].
- Atomic charge, atomic dipole and atomic quadrupole are obtained via the Distributed Multipole Analysis (DMA) method [227], in which the charge density is originally computed by the density function theory with B3LYP and 6-31G basis in the Gaussian quantum chemistry software [80, 18, 148].
- Atomic electrostatic solavtion energy is calculated by our in-house MIBPB software [253, 39, 90].

With the above selected atomic features, intuitively, we measure the similarity of molecules by the Pearson correlation coefficient of atomic feature vectors. Specifically, we denote the features of two molecules as vector $X \doteq \{x_1, x_2, \cdots, x_k\}$ and vector $Y \doteq \{y_1, y_2, \cdots, y_k\}$, then their similarity is measured by

$$C_{XY} = \frac{|\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y})|}{\sqrt{\sum_{i=1}^{k}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{k}(y_i - \bar{y})^2}},$$

where $\bar{x} \doteq \frac{1}{k}\sum_{i=1}^{k} x_i$ and $\bar{y} \doteq \frac{1}{k}\sum_{i=1}^{k} y_i$, and k is the dimension of the feature space. The higher the correlation between two atomic feature vectors indicates the more similarity between molecules.

6.3.2.2 Nearest neighbor solvation free energy prediction

By using the above nearest neighbor based nearest molecule searching procedure, for a given molecule, we obtain the similarity ranking of the remaining molecules with this specified one, as well as correlations between molecules. In the nearest neighbor prediction of the nonpolar solvation free energy, we learn the parameters in the nonpolar solvation free energy expression, Eq. (6.3.2), by a given number of nearest molecules found in the training set. The number of the nearest molecules used is based on the following principles

- All the molecules whose correlation with a given molecule is greater than 0.99 are used for the parameter learning for this given molecule;

- If the above criterion yields less than 5 molecules, we use 5 nearest molecules for the parameter learning. Here the 5-molecule nearest neighbor method is found to provide the best leave-one-out prediction.

6.3.3 Unified protocol for blind solvation free energy predictions

We utilize a unified protocol for the prediction of solvation free energy. First, for the polar solvation free energy $\Delta G^{\rm p}$, we select either the Poisson model with a given point charge force field or the polarizable Poisson model. Second, for the calculation of the nonpolar solvation free energy we note that in general, the functional group scoring approach can deliver better blind predictions than the nearest neighbor approach. Therefore, in our nonpolar energy prediction step, the functional group scoring method is used whenever it works. Otherwise, we utilize the nearest neighbor approach.

6.4 Results and discussions

6.4.1 Date processing and model validation

6.4.1.1 Data sets and force fields

We consider a total of 668 molecules, the structures of these molecules are downloaded from the Pubchem project. The experimental solvation free energies of these molecules are collected from the literature. This dataset contains molecules from the SAMPL blind solvation prediction projects, ranging from SAMPL0 to SAMPL4 [184, 105, 86, 87, 106]; and the remaining molecules in our data set are collected from the literature [30, 261, 153]. Coincidentally, there is a considerable overlap of our database with Mobley's solvation database, which is available from http://mobleylab.org/resources.html. More precisely, 589 molecules in our database are already covered by Mobley's. The detail information of our dataset is provided in the supporting material.

For both the standard Poisson model and the KSDFT based polarizable Poisson model, we consider four types of atomic radii, namely, Amber 6, Amber bondi, Amber mbondi2 [35] and ZAP9 [183] parametrizations. Additionally, for the standard Poisson model, three sets of charge assignments, namely, OpenEye-AM1-BCC v1 parameters [123], Gasteiger [85], and Mulliken [35], are tested. Our MIBPB solver [253, 39, 90] is utilized to solve the Poisson interface problem to obtain electrostatic solvation energies. The probe radius is set to 1.4Å for the ESES surface generation in all PB calculations.

For the KSDFT based polarizable Poisson model, the Poisson interface problem is coupled with the generalized KSDFT in a self-consistent approach. The charge density used by Poisson interface problem is calculated in *Ab Initio* approach by the generalized KSDFT,

Table 6.3: Molecules in SAMPLx sets involving bromine and/or iodine atoms.

Test set	Molecule
SAMPL0	benzyl bromide
SAMPL1	bromacil
SAMPL2	5-bromouracil 5-iodouracil
SAMPL3	None
SAMPL4	None

here generalization due to the additional solute-solvent reaction field energy is included in the effective KS potential. The Poisson interface problem is employed for calculating the reaction field energy. This approach is modified from our earlier differential geometry based polarizable PB model [45]. We iteratively couple the SIESTA [224] and the MIBPB solver [253, 39, 90], in which a sharp solvent-solute interface is employed. The SIESTA software with additional reaction field energy is used for charge density calculation, and our in-house software is used for reaction field energy calculation. A uniform mesh size of 0.25 Å is used for solving the Poisson equation in both the standard Poisson and Polarizable Poisson model. In the polarizable Poisson model, the computed change densities are mapped to the uniform mesh with the conservation of the total charge.

For the molecules containing iodine atoms, the current level of DFT method used in this work, including the Gaussian software, cannot handle this element appropriately. Therefore, for a uniform comparison, we ignore molecules containing iodine atoms. There is a similar situation for bromine atoms — there is no appropriate pseudo-potential for this atom in the SIESTA software. Therefore, molecules involving bromine atoms are excluded in our KSDFT based polarizable PB calculations. Table 6.3 lists four molecules in SAMPLx that are not considered in some of our predictions.

6.4.1.2 Atomic surface area and molecular volume calculation

In our nonpolar solvation free energy model, both atomic areas and surface enclosed volume are required. In a recently developed Eulerian solvent excluded surface (ESES) package, a second-order accurate numerical scheme for surface area calculations and a third-order accurate numerical scheme for volume estimations have been developed [157]. A weighted Voronoi diagram algorithm is implemented to partition a molecular surface area into atomic surface areas. These schemes have been intensively validated by a large number of test examples. In this work, both atomic areas and molecular volume are calculated directly by using our ESES software package [157].

6.4.1.3 Validation of atomic surface area based nonpolar model

In this work, instead of using the classical nonpolar model that includes the van der Waals interaction between the solvent and solute nonpolar interaction [248] as shwon in Eq. (6.3.1), we utilize the atomic surface area model given in Eq. (6.3.2) for the corresponding effects. In this part, we consider a set of numerical test to verify the validity of this treatment. Our numerical results indicate that the atomic surface area model provides results as good as those obtained by the van der Waals based nonpolar model.

The van der Waals interaction Eq. (6.3.1) is modeled as a semi-continuous and semi-atomic potential. More specifically, we consider the 6-12 Lennard potential for modeling the van der Waals interaction between the continuum solvent and atomistic solute molecules [248]. The solvent radii in the Lennaed potential is set to be 1.4 Å, which is the same as that used in all other calculations. Table 6.4 lists the RMSEs with four types of atomic radii and four charge parametrization methods for the prediction of SAMPL0's solvation free energies. From these results, it is seen that for a given force field parametrization, the

Table 6.4: The RMSEs of the solvation free energy prediction with atomic surface area and van der Waals interaction models of nonpolar solvation free energy for SAMPL0 test set. The molecule in the SAMPL0 set that contains Br atom is excluded from this comparison. All results are in unit kcal/mol

Nonpolar model	Radius	BCC	Mulliken	Gasteiger	SIESTA
Atomic surface area	Amber 6	1.30	1.27	1.23	0.99
	Amber Bondi	1.40	1.30	1.30	0.93
	Amber Mbondi2	1.41	1.34	1.33	1.10
	ZAP9	1.33	1.39	1.37	1.08
van der Waals	Amber 6	1.42	1.31	1.34	0.93
	Amber Bondi	1.44	1.30	1.30	1.05
	Amber Mbondi2	1.45	1.38	1.30	1.18
	ZAP9	1.44	1.39	1.32	1.32

Table 6.5: The RMSEs of the leave-one-out test of the solvation free energy prediction with different methods, all with unit kcal/mol

Radius	BCC	Mulliken	Gasteiger	SIESTA
Amber 6	1.47	1.49	1.65	1.65
Amber Bondi	1.34	1.48	1.66	1.40
Amber Mbondi2	1.33	1.49	1.68	1.42
ZAP9	1.33	1.39	1.70	1.53

atomic surface area approach provides slightly more accurate nonpolar solvation free energy prediction. As shown in our earlier work [248], the performance of the semi-continuous and semi-atomic Lennard Jones potential is sensitive to the choice of the probe radius, due to its nonlinear dependence on the radius, whereas the SES based atomic surface area approach is less sensitive to the probe radius. We conclude that in general the atomic surface area based nonpolar model is at least as good as the Lennard Jones potential based one for modeling the nonpolar solvation free energy. Therefore, the atomic surface area based nonpolar model is employed in the rest of this chapter.

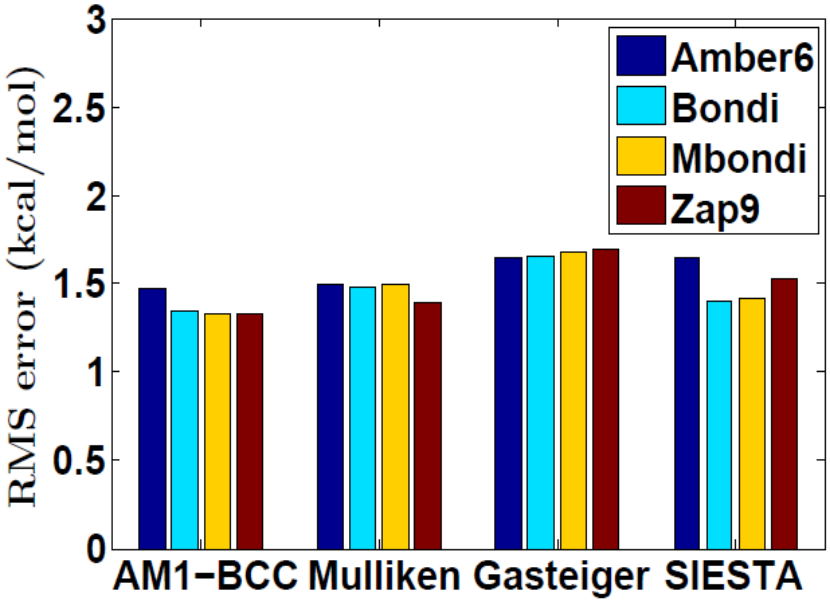


Figure 6.3: The leave-one-out error of the whole training set with 16 different charge and radius parametrizations. The nearest neighbor approach is employed for solvation free energy prediction.

6.4.2 Solvation predictions

6.4.2.1 Leave-one-out prediction

We first examine the proposed nearest neighbor approach by the leave-one-out test of 668 molecules. In this examination, we select one molecule at a time and use all other molecules as the training set to predict the selected one's solvation free energy. This process is applied systematically to all the molecules in the whole dataset of 668 molecules. Four different atomic radius sets are considered, together with four different charge force fields. Our results are illustrated in Fig. 6.3, the associated values are listed in Table 7.4. In general, all radius sets and charge force fields perform similarly well. The maximum RMSE is below 1.7 kcal/mol for all methods over all molecules. More specifically, Bondi and Mbondi radii offer the best overall results. For charge force field AM1-BBC charges appear to provide the best predictions and their RMSEs are less than 1.5kcal/mol for all the four radius sets.

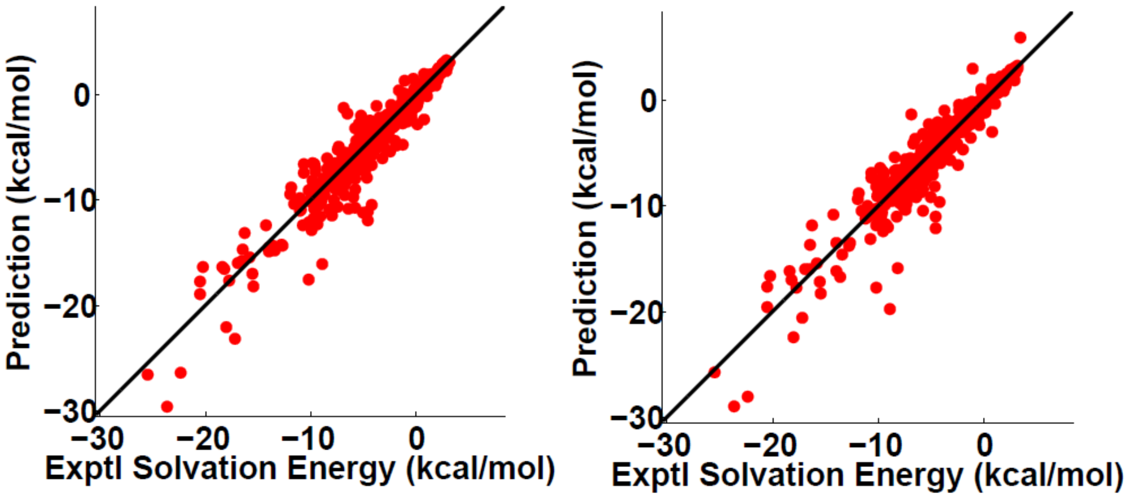


Figure 6.4: The plot of optimal leave one out test results, where the optimal prediction achieves when BCC charge is used in conjunction with either the Amber MBondi2 or the ZAP9 radius. In both these cases, the prediction with RMSE 1.33 kcal/mol. The correlation between prediction and experimental solvation free energies are 0.956 and 0.955, respectively, for the Amber MBondi2 and ZAP9 force fields. The corresponding R^2 are 0.913 and 0.911, respectively.

The optimal result obtained with AM1-BBC charges and ZAP9 radii has an RMSE of 1.33 kcal/mol. Figure 6.4 plots the optimal results on the leave one out test.

For a set of 643 molecules, which largely overlaps with present dataset, Mobley and Guthrie reported an RMSE of 1.51 kcal/mol [175]. Our results indicate that the present nearest neighbor approach can achieve highly accurate predictions for the solvation free energies for all the atomic radius and charge force fields.

6.4.2.2 SAMPLx blind predictions

In this section, we present results of the blind solvation free energy predictions based on the proposed protocol. All the SAMPL0-SAMPL4 challenges for solvation free energies are considered. We predict the solvation free energies with a leave-SAMPLx-out approach, in which the SAMPLx molecules and their experimental solvation free energies are regarded

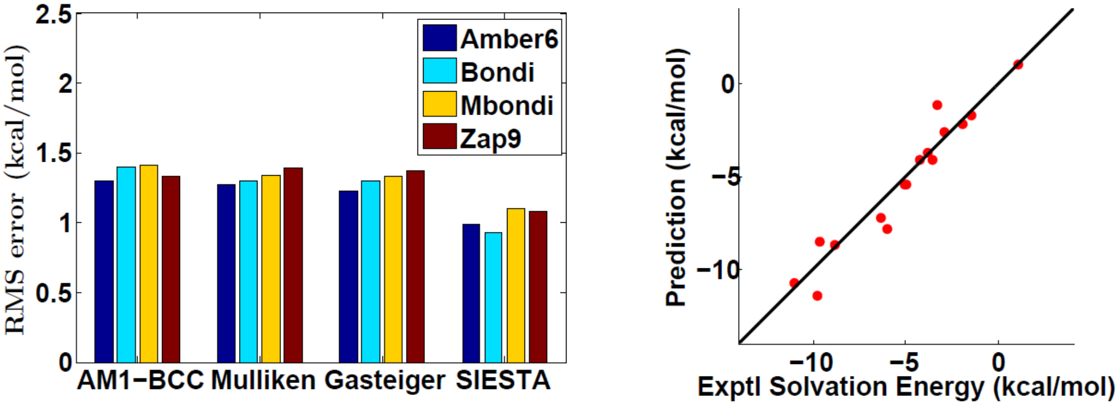


Figure 6.5: The prediction results for the SAMPL0 blind test set, the left chart shows the RMSEs between the experimental and prediction solvation free energies. The right chart shows the optimal predictions of the solvation free energies for the SAMPL0 test set.

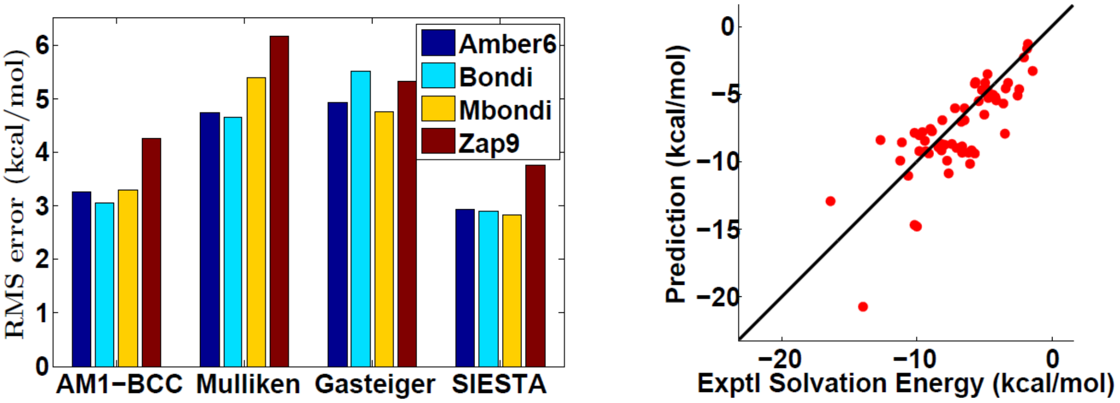


Figure 6.6: The prediction results for the SAMPL1 blind test set, the left chart shows the RMSEs between the experimental and prediction solvation free energies. The right chart shows the optimal predictions of the solvation free energies for the SAMPL1 test set.

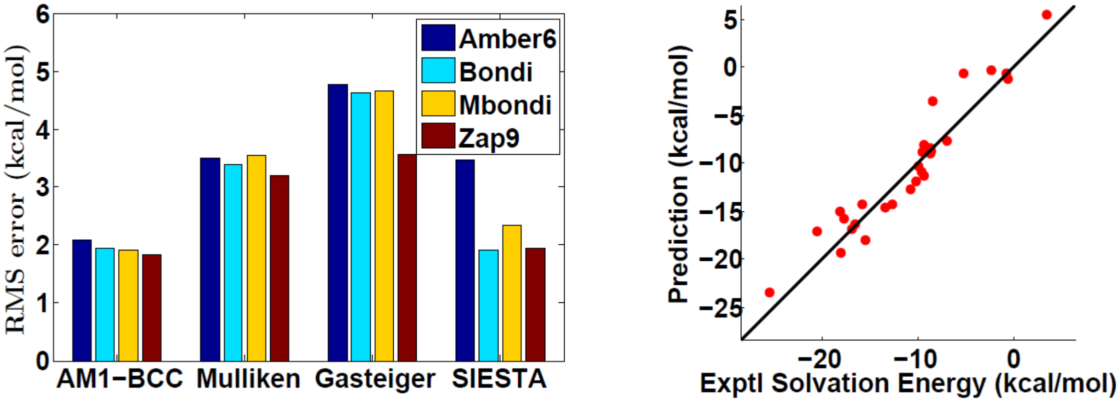


Figure 6.7: The prediction results for the SAMPL2 blind test set, the left chart shows the RMSEs between the experimental and prediction solvation free energies. The right chart shows the optimal predictions of the solvation free energies for the SAMPL2 test set.

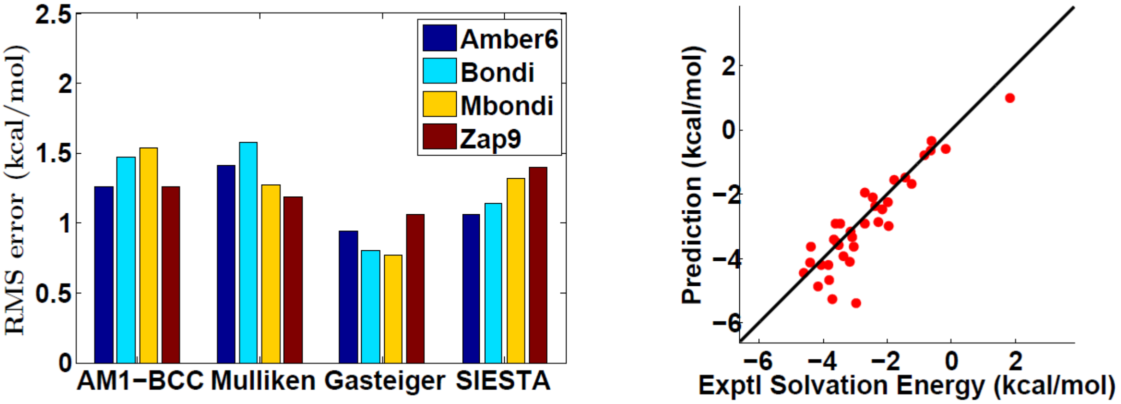


Figure 6.8: The prediction results for the SAMPL3 blind test set, the left chart shows the RMSEs between the experimental and prediction solvation free energies. The right chart shows the optimal predictions of the solvation free energies for the SAMPL3 test set.

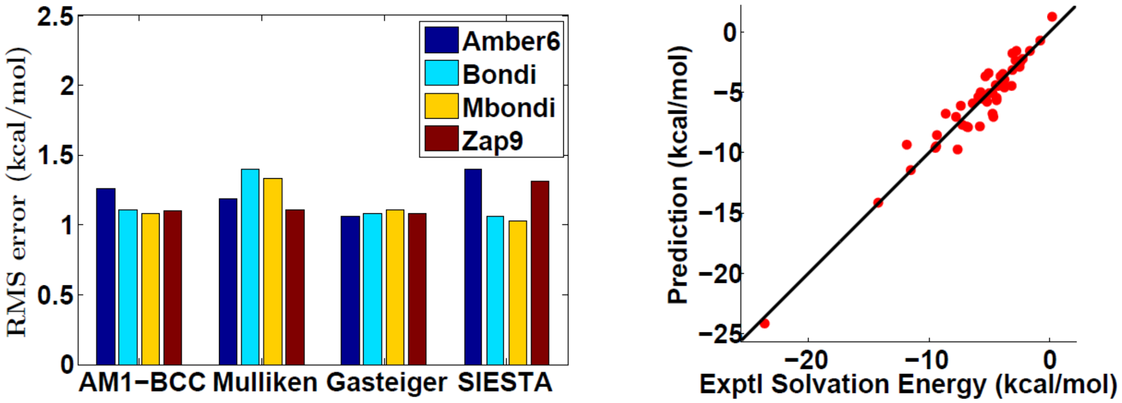


Figure 6.9: The prediction results for the SAMPL4 blind test set, the left chart shows the RMSEs between the experimental and prediction solvation free energies. The right chart shows the optimal predictions of the solvation free energies for the SAMPL4 test set.

as unknown while information of other molecules is utilized to predict selected SAMPL test set, based on molecular formulas.

We first consider SAMPL0 test set, in which molcules are diverse. One molecule in this test set contains Br atom, for which our polarizable PB model does not an appropriate pseudo-potential. The structures of the SAMPL0 molecules can be found in the literature [183]. The left chart of Fig. 6.5 shows the plot of RMSEs for each charge and atomic radius combination. It is clear that the change in atomic radii has a minor influence on the accuracy of predictions, while the change in charge force fields has a much more significant influence. In other words, the solvation free energy prediction for this test set is more sensitive to the charge parametrization than the atomic radii parametrization. In general, our KSDFT based polarizable Poisson model provides better predictions than those of other charge force fields. It is noted that the SIASTA based polarizable Poisson model with Amber Bondi radius parameters delivers the best solvation free energy prediction. This result is depicted in the right chart of Fig. 6.5. The associated RMSE (0.93 kcal/mol) appears to be better than that in the literature [183] (i.e., 1.71 kcal/mol for the *full* SAMPL0 test set of 17 molecules).

Table 6.6: The RMSEs of the solvation free energy prediction with different methods. The RMSEs inside and outside the parenthesis denote for the prediction errors include and exclude the molecules contains Br atom. All with unit kcal/mol

Test set	Radius	BCC	Mulliken	Gasteiger	SIESTA
SAMPL0	Amber 6	1.30 (1.26)	1.27 (1.25)	1.23 (**1.20**)	0.99 (NA)
	Amber Bondi	1.40 (1.37)	1.30 (1.27)	1.30 (1.27)	**0.93** (NA)
	Amber Mbondi2	1.41 (1.37)	1.34 (1.32)	1.33 (1.29)	1.10 (NA)
	ZAP9	1.33 (1.29)	1.39 (1.37)	1.37 (1.33)	1.08 (NA)
SAMPL1	Amber 6	3.26 (3.27)	4.74 (4.77)	4.92 (4.96)	2.92 (NA)
	Amber Bondi	3.06 (**3.07**)	4.65 (4.68)	5.52 (5.55)	2.89 (NA)
	Amber Mbondi2	3.29 (3.30)	5.39 (5.41)	4.76 (4.82)	**2.82** (NA)
	ZAP9	4.26 (4.35)	6.16 (6.16)	5.33 (5.45)	3.76 (NA)
SAMPL2	Amber 6	2.09 (2.11)	3.51 (3.59)	4.78 (4.86)	3.46 (NA)
	Amber Bondi	1.95 (1.97)	3.38 (3.47)	4.62 (4.72)	**1.90** (NA)
	Amber Mbondi2	**1.90** (**1.96**)	3.55 (3.66)	4.65 (4.76)	2.35 (NA)
	ZAP9	2.05 (2.03)	3.19 (3.15)	4.56 (4.51)	1.93 (NA)
SAMPL3	Amber 6	1.28	1.42	0.97	1.08
	Amber Bondi	1.47	1.58	0.82	1.16
	Amber Mbondi2	1.47	1.58	0.82	1.16
	ZAP9	1.55	1.28	**0.78**	1.33
SAMPL4	Amber 6	1.28	1.20	1.08	1.41
	Amber Bondi	1.12	1.41	1.10	1.07
	Amber Mbondi2	1.09	1.33	**1.03**	1.04
	ZAP9	1.12	1.12	1.09	1.32

The best prediction in the literature has an RMSE of 1.34 kcal/mol [134]. Even though for the SIESTA based *Ab Initio* charge calculation, the molecule contains the Br atom is neglected, based on the general pattern that when any other force field used, the prediction for that omitted molecules is very well. We believe with proper pseudopotential for Br atom, our polarizable Poisson model can deliver an accurate prediction for this molecule also. Moreover, for this test set, all the other three charge parametrizations can provide a similar level of solvation free energy prediction.

We next consider the SAMPL1 challenge set for solvation predictions [105]. This set contains not only a largest number of 63 molecules, which is the largest among all the SAMPL test sets, but also many molecules with extremely complicated structures. The detail description of this set is given in the literature [105]. Most molecules in this set are druggable and very complex. The difficulty of this test set comes from two aspects: on the one hand, the structures of molecules in this set are very complicated. On the other hand, the knowledge in the training set that can be used for predicting the solvation free energies of these molecules is rare or insufficient. In addition to the molecular complexity, the reported experimental solvation free energies also admit large uncertainty [105]. Most earlier computational predictions report RMSEs of 3 to 4 kcal/mol when some extremely complex molecules are excluded. Some best prediction for the whole set has an RMSE of 2.45 kcal/mol [134]. The best performance was shown to give an RMSE of 2.4 kcal/mol on a subset of the SAMPL1 test set that contains only 56 molecules [165].

Figure 6.6 plots the present blind predictions. One molecule (bromacil) that contains a Br atom is considered by all the charge models except for the polarizable Poisson method. It is noted that two of present approaches, namely, the AM1-BCC semi-empirical charges and the *Ab Initio* charge provide relatively accurate predictions for this test set. When Amber Mbondi2 atomic radii together with the SIESTA charge calculation applied, the optimal solvation free energy prediction is achieved with RMSE 2.82 kcal/mol. When BCC charges and the Amber Bondi radii are used, the prediction for the whole SAMPL1 set without a molecule that contains Br atom has an RMSE of 3.06 kcal/mol, the RMSE is 3.07 kcal/mol for the same test with all molecules involved. For this test set, the large prediction RMSE mainly due to the extremely large error in predicting solvation free energies for some complex molecules, for which the RMSE can be as large as 15 kcal/mol. This unreasonable prediction is due to the fact that inappropriate molecule force field parametrization yields unreasonable electrostatic solvation free energies, which in turn leads to erroneous solvation free energy prediction. Similar to the SAMPL0 test set, the prediction is more sensitive to the charge parametrization. Nevertheless, the atomic radius parametrization also plays an critical role in the prediction accuracy for this test set.

SAMPL2 is another difficult test set with almost the same level of difficulty as the SAMPL1 test set [137]. Compared with SAMPL1 test set it contains a few complex molecules, and most molecules in this test set are drug like ones as well. Contrary to molecules in SAMPL1 test set, this set has less uncertainty in the experimental solvation free energies. The experimental solvation free energies of this test set distribute over a wide range. Using all-atom molecular dynamics simulations and multiple starting conformations for blind prediction, Klimovich and Mobley reported an RMSE of 2.82 kcal/mol over the whole set and 1.86 kcal/mol over all the molecules except several hydroxyl-rich compounds [137]. Some best prediction has an RMSE of 1.59 kcal/mol [134]. In the present work, the molecule containing an I atom (5-iodouracil) is excluded in all calculations. Additionally, 5-bromouracil has a Br atom is excluded in the polarizable Poisson model. The RMSEs from various radius and charge force fields are given Fig. 6.7. Apparently, this test set has a strong force field dependence as well. The RMSEs vary very much from one charge force field to another. However, the performance of these predictions has a weak radius dependence. The best prediction, obtained from a combination of Amber MBondi2 radius parameters and BCC charge force field, or Amber Bondi radius together with modified SIESTA charge, both have the RMSE of 1.90 kcal/mol when the molecule with a Br atom is excluded in the prediction. When all molecules are included, the combination of Amber MBondi2 radius and BCC charge parametrization gives the optimal prediction of RMSE 1.96 kcal/mol as depicted in the right chart of Fig. 6.7. Compared to the prediction of SAMPL1 test set, these predictions are more accurate, due to two reasons. First, the molecules in this set are slightly simpler and experimental uncertainly is less severe for the deterministic prediction. Second, in our knowledge based models for the nonpolar solvation free energy prediction, we have more similar molecules in training set, which enables us to obtain better nonpolar solvation free energy parameters in the nearest neighbor based approach. In contrast, in the SAMPL1 prediction, when the nearest neighbor approach is applied, the nearest molecules selected from the training set differ much from SAMPL1 molecules.

SAMPL3 test set is a relatively easy one with 36 solute molecules [87]. Its molecular structures are less versatile than the earlier test sets. Additionally, in this test set, there is no molecule that involves Br or I atom. One of difficulties in the prediction of this set is the lack of similar molecules in our database when all the SAMPL3 molecules are left out. The best prediction in the literature offers an RMSE of 1.29 kcal/mol [134]. The RMSEs of our blind predictions for the SAMPL3 test set are plotted in Fig. 6.8. In this case, the accuracy of predictions also depends strongly on charge force fields and weakly on radius parameters. In general, Gasteiger charges are superb for this test set and their RMSEs are always smaller than 1 kcal/mol. Our best prediction, obtained from the combination of ZAP9 radius parameters and Gasteiger charge force field, has a small RMSE of 0.78 kcal/mol and is depicted in the right chart of Fig. 6.8. In general Gasteiger charge performs very poor in the solvation free energy prediction for previous tests on complex molecules, while we see that this charge parametrization method is superior for chlorinated hydrocarbon molecules in the present test set. Unfortunately, there is no uniformly optimal parametrization for all the molecules. This fact motivates us to seek a solvation free energy prediction that does not heavily depends on the force fields parametrization.

SAMPL4 test set is studied by using the proposed methods as well. A key feature of this test set is that its molecules are diversified. This test set also involves a wide range of solvation free energies. However, the structures of these molecules are not as complex as those in both SAMPL1 and SAMPL2 test sets, which indicates a slightly easier task for solvation free energy prediction. This test set was studied by a number of researchers in the literature and the best prediction in the literature has the RMSE of 1.2 kcal/mol [177]. Figure 6.9 illustrates the RMSEs of our predictions for a total of 16 charge and radius combinations. Compared with those in the literature, all of our predictions are of high quality. Our best result has the RMSE of 1.03 kcal/mol and the corresponding results are given in the right chart of Fig. 6.9.

Table 6.6 lists the RMSEs of our blind predictions of solvation free energies for all five test sets using a total of 16 charge and radius implementations. Some comments are in order. First, all predictions are quite sensitive to charge force fields and relatively less dependent on radius parameter selection. Additionally, it is difficult to identify a clear winner over all the test sets — some approaches perform better in one or two test sets, but do not do well in the rest test sets. This phenomenon highlights the difficulty of designing optimal models for solvation analysis. Moreover, the KSDFT based polarizable Poisson model is more often to provide better predictions over most charge force fields and radius parameters. This indicates a potential to develop better physical models by improving the quantum charge density calculation. Finally, we point out that the blind prediction results presented in the present work are the state-of-the-art compared to those in the literature [183, 177, 106, 205, 176, 173, 207, 174, 81, 105, 86, 87, 134].

6.5 Concluding Remarks

In this work, we proposed a hybrid physical and knowledge based protocol for the blind prediction of solvation free energies. The proposed protocol predicts nonpolar solvation free energies by statistical based models while utilizing either the Poisson model or a Density Functional Theory (DFT) based polarizable Poisson model for polar solvation free energy calculations. For the knowledge based modeling of nonpolar solvation free energies, we first utilized the assumption that molecules with the same functional group admit the same parametrization of the nonpolar solvation energy functional. For complex poly-functional-group molecules, we develop a scoring procedure to determine the optimal relative weight of each functional group. For extremely complex molecules that fail the functional group scoring method, we further develop a molecule ranking algorithm to select an optimal set of nearest neighbor molecules for parameter training. We construct atomic features for the molecule ranking. Finally, we systematically integrate the above mentioned models and methods into a robust protocol for blind solvation free energy prediction.

In the present work, we considered an experimental database of 668 solvation molecules, the largest database ever constructed for solvation, to validate our approach. Among them, SAMPL0 to SAMPL4 test sets are paid special attention. For the Poisson model or DFT based polarizable Poisson model, four sets of atomic radius parameters (i.e., Amber 6, Amber bondi, Amber mbondi and ZAP9 radii) are combined with four sets of charge force fields (i.e., AM1-BCC, Mulliken, Gasteiger and SIESTA DFT) to arrive at a total of 16 different implementations. The resulting polar solvation free energies are utilized in our statistical approaches for blind predictions. We first carry out the leave-one-out validation of the whole database. The AM1-BCC charge force field delivers a low RMSE of 1.33 kcal/mol, which is the lowest for such a large test database, to our best knowledge. We further conduct a series of leave-SAMPLx-out blind tests. On average, the BCC parametrization in the Poisson model and DFT based polarizable Poisson model performs better than other charge force fields, especially for predicting the solvation free energies of the complex molecules.

We obtain some of best known results. The optimal RMSEs for SAMPL0-SAMPL4 are respectively, 0.93, 2.82, 1.90, 0.78, and 1.03 kcal/mol, which again, are some of the best to our best knowledge.

From the solvation free energy predictions, particularly on SAMPL1 and SAMPL2 test sets, we conclude that atomic charge parametrization is extremely important for the present physical models, namely, the Poisson model or the KSDFT based polarizable Poisson model. Without an appropriate charge parametrization, the prediction errors can be amplified for molecules with complex structures. In general, for four charge parametrization methods, both the semi-empirical BCC charge and the *Ab Initio* charge calculation from the generalized KSDFT can provide relatively reliable charge assignments. For this reason, a solvation free energy prediction method that does not heavily depend on the molecule parametrization is under our consideration. The essential idea is that if one does not partition the solvation free energy into two isolated parts, the prediction errors in the polar solvation free energy will not be propagated to the nonpolar solvation free energy prediction. Alternative, our differential geometry based solvation model that dynamically couple polar and nonpolar models [270, 43] might also provide a less sensitive approach.

This work can be improved in a number of ways. First, the classification of the database into functional groups is not unique. Future studies will explore optimal molecules partitioning. Additionally, the selection and computation of atomic features need to be further investigated. It is possible to construct an optimal set of atomic features for solvation analysis and prediction. Further, in the current work, molecular ranking for nearest neighbor searching is not optimal yet. More sophisticated machine learning and/or deep learning algorithms can be developed for this purpose. Finally, a more versatile DFT solvers can be utilized to further improve our in-house polarizable Poisson model.

Chapter 7

Learning to Rank for Solvation Prediction

7.1 Introduction

Last chapter, we briefly reviewed many physical based solvation models. Besides the physical solvation models, the knowledge based models which are based on the combination of the statistical learning theory with the chemical and biological intuition. Many knowledge based models have been presented in the literatures for solvation free energy prediction. For instance, the solvation free energy modeled by the atomic solvent accessible surface area [260]; another typical model is the movable type model in which the solvation free energy itself is decomposed into the contribution from atomic solvation free energy. A partition function is trained through the whole molecule database for atomic solvation free energy modeling [290]. There are many other interesting knowledge based solvation models that exist, the comprehensive reviewing of these models is beyond the scope of this work.

In our previous work [256], we proposed a hybrid physical and statistical model for solvation free energy prediction. In which we follow the classical polar and nonpolar decomposition of the whole solvation free energy. For the polar solvation free energy prediction, we applied the Poisson model or the polarizable Poisson model with different atomic parametrization, especially the charge models ranging from semi-empirical to *Ab Initio* calculation. In terms of the nonpolar solvation free energy prediction, we modeled it by the molecular solvent excluded atomic surface area and volume enclosed by the molecular surface. We follow the classical assumption, that in the nonpolar model, molecules of different functional groups have different sets of parameters. However, this approach only works for simple monofunctional group molecules, for polyfunctional group molecules, the molecule shares the same functional groups is insufficient for accurate solvation free energy prediction. To make the method applicable to arbitrary complex molecules, we further introduced the nearest neighbor searching algorithm to detect the closest set of molecules to a given target molecule, where the similarity of the molecule is measured by the cosine similarity between the atomic features of the molecules. To make the nearest neighbor searching effective, we elaborately selected the atomic features that can distinguish molecules with different functional groups. The proposed nearest neighbor searching method is consistent with the classical functional group approach due to the fact that the nearest molecules for monofunctional group molecule found have the same functional group as the target molecule. Besides this basic consistency in terms of the chemical property, i.e., the nearest neighbors have the same functional group to their central molecule, we further noticed that the nearest neighbors possess quite close solvation free energy to its central molecule. The proposed protocol can provide accurate solvation free energy prediction provided the molecules are properly parameterized. However, when the molecular charge assignment is inappropriate, the error can be extremely large. In other words, the prediction by the previous model is sensitive to the molecular force field assignment. This motivates us to seek a method that is insensitive, or at least much less sensitive to the molecule parametrization. By an analysis of the model, we found that the main issue on the large prediction error comes from the decoupling of the polar and nonpolar solvation free energies. In this approach, once the molecule is not properly parametrized, the

Poisson model or its polarizable model will lead to huge error in the electrostatic calculation, which will be further propagated to the nonpolar solvation prediction.

Motivated by our previous work, we propose a new solvation free energy modeling paradigm. In this model, instead of treating the solvation free energy as two separated parts, i.e., polar and nonpolar parts, we consider the solvation free energy as a unity. The solvation free energy itself is modeled as a function of the molecular descriptors, where these descriptors describe the molecule at the atomic level instead of at the molecular level; the finer scale enables more accurate description, more importantly, the molecular physical properties will be captured by these descriptors. Based on the above assumption, we introduced a local solvation free energy learning framework. The procedure can be divided into two stages. First, applying Learning To Rank (LTR) theory for nearest neighbor searching; second, local hyperplane approximation for solvation free energy function, where the hyperplane is learned by a regularized least square fitting based on the nearest neighbors information to a given target molecule.

This chapter is structured as follows: Section 7.2 is devoted to methods and algorithms. We provide a brief review of our previous nearest neighbor searching algorithm in Section 8.2.1, followed by the basic ansatz for building the solvation model in this work. The LTR based nearest neighbor searching method is described in Section 7.2.2, which is divided into three parts: i) query construction which incorporates the previous nearest neighbor searching results; ii) feature selection; and iii) LTR for molecular neighbor detection. The nearest neighbor information based algorithm for solvation free energy prediction is presented in Section 7.2.3. Section 7.3 presents numerical results and discussions. After describing the datasets and force fields in Section 7.3.1 , we offer the leave-one-out validation of the proposed model in Section 7.3.2. We demonstrate that the present model is not very sensitive to atomic

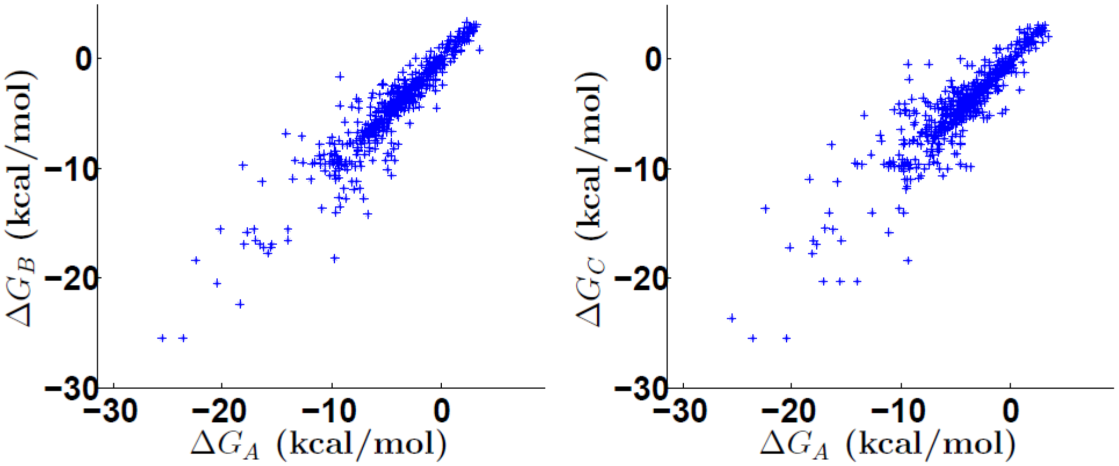


Figure 7.1: The plot of solvation free energies of the central and its neighbor molecules, the left chart for the first nearest neighbor, the right chart for the second nearest neighbor. In both cases, the horizontal axis represent the solvation free energy for the central molecule, the vertical axis stands for that of the nearest neighbor molecule.

force field parametrization. SAMPLx blind challenges are presented in Section 7.3.3. Some of the best results in solvation free energy prediction are obtained.

7.2 Methods and algorithms

7.2.1 Basic assumptions

As mentioned above, our earlier HPK model suffers from its sensitivity to charge and radius parametrizations. A major goal of the present work to develop a model that is less or not sensitive to feature parametrizations. We have observed that the solvation free energy of a target molecule is quite close to that of its nearest neighbors. Figure 7.1 depicts the correlation between experimental solvation free energy from a molecule and that of its nearest neighbors. The RMSE of solvation free energies between molecules and their first and second nearest neighbors are 1.44 and 1.77 kcal/mol, respectively.

Motivated by the above observation, we assume that there exists a feature vector to uniquely characterize and distinguish each solute molecule. Moreover, we assume that the solvation free energy of a given molecule is a functional of its atomic feature vector.

$$\Delta G_{\mathrm{A}} = f(\mathbf{x}_{\mathrm{A}}) \tag{7.2.1}$$

where ΔG_{A} is the solvation free energy of solute A, f is a unknown functional for modeling the relationship between solvation free energy and molecular descriptors, or atomic features, and $\mathbf{x}_{\mathrm{A}}$ is the feature vector of the given solute molecule. Finally, we assume that similar molecules have similar solvation free energies.

It is well known that for a given function with suitable regularity, its first order Taylor polynomial provides a very good approximation of the function, locally. For the solvation free energy function $\Delta G_{\mathrm{A}} = f(\mathbf{x}_{\mathrm{A}})$, it can be locally approximated by $\Delta G_{\mathrm{A}} \approx \nabla f(\mathbf{x}_{\mathrm{A0}})(\mathbf{x}_{\mathrm{A}} - \mathbf{x}_{\mathrm{A0}}) + \Delta G_{\mathrm{A0}}$ around the molecular atomic feature vector $\mathbf{x}_{\mathrm{A0}}$ and solvation free energy ΔG_{A0}. For a given molecule, we can predict its solvation free energy in the following featuring, learning-to-rank and learning-to-predict protocol:

- construct feature vectors for all molecules, including the target one;
- find nearest neighbor molecules to the target molecule in the database with the known solvation free energies by using a learning-to-rank algorithm;
- learn the linear functional relation between the solvation free energy and features according to a group of nearest neighbor molecules, and then predict the solvation free energy for the target molecule by the linear functional.

The fundamental important ansatz in this work is that similar molecules have close solvation free energies, which in contrast to the ansatz used in our HPK model: similar molecules share the same set of parameters in the nonpolar solvation modeling. Therefore, the solvation free energy is no longer modeled by decoupled polar and nonpolar parts. As a result, the present method is not very sensitive to charge and radius parametrization.

In the following sections, we provide detailed descriptions on feature selection, nearest neighbor searching algorithm based on LTR and solvation free energy prediction.

7.2.2 Learning-to-rank algorithm

In this section, we introduce the LTR method. The listwise LTR algorithm is employed to rank molecules. In the training procedure of the LTR model, the solvation free energy of molecule is used as the molecular label, which is consistent with our basic ansatz. A scoring function is learned in the listwise LTR method on the set of training molecules and is utilized for ranking the molecules in the set of testing molecules. The nearest neighbor searching can be regarded as a top-N recommendation problem, which is mathematically the same as the item searching in the world-wide-web.

7.2.2.1 Query construction

In our LTR model, the use of solvation free energy as a label, based on our ansatz that similar molecules have similar solvation free energies, automatically implies that similar solvation free energies indicate similar molecules. This, however, is not reasonable in general. To circumvent this deficiency in our LTR model for nearest neighbor searching, we need to partition the whole dataset (which contains a total 668 molecules as described later) into a number of subsets, where each subset is regarded as a query in the LTR terminology. The basic requirement on the query construction is that molecules in each query should have some chemical similarity. Additionally, we require that each query is invariant to the nearest molecule detection based on the cosine similarity of the atomic features proposed in our earlier work [256]. To achieve this, the most straight forward approach is that each query of molecules contains the same functional group. However, the complexity of the molecules in the dataset makes this partition impractical. A direct relaxation is that each query of the molecules have the same element types, which will be used for query construction in this work.

We first construct six groups of molecules, they are formed by the molecules with element type: i) H, C; ii) H, C, O; iii) H, C, N/H, C, N, O; iv) H, C, Cl; v) H, C, O, Cl; and vi) H, S, respectively. The third group contains molecules either with H, C, and N elements or with H, C, N, and O. This classification is due to the fact that based on the cosine similarity, molecules in these two categories may have their nearest neighbors overlapped. For the remaining molecules we iteratively add them into the above six groups based on their nearest neighbor's class label. The molecules that cannot be classified into any of the above category are regarded as a new query. Let label molecules in the dataset from 1 to 668. Figure 7.2 shows that the queries constructed based on the above procedure are invariant to the nearest neighbor searching based on the measure proposed in our earlier work [256], where each block denotes a query of molecules. It is easy to see that molecules' nearest neighbors are localized into each block. This invariance indicates that our query construction preserves the molecular chemical similarity, i.e., each query of molecules is of same similarity in the physical sense. Based on the above query construction, we can approximately regard that close solvation free energies indicate similar molecules in each query, which makes the LTR based nearest neighbor searching physically sound. We list all the queries in Supplementary material.

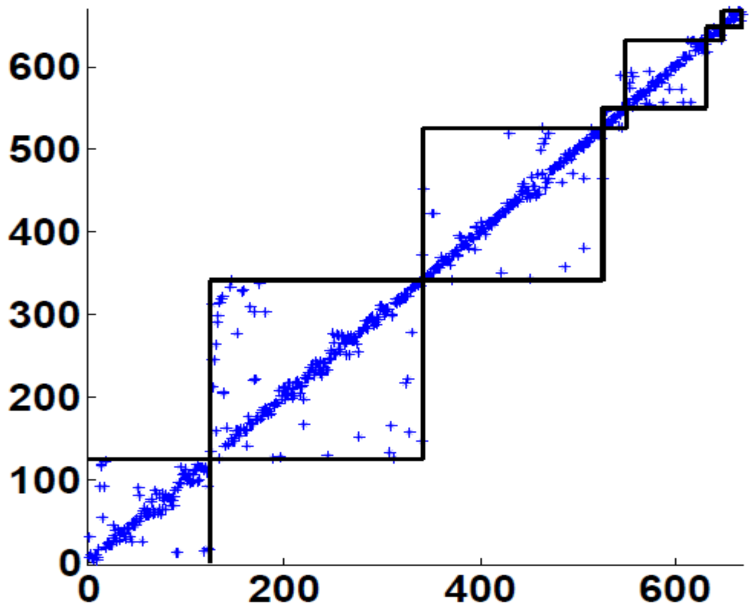


Figure 7.2: Localization of nearest neighbor molecules. The horizontal axis stands for the index of a target molecule and the vertical axis denotes the index of the nearest neighbor of the target molecule. Each block contains a query of molecules.

Since our partition of the dataset is based on similarity with chemical constrains, we discuss the similarity measure, atomic feature selection based on chemical and physical properties that facilitate the measure and LTR algorithm for ranking the molecules. For nearest neighbor searching in each query, we emphasize that the nearest neighbor is measured based on the closeness of the solvation free energies, instead of the similarity measure used before.

7.2.2.2 Feature selection

A fundamental assumption of our approach is that there exists a feature vector that can uniquely characterize and distinguish one molecule from another. Obvious, finding such as feature vector is one of most important tasks. In our previous work [256], the goal of the feature selection is to find the closest molecule to a given target molecule in the sense of functional group similarity, so we selected the features that can distinguish molecules with different functional groups, and designate molecules having the same functional groups as similar. Nevertheless, this may not be suitable to the fundamental assumption in this work, namely, similar molecules having similar solvation free energies. The desired features should reflect the similarity in solvation free energy. To this end, we select those features that their Pearson correlations to solvation free energies are larger than 0.65 or less than -0.65. Based on this criterion, we select the following features, as listed in Table 7.1.

Table 7.1: Atomic features used for LTR nearest neighbor searching.

Feature Name
Sum of atomic reaction field energy
Sum of absolute value of atomic reaction field energy
Sum of H atomic reaction field energy
Sum of absolute value of H atomic reaction field energy
Sum of O atomic reaction field energy
Sum of absolute value of O atomic reaction field energy
Minimum value of atomic reaction field energy
Maximum of the absolute value of reaction field energy
Minimum value of H atomic reaction field energy
Maximum of the absolute value of H atomic reaction field energy
Average value of atomic reaction field energy
Average of absolute value of atomic reaction field energy
Variation of atomic reaction field energy
Variation of the absolute value of reaction field energy
Variation of H atomic reaction field energy
Variation of absolute value of H atomic reaction field energy
Sum of absolute value of atomic charge
Sum of H atomic charge
Sum of absolute value of H atomic charge
Sum of O atomic charge
Sum of absolute value of O atomic charge
Minimum of atomic charge
Maximum of absolute value of atomic charge
Maximum of H atomic charge
Maximum of absolute value of H atomic charge
Average of absolute value of atomic charge
Variation of the atomic charge
Variation of the absolute value of atomic charge
Variation of the absolute value of H atomic charge
Variation of H atomic dipole

Mathematically, for a given molecule A, the sum of atomic reaction field energy is defined as

$$\Delta G_{\mathrm{rf}} = \sum_{i=1}^{N_m} \Delta G_{\mathrm{rf},i},$$

which is the same as the electrostatic solvation free energy of the solute molecule A, where N_m is the number of atoms in solute M, $\Delta G_{\mathrm{rf},i}$ is the reaction field energy contributed from the ith atom.

The sum of the absolute value of atomic reaction field energy is

$$\Delta G_{\mathrm{rf}}^{\mathrm{abs}} = \sum_{i=1}^{N_m} |\Delta G_{\mathrm{el},i}|.$$

The other features can be defined mathematically in the same manner.

The above atomic features are calculated by the following methods.

- Atomic charges and dipoles can be computed by using quantum mechanical theory.
- Atomic reaction field energies can be computed by using PB theory.
- The calculations of maximum, minimum, sum, mean, and variance are based on straightforward statistical theory.

Figure 7.3 plots some representative features compared to experimental solvation free energies. From left to right, three charts are the correlations of experimental solvation free energies with total reaction field energies, the absolute value of the mean reaction field energies of all atoms, and the absolute value of the total reaction field energy of hydrogen atoms, respectively. Their Pearson correlations are 0.87, -0.76, and -0.80, respectively.

In many physical based solvation models, the nonpolar solvation free energy is usually modeled by the solvent excluded surface area, volume enclosed by the molecular surface,

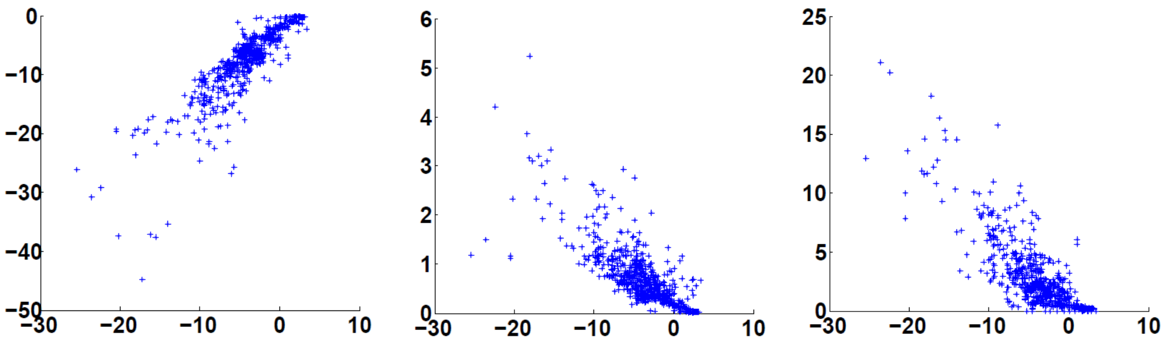


Figure 7.3: Correlations between features and experimental solvation free energies. The horizontal axes represent for the experimental solvation free energies. From left to right, three charts in the vertical axes represent total reaction field energies, the absolute value of the mean reaction field energies of all atoms, and the absolute value of the total reaction field energy of hydrogen atoms, respectively.

and the van der Waals interaction. This terms usually highly correlated with the nonpolar solvent solute interaction. However, here we note that it is not highly correlated to the total solvation free energy, some typical terms are depicted in Fig. 7.4, here the van der Waals interaction of the H and C atoms are calculated in the same way with exactly the same parameters used in [256], however the surface tension parameters are ignored, which do not matter in terms of the correlation measure. The correlations between the solvation free energy and these four features are -0.27, -0.26, 0.02, and -0.60, respectively. Here even the van der Waals interaction between the solvent and H atoms of solute molecule is slightly high, but we tested that add this to our LTR ranking feature set do not affect the ranking results. Here we should emphasis again, we have not adopt nonpolar solvation free energy related features for LTR do not mean these features are irrelevant to the solvation free energy, only in terms of LTR, they are not good molecular descriptors.

Remark 7.2.1. *The high correlation of reaction field energy calculated by the PB model with the solvation free energy indicates that the PB is an effective approach for modeling the solvation effects, the reaction field energy calculated by the PB is consistent with the*

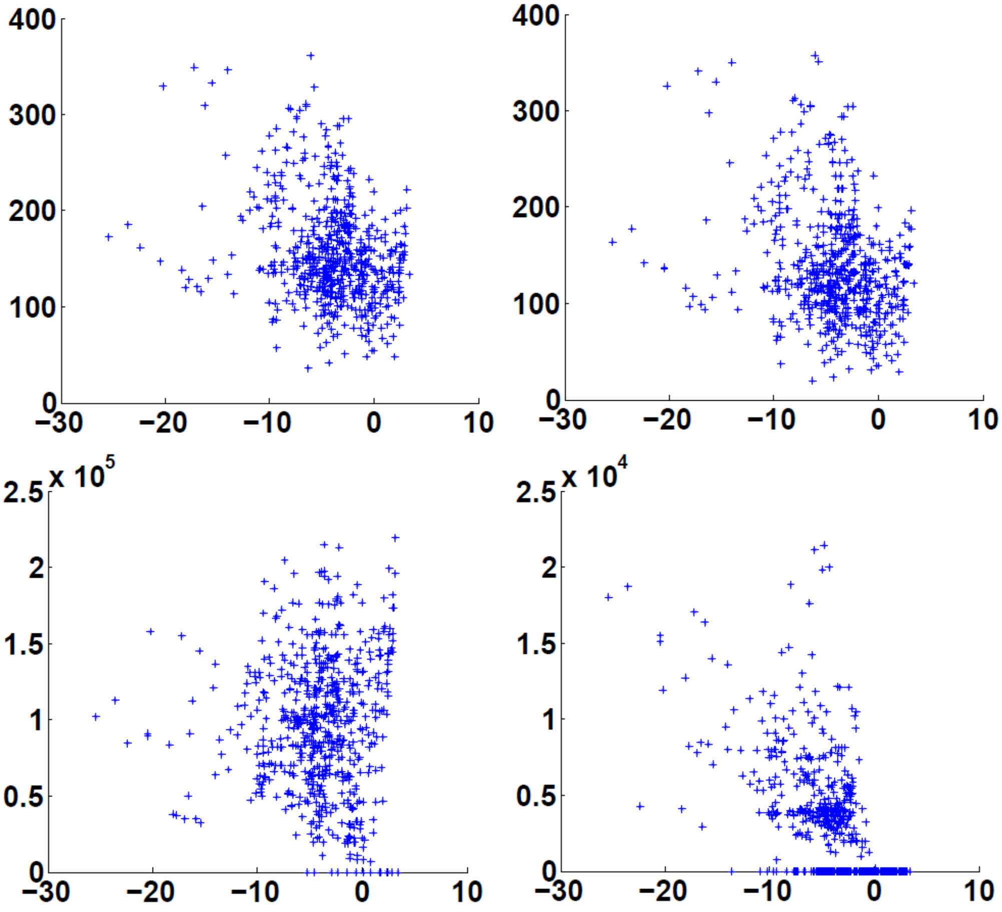


Figure 7.4: Correlations between features and experimental solvation free energies. The horizontal axises represent for the experimental solvation free energies. From left to right and up to down four sub-figures, the vertical axes represent solvent excluded surface area, volume enclosed by the solvent excluded surface, the van der Waals interaction of H and O atoms, respectively.

experimental solvation free energy.

7.2.2.3 LambdaMART for ranking

In general, LTR algorithms can be classified into three categories, i.e., pointwise, pairwise, and listwise approaches. Among them, listwise LTR algorithms are commonly regarded as the most advanced ones. In our model, each query of molecules forms a list, and thus with the accurate listwise LTR algorithm, the nearest neighbor for a given target molecule can be found accurately. In this work, we specifically select the state of the art listwise ranking algorithm, LambdaMART for ranking molecules in different queries. In this part, we provide a brief introduction to the LambdaMART algorithm. We also discuss how to apply the LambdaMART algorithm to our solvation modeling.

7.2.2.3.1 Information retrieval measures LambdaMART is the boosted tree version of the LambdaRank algorithm, where lambda is the physical gradient which makes the gradient descent applicable to the solution of the LTR problem. MART (multiple additive regression tree, also named Gradient Boosted Regression Tree (GBDT)) is a boosted tree for fitting the ranking score of the training set. In this part, we give a short review of the LambdaMART method. For more detail about this algorithm, the reader is referred to the literature [26, 78, 27, 25].

First, let us introduce the measure that is used in information retrieval research. There are several ranking quality measures used in this field, such as, Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), Expected Reciprocal Rank (ERR), and Normalized Discounted Cumulative Gain (NDCG). Among them, NDCG and ERR have the advantage in handling multiple levels of relevance, while MAP and MRR are suitable for binary relevance levels. The frequently used ranking measure for the LambdaMART algorithm is NDCG. To define NDCG, we first give the definition of the discounted cumulative gain (DCG) for a given set of search results:

$$\mathrm{DCG}@T := \sum_{i=1}^{T} \frac{2^{l_i} - 1}{\log(1+i)} \tag{7.2.2}$$

where T is the truncation level, and l_i is the label of the ith listed molecule. We typically use five levels of relevance: $l_i \in \{0, 1, 2, 3, 4\}$.

The NDCG is the normalized version of DCG

$$\mathrm{NDCG}@T := \frac{\mathrm{DCG}@T}{\max \mathrm{DCG}@T} \tag{7.2.3}$$

where the denominator is the maximum DCG@T attainable for the query, so that NDCG@$T \in [0, 1]$.

7.2.2.3.2 LambdaRank The main drawback of the above mentioned ranking quality measure is that, the gradient of them is everywhere either zero or not defined. This deficiency makes gradient descent type of optimization techniques failed for direct solving ranking problems. The key idea of LambdaRank is to introduce a physical gradient λ which directly optimizes the ranking objective function. In this part, we will briefly introduce the principle of the LambdaRank algorithm.

Consider two arbitrary molecules M_i and M_j in a given query. Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be feature vectors, and y_i and y_j be the corresponding solvation free energies, respectively. For the sake of notation simplicity, let $s_i := F(\mathbf{x}_i)$ and $s_j := F(\mathbf{x}_j)$ be model learned solvation free energies for ith and jth molecules, respectively. Further, we define the learned probability that M_i should have larger solvation free energy than M_j to be

$$P_{ij} := P(M_i \rhd M_j) := \frac{1}{1+e^{-(s_i-s_j)}} := \frac{1}{1+e^{-\sigma_{ij}}} \tag{7.2.4}$$

which models the probability by a simple sigmoid function. The associated known probability that M_i should be larger than that of M_j is

$$\bar{P}_{ij} := \frac{1+S_{ij}}{2} \tag{7.2.5}$$

where

$$S_{ij} := \begin{cases} 0, & \text{for } \; y_i = y_j \\ 1, & \text{for } \; y_i > y_j \\ -1, & \text{for } \; y_i < y_j. \end{cases}$$

In the LambdaRank, the objective function to maximize is chosen to be

$$\begin{aligned} C_{ij} &:= |\Delta Z_{ij}| \left[-\bar{P}_{ij}\log(P_{ij}) - (1-\bar{P}_{ij})\log(1-P_{ij})\right] && (7.2.6) \\ &= |\Delta Z_{ij}| \left[\frac{1}{2}(1-S_{ij})\sigma_{ij} + \log(1+e^{-\sigma_{ij}})\right] && (7.2.7) \\ &= \begin{cases} |\Delta Z_{ij}|\log(1+e^{-\sigma_{ij}}), & \text{for } \; S_{ij}=1 \\ |\Delta Z_{ij}|\log(1+e^{-\sigma_{ji}}), & \text{for } \; S_{ij}=-1 \end{cases} && (7.2.8) \end{aligned}$$

where ΔZ_{ij} is the change of NDCG@T by swapping the rank positions of M_i and M_j. It should be pointed out that the above target function incorporates both a ranking quality measure and a cross entropy cost function.

From now on, we assume all $(i,j) \in I$. Here $I := \{(i,j)|S_{ij} = 1\}$: the set of pairs of indices (i,j), for which desired M_i have a larger solvation free energy than M_j. Thus, the expression of C_{ij} can be simplified as

$$C_{ij} := |\Delta Z_{ij}| \log(1 + e^{-\sigma_{ij}}). \tag{7.2.9}$$

By treating ΔZ_{ij} as a constant, we have

$$\frac{\partial C_{ij}}{\partial s_i} = \frac{\partial C_{ij}}{\partial \sigma_{ij}} = -\frac{|\Delta Z_{ij}|}{1 + e^{\sigma_{ij}}} := -|\Delta Z_{ij}|\rho_{ij} = -\frac{\partial C_{ij}}{\partial \sigma_{ji}} = -\frac{\partial C_{ij}}{\partial s_j}.$$

We define the following λ-gradient

$$\lambda_{ij} := -\frac{\partial C_{ij}}{\partial \sigma_{ij}} = \frac{|\Delta Z_{ij}|}{1 + e^{\sigma_{ij}}} = |\Delta Z_{ij}|\rho_{ij}. \tag{7.2.10}$$

Then we have

$$\frac{\partial C_{ij}}{\partial s_i} = -\lambda_{ij}$$

and the second gradient is

$$\frac{\partial^2 C_{ij}}{\partial s_i^2} = -\frac{\partial \lambda_{ij}}{\partial s_i} = \frac{|\Delta Z_{ij}| e^{\sigma_{ij}}}{(1 + e^{\sigma_{ij}})^2} = |\Delta Z|\rho_{ij}(1 - \rho_{ij}). \tag{7.2.11}$$

For a given query, assembling the pairwise cost yields the total cost function

$$\begin{aligned} C = \sum_i \sum_{j:(i,j)\in I} C_{ij} + \sum_i \sum_{j:(j,i)\in I} C_{ji} = \sum_i \sum_{j:(i,j)\in I} |\Delta Z_{ij}| \log(1 + e^{-\sigma_{ij}}) + \\ \sum_i \sum_{j:(j,i)\in I} |\Delta Z_{ji}| \log(1 + e^{-\sigma_{ji}}) \end{aligned} \tag{7.2.12}$$

The first order gradient of C is

$$\frac{\partial C}{\partial s_i} = \sum_{j:(i,j)\in I} \frac{\partial C_{ij}}{\partial s_i} + = \sum_{j:(j,i)\in I} \frac{\partial C_{ji}}{\partial s_i} = \tag{7.2.13}$$

$$\sum_{j:(i,j)\in I} (-|\Delta Z_{ij}|\rho_{ij}) + \sum_{j:(j,i)\in I} (-|\Delta Z_{ji}|\rho_{ji}) = \sum_{j:(i,j)\in I} (-\lambda_{ij}) + \sum_{j:(j,i)\in I} (\lambda_{ji})$$

Thus, we define the λ-gradient for a given molecule M_i in the query

$$\lambda_i := \sum_{j:(i,j)\in I} \lambda_{ij} + \sum_{j:(j,i)\in I} \lambda_{ij} = \sum_{j:(i,j)\in I} \lambda_{ij} - \sum_{j:(j,i)\in I} \lambda_{ji}. \tag{7.2.14}$$

Hence, we simply have

$$\frac{\partial C}{\partial s_i} = -\lambda_i. \tag{7.2.15}$$

Furthermore, it is easy to verify the following second order gradient expression

$$\frac{\partial^2 C}{\partial s_i^2} = -\frac{\partial \lambda_i}{\partial s_i}. \tag{7.2.16}$$

7.2.2.3.3 Gradient boosting Before introducing the MART, we first present the principle of the gradient boosting. We consider the loss function $L = L(y, F)$ of input score y and model function $F = F(\mathbf{x})$. The goal of the gradient boosting is to minimize the loss function

$$\min_F L(y, F). \tag{7.2.17}$$

Borrowing the idea from the classical gradient descent, the gradient boosting iteratively updates the model function F in a given functional space

$$F_{n+1} = F_n + \rho f_n(\mathbf{x}) \tag{7.2.18}$$

where $f_n(\mathbf{x})$ is a model from a given functional space, i.e., regression tree for the MART, that fits the residual $\{\tilde{y}\}_i$

$$\tilde{y}_i = -\frac{\partial L(y_i, F)}{\partial F}|_{F=F_n}. \tag{7.2.19}$$

The shrinkage parameter is obtained by solving the following optimization problem

$$\rho = \arg\min_{\rho} L(y, F_n + \rho f_n(\mathbf{x})), \tag{7.2.20}$$

via a simple line searching algorithm.

7.2.2.3.4 Gradient boosted regression trees In the aforementioned gradient boosting algorithm, if we select the functional space to be the regression tree, this results in a GBDT algorithm. Mathematically, the regression tree is formulated as

$$f(\mathbf{x}) = f(\mathbf{x}, \{\gamma_j, R_j\}_i^J) = \sum_{j=1}^{J} \gamma_j \mathbf{1}(\mathbf{x} \in R_j), \tag{7.2.21}$$

where J is the number of leaves, and $\{R_j\}_i^J$ are disjoint regions that cover all feature space, with each of them being covered in one leaf. Here $\{\gamma_j\}$ are the values in the corresponding leaf.

In this case, we first construct $\{R_j\}_i^J$ to fit $\{\mathbf{x}_i, y_i\}$ by a least square algorithm. According to Eq. (7.2.18), we have

$$F_{n+1} = F_n + \sum_{j=1}^{J} \gamma_j \mathbf{1}(\mathbf{x} \in R_j).$$

Additionally, Eq. (7.2.20) indicates that

$$\gamma_j = \arg\min_{\gamma} \sum_{i:\mathbf{x}_i \in R_j} L(y_i, F_n(\mathbf{x}_i) + \gamma) := \arg\min_{\gamma} \sum_{i:\mathbf{x}_i \in R_j} g \tag{7.2.22}$$

where $g = g(y_i, F_n(\mathbf{x}_i)) = L(y_i, F_n(\mathbf{x}_i) + \gamma)$. When there is no closed form solution to Eq. (7.2.22), we can approximate it by a single Newton step

$$\gamma_j = -\frac{\sum_{i:\mathbf{x}_i \in R_j} \frac{\partial g}{\partial F_n}}{\sum_{i:\mathbf{x}_i \in R_j} \frac{\partial^2 g}{\partial F_n^2}}. \tag{7.2.23}$$

7.2.2.3.5 LambdaMART The goal of the LambdaMART is to maximize

$$\max_{F} C \tag{7.2.24}$$

where C is the total cost function defined by Eq. (7.2.12), and F is a MART.

7.2.2.3.6 LambdaMART for molecules ranking Now let us turn to the application of LambdaMART to the solvation prediction. In each query of the molecules, the solvation free energies themselves are regarded as the labels of molecules, and the corresponding features are discussed in the next part. Our method can be summarized as ranking the nearest neighbors of a target molecule based on their solvation free energies and then, learning a relation between features and solavtion free energies for predicting the solvation free energy of the target molecule.

7.2.3 Functional estimation for solvation free energy prediction

We discuss the solvation free energy prediction for a given target molecule in this section. Based on our assumption that solute solvation free energy is a functional of the feature vector, solvation free energy prediction is actually to construct a energy functional around the target molecule. This functional will be utilized for solvation free energy prediction for the target molecule.

Consider the solvation free energy for a target molecule A characterized by its feature vector $\mathbf{x}_{\mathrm{A}} = (\mathbf{x}_{\mathrm{A}1}, \mathbf{x}_{\mathrm{A}2}, \cdots, \mathbf{x}_{\mathrm{A}n})$, where n is the dimension of the feature space, i.e., the space of all feature vectors. Here, we simply use the features that utilized in the LTR procedure for training the local solvation function, since these features in some sense also reflects the order of the solvation free energy. However, we believe that optimal selection of the features for training this function can leads to better prediction, since we note that in our features selection the nonpolar features were not used. For global solvation free energy function training, neglecting nonpolar features may leads to tremendous error, but in our LTR model, we are restricted our learning in the local sense. From the LTR framework, we find m feature vectors of the nearest neighbors and corresponding known solvation free energies $(\mathbf{x}_1, \Delta G_1), (\mathbf{x}_2, \Delta G_2), \cdots, (\mathbf{x}_m, \Delta G_m)$. Note that in general, the number of nearest neighbors found is far less than the dimension of the feature space, i.e., $m \ll n$.

In this work, we assume the functional relation between features and solvation free energies for target molecule A has the form

$$\Delta G_{\mathrm{A}} = b + \sum_{i=1}^{n} w_i \mathbf{x}_{\mathrm{A}i}, \tag{7.2.25}$$

where w_i is the weight for feature $\mathbf{x}_{\mathrm{A}i}$ and b can be intuitively understood as the height of hyperplane embedded in the Euclidean space. Equation (8.2.21) can be regarded as a first order Taylor polynomial approximation of the solvation free energy $\Delta G_{\rm A} = f(\mathbf{x}_{\rm A})$.

Since the fact that $m \ll n$, the directly regression based on the least square approach may leads to over-fitting. To avoid over-fitting there are generally two strategies for determining $\{w_i\}$ and b:

- sparse solution via a compressed sensing approach;

- Tikhonov regularization based least square fitting.

In this work, we use the second strategy for training the local regression model for solvation free energy prediction. The local regression problem is equivalent to solve the linear system in Eq. (7.2.26) in the L_2 sense

$$\begin{pmatrix} \Delta G_1 \\ \Delta G_2 \\ \vdots \\ \Delta G_m \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} + \begin{pmatrix} b \\ b \\ \vdots \\ b \end{pmatrix}. \tag{7.2.26}$$

Equation (8.2.15) can be written as

$$\mathbf{\Delta G} = \mathbf{x}\mathbf{w} + b\mathbf{1}, \tag{7.2.27}$$

where $\mathbf{\Delta G} = (\Delta G_1, \Delta G_2, \cdots, \Delta G_m)^T$, $\mathbf{w} = (w_1, w_2, \cdots, w_n)^T$, $\mathbf{1}$ is a m-dimensional column vector with all elements equal 1, and matrix $\mathbf{x}$ is given by

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}.$$

To avoid overfitting, we add the L_2 penalty to the weight vector $\mathbf{w}$, and thus Eq. (8.2.16) can be solved by the following optimization problem

$$\min_{\mathbf{w},b} ||\Delta\mathbf{G} - \mathbf{x}\mathbf{w} - b\mathbf{1}||_2^2 + \lambda||\mathbf{w}||_2^2 := \min_{\mathbf{w},b} F, \tag{7.2.28}$$

where λ is the regularization parameter, which is set to 100 in this work, $|| * ||_2$ denotes the L_2 norm of the quantity $*$.

By solving $\frac{\partial F}{\partial \mathbf{w}} = 0$, we have

$$\mathbf{w} = \left(\mathbf{x}^T\mathbf{x} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{x}^T\Delta\mathbf{G} - \mathbf{x}^T(b\mathbf{1})\right), \tag{7.2.29}$$

where I is $m \times m$ identity matrix.

To find the value b that solves the optimization problem Eq. (8.2.17), we relax $b\mathbf{1}$ to arbitrary vector $\mathbf{b} = (b_1, b_2, \cdots, b_m)^T$, by solving $\frac{\partial F}{\partial \mathbf{b}} = 0$, we have

$$\mathbf{b} = \Delta\mathbf{G} - \mathbf{x}\mathbf{w}. \tag{7.2.30}$$

Therefore, we obtain the unbiased estimation of b as

$$b = \frac{\sum_{i=1}^{m}(\Delta\mathbf{G} - \mathbf{xw})_i}{m}, \tag{7.2.31}$$

where $(\Delta\mathbf{G} - \mathbf{xw})_i$ is the ith component of the vector $\Delta\mathbf{G} - \mathbf{xw}$.

We can solve the optimization problem Eq. (7.2.28) by alternating iterations between Eq. (7.2.29) and Eq. (7.2.31), which is essentially an expectation-Maximization (EM) algorithm.

After obtaining optimized parameters $\mathbf{w}$ and b, the solvation free energy of target molecule A, is predicted by Eq. (7.2.25).

7.3 Numerical results and discussions

7.3.1 Dataset and feature parametrization

7.3.1.1 Dataset

In order to assess the performance of the present method, we consider the same dateset that has been constructed in our earlier work [256]. With a total of 668 molecules, this dataset is the largest to date, to our knowledge, and contains both monofunctional group and polyfunctional group molecules. Experimental solvation free energies are collected from the literature [30, 261, 153]. The main part of our dataset, i.e., 589 molecules, overlaps with Mobley's solvation database (http://mobleylab.org/resources.html). All the structures of this dataset are downloaded from the Pubchem project (https://pubchem.ncbi.nlm.nih.gov/). More detailed description of the dataset can be found in our earlier work [256].

7.3.1.2 Atomic feature parametrization

In atomic feature generation, atomic charges and atomic dipoles are calculated via the distributed multipole analysis (DMA) method [227], in which the charge density is originally computed by the density function theory with B3LYP and 6-31G basis selection in Gaussian quantum chemistry software [80, 18, 148]. Atomic reaction field energies (i.e., atomic electrostatic solvation energies) are calculated by our in-house MIBPB software [253, 39, 90] with a probe radius of 1.4 Å and dielectric constants being 1 and 80, respectively for the solute and solvent domains. A uniform grid size of 0.25 Å is used in all atomic reaction field energy calculations. To examine the sensitivity of the present approach to charge force fields, which was a major issue in our earlier HPK model, we utilize three types of atomic radii, namely, Amber 6, Amber bondi, and Amber mbondi2 [35]. Additionally, we consider three type of charge assignments, namely, OpenEye-AM1-BCC v1 parameters [123], Gasteiger [85], and Mulliken [35]. The combination of radius sets and charge sets gives rise to a total of nine different parametrizations, which have already been utilized in our earlier work [256] to offer some of the best solvation prediction results. For the regularized least square hyperplane fitting, the regularization parameter λ is set to 100.

7.3.2 Leave-one-out prediction

First, we consider the leave-one-out test on the whole dataset of 668 molecules. In this test, we regard the solvation free energy of one molecule as unknown, and use the remaining molecules to predict the solvation free energy for the target molecule. The purpose of the leave-one-out test is two twofold. First, it helps for the parameter selection, i.e., the number of nearest neighbors to be used for the prediction of the target molecule's solvation free energy and the parameters used in training the LambdaMART. Second, the leave-one-out test can demonstrate the performance of the proposed model for solvation free energy prediction. The performance of leave-one-out test is measured by both the root mean square error (RMSE) and mean error (ME), respectively defined by

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(\Delta G_i^{\mathrm{Pred}} - \Delta G_i^{\mathrm{Expl}}\right)^2}{N}} \tag{7.3.1}$$

and

$$\mathrm{ME} = \frac{\sum_{i=1}^{N}\left(\Delta G_i^{\mathrm{Pred}} - \Delta G_i^{\mathrm{Expl}}\right)}{N} \tag{7.3.2}$$

where N is the total number of molecules in our dataset, $\Delta G_i^{\mathrm{Expl}}$ and $\Delta G_i^{\mathrm{Pred}}$ stand for the experimental and predicted solvation free energies for the ith molecule, respectively.

The RMSE measures the accuracy of the prediction. A small RMSE indicates the predictions for the whole dataset are uniformly accurate. ME is used to determine whether the prediction is biased or not. If the ME is close to zero, it means that the prediction is unbiased.

7.3.2.1 Number of neighbors involved

In applying our model, one has to determine how many nearest neighbors to be involved for the solvation free energy prediction. In general, this number depends on the training dataset and parametrization. Numerically, one can use either leave-one-out or five-fold cross validation to determine the optimal number of nearest neighbors. Table 7.2 lists RMSE and ME of our leave-one-out prediction using a total of 9 different combinations of atomic radii and charge force fields. The use of different numbers of nearest neighbors is examined as

Table 7.2: The RMSE and ME of the solvation free energy prediction with different parametrization on the molecules, the errors are calculated by the proposed solvation models with different number of nearest neighbors involved. All with unit kcal/mol.

Parametrization	Error	1	2	3	4	5	6	7	8
BCC+Amber6	RMSE	1.059	1.068	1.079	1.086	1.111	1.133	1.128	1.109
	ME	0.030	0.043	0.037	0.029	0.037	0.035	0.050	0.045
BCC+Bondi	RMSE	1.010	1.071	1.078	1.093	1.010	1.084	1.086	1.095
	ME	0.036	0.058	0.061	0.051	0.070	0.073	0.071	0.078
BCC+MBondi2	RMSE	1.099	1.106	1.136	1.139	1.177	1.177	1.175	1.148
	ME	0.040	0.022	0.017	0.029	0.030	0.047	0.047	0.469
GAS+Amber6	RMSE	1.331	1.294	1.267	1.278	1.268	1.284	1.321	1.336
	ME	0.002	-0.005	-0.008	-0.003	0.019	-0.003	0.019	0.038
GAS+Bondi	RMSE	1.203	1.193	1.227	1.249	1.276	1.302	1.311	1.340
	ME	-0.011	-0.004	0.012	0.028	0.042	0.054	0.055	0.065
GAS+MBondi2	RMSE	1.165	1.200	1.179	1.163	1.175	1.192	1.207	1.220
	ME	-0.018	-0.037	-0.031	-0.018	0.003	0.017	0.024	0.036
MUL+Amber6	RMSE	1.356	1.328	1.313	1.341	1.332	1.338	1.347	1.368
	ME	0.025	0.037	0.036	0.037	0.037	0.057	0.071	0.076
MUL+Bondi	RMSE	1.281	1.252	1.229	1.238	1.227	1.250	1.280	1.305
	ME	0.030	0.027	0.031	0.033	0.038	0.045	0.053	0.070
MUL+MBondi2	RMSE	1.264	1.286	1.277	1.261	1.275	1.285	1.316	1.326
	ME	-0.012	-0.022	-0.020	-0.010	0.013	0.012	0.028	0.032

well. We note that judged by RMSEs, our method is not sensitive to the number of nearest neighbors. All of the top ten recommendations have the same level of accuracy. However, when MEs are also taken into consideration, it is found that a large number of nearest neighbors typically contributes to a large EM. We propose to select the number of nearest neighbors based on the following criteria:

- The RMSE should be as small as possible to give an accurate prediction.
- The ME should be as close to zero as possible to give an unbiased prediction.
- At the same level of RMSE and ME, it is preferred to involve more molecules, which makes it easy to determine solvation free energy functional.

Usually, there is a trade off among the aforementioned criteria in selecting the number of molecules for solvation prediction.

Based on the above criteria, the final choice of the number of nearest neighbors that is utilized for each force field parametrization for solvation free energy prediction is listed in Table 7.3. We emphasize that the proposed method is quite robust with respect to different choices.

Table 7.3: The number of nearest neighbor involved for the solvation free energy prediction for different force field parametrization.

Charge	Radius	Number of Nearest Neighbors
BCC	Amber6	4
	Amber Bondi	4
	Amber MBondi2	3
GAS	Amber6	4
	Amber Bondi	2
	Amber MBondi2	4
MUL	Amber6	3
	Amber Bondi	3
	Amber MBondi2	4

Table 7.4: The RMSE and ME of the leave-one-out test of the solvation free energy prediction with different methods. For a comparison, the numbers in the parenthesis is calculated by the method proposed in the work [256]. All with unit kcal/mol

Radius	Charge	BCC	Mulliken	Gasteiger
Amber 6	RMSE	1.08 (1.47)	1.27 (1.49)	1.31 (1.65)
	ME	0.03 (-0.13)	-0.00 (-0.20)	0.03 (-0.19)
Amber Bondi	RMSE	1.09 (1.34)	1.19 (1.48)	1.22 (1.66)
	ME	0.05 (-0.14)	-0.00 (-0.21)	0.03 (-0.13)
Amber MBondi2	RMSE	1.10 (1.33)	1.16 (1.49)	1.26 (1.68)
	ME	0.02 (-0.14)	-0.02 (-0.22)	-0.01 (-0.22)

7.3.2.2 Accuracy and sensitivity analysis

In this part, we compare the present leave-one-out predictions with those of our earlier HPK model [256] under the same radius and charge parametrizations.

Table 7.4 lists the RMSEs and MEs of the current model predictions. For a comparison, corresponding RMSEs obtained by our previous HPK model is also listed in parentheses. From Table 7.4, we can conclude the follows:

- The LTR solvation model is in much more accurate than our previous HPK model. The best prediction by the present model has an RMSE of 1.08 kcal/mol, compared to the lowest RMSE of 1.33 kcal/mol achieved by the previous model. The the worst RMSE of the present prediction is 1.31 kcal/mol, which is still better than the best result obtained by the previous model. Note that worst earlier result has an RMSE of 1.68kcal/mol [256].

- The LTR solvation model provides unbiased solvation predictions, as indicated from ME results. The predictions with different molecular parametrizations all achieve near zero MEs. The MEs of the previous model are almost ten times larger than those of the LTR solvation model. Additionally, we note that no matter with what type of molecular parametrization, the previous predictions are biased toward one direction, whereas the present model has MEs of both signs.

- The LTR solvation model is less sensitive to the atomic feature parametrization compared to the HPK solvation model. The ranges of the RMSEs for LTR and HPK models due to 9 different parametrizations are 1.08-1.31 and 1.33-1.68 kcal/mol, respectively. Obviously, larger range in prediction RMSEs indicates that the HPK model is more sensitive to atomic feature parametrization.

7.3.3 Blind prediction of SAMPLx challenges

In this part, we consider the blind prediction of solvation free energy for the SAMPLx challenge sets. Our LTR solvation model is applied to all of five SAMPL test sets, i.e., SAMPL0-SAMPL4. We adopt the same protocol used in our previous leave-SAMPLx-out prediction [256]. Specifically, in each SAMPL test prediction, we exclude all the molecules in the given SAMPL in our LTR process, and use the remaining molecules as our training set to find a set of the nearest neighbors to each each molecule in the SAMPL test set. Both RMSE and ME measures are evaluated to assess the performance of the proposed LTR model. The same 9 sets of charge and radius paramerizations are implemented in leave-SAMPLx-out tests.

First, let us consider the solvation free energy prediction for the SAMPL0 test set, which contains a total of 17 molecules. All structures of this test set are relatively simple. However, the molecule species of this set is quite diverse. Many researchers have reported their solvation free energy predictions for this challenge set [183, 134]. In our earlier work [256], we have shown that when Gasteiger charge and Amber 6 radius are used for parametrization, our HPK model gives a blind prediction RMSE of 1.20 kcal/mol for the whole set. When one molecule that contains a Br atom was excluded, by using Amber Bondi radii and the DFT based polarizable Poisson model for electrostatic calculation, our optimal prediction has an RMSE of 0.93 kcal/mol. Prior to our earlier work, the optimal prediction to this test set has an RMSE of 1.34 kcal/mol for the whole set [134]. Figure 7.5 depicts the present LTR results for a total of 9 charge and radius combinations. When BCC charge is used, the RMSEs of our predictions with three radius parametrizations are all less than 0.90 kcal/mol. Our optimal prediction has an RMSE of 0.76 kcal/mol, obtained from both Amber 6 and Amber Mbondi2 radius parametrizations in conjugation with the BCC charge assignment. Apart from delivering the same RMSE, these two parametrizations also offer quite close MEs. For a comparison with our earlier work, we have also plot the RMSEs from our previous HPK model in Fig. 7.5. Obviously, except for Mulliken charge and Amber 6 or Amber Bondi radius, the present predictions are more accuracy than those of our previous HPK model.

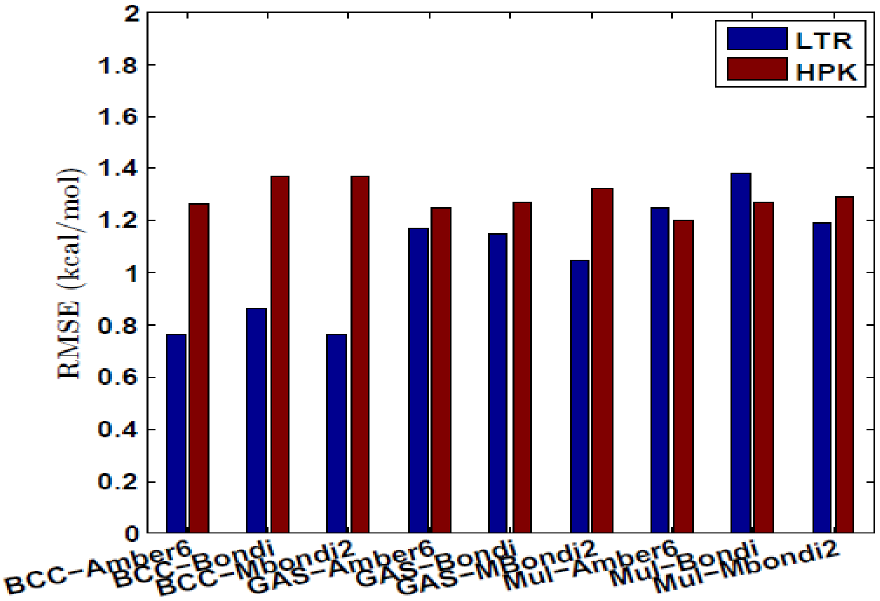


Figure 7.5: Illustration of prediction RMSEs obtained with different molecular parametrizations by the LTR and HPK models for SAMPL 0 test set.

Having demonstrated the superiority of the proposed LTR model for the blind prediction of the SAMPL0 challenge set, we further consider the SAMPL1 test set, which is generally believed to be the most difficulty one, due to the following two reasons. First, the molecular structures of this test set are extremely complex compared to other molecules with known experimental solvation free energies. Second, the uncertainty of SAMPL1 experimental data is very large. For some molecules the uncertainty is as large as 2.0 kcal/mol [105]. Nevertheless, it is extremely desirable to develop an accurate modeling paradigm for this test set because most molecules in this test set are druggable. The best prediction for the whole set has an RMSE of 2.45 kcal/mol [134]. On a subset of the SAMPL1 test set that contains only 56 molecules, the best performance was shown to give an RMSE of 2.4 kcal/mol. Our previous HPK model can provide best prediction for this set with RMSE 2.82 kcal/mol when the DFT is used for the charge assignment. However, this result has a prerequisite that one molecule that contains a Br element was ignored due to the lack of a proper Br pesudopotential for the DFT software. The optimal prediction for the whole test set has an RMSE of 3.07 kcal/mol based on the AM1-BCC charge and Amber Bondi radius parametrization [256]. Additionally, our previous HPK model is very sensitive to the force field assignment. Our earlier RMSEs for 16 charge and radius combinations vary from 3.07 to 6.16 kcal/mol. Figure 7.7 illustrates the comparison of the LTR and HPK predictions on the whole SAMPL1 test set. It is easy to see that the LTR model is much more accurate. The optimal prediction has an RMSE as small as 2.14 kcal/mol, which is the best to our knowledge. Additionally, the present LTR model is very robust with respect to the change in force fields. The maximum and minimum prediction RMSEs over 9 sets of charge and radius parametrizations are 2.14 and 2.81 kcal/mol, respectively. The difference between the maximum and minimum is 0.67 kcal/mol, which is much smaller than experimental uncertainty of 2 kcal/mol.

Another difficult test set is SAMPL2, which contains a total of 30 molecules [137]. The

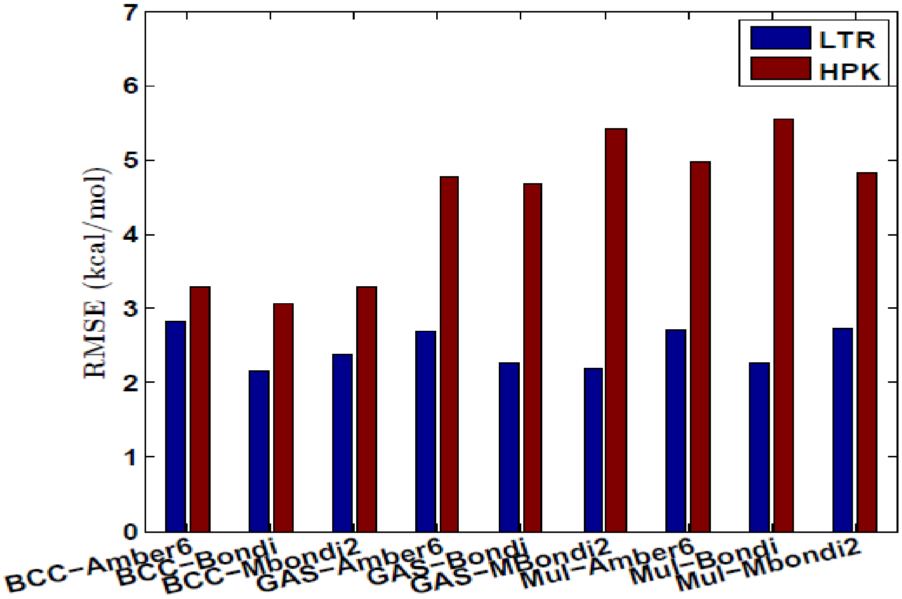


Figure 7.6: Illustration of prediction RMSEs obtained with different molecular parametrizations by the LTR and HPK models for SAMPL1 test set.

experimental uncertainty on these molecules is much less than that of the SAMPL1 test set. Nevertheless, accurate solvation prediction for this set is rare. Using all-atom molecular dynamics simulations and multiple starting conformations for blind prediction, Klimovich and Mobley reported an RMSE of 2.82 kcal/mol over the whole set and 1.86 kcal/mol over all the molecules except several hydroxyl-rich compounds [137]. Some best prediction has an RMSE of 1.59 kcal/mol [134]. In our previous test, the molecule containing an I atom (5-iodouracil) is excluded in all calculations due to the lack of appropriate charge force field. In this work, we also ignore this molecule for the same reason. The HPK model gives an optimal prediction with RMSE 1.96 kcal/mol. However, the RMSEs of the prediction vary over a large range, from 1.96 to 4.86 kcal/mol, when different charge and radius force fields are applied. In the present work, an optimal LTR prediction has an RMSE of 1.90 kcal/mol, offering a slight improvement over our earlier prediction. However, the variation of RMSEs under different charge and radius parametrizations is only 1.2 kcal/mol (i.e., from 1.90 to

3.10 kcal/mol), which indicates the robustness of the present LTR model compared to the earlier HPK model. A comparison of HPK and LTR predictions is given in Fig. 7.7.

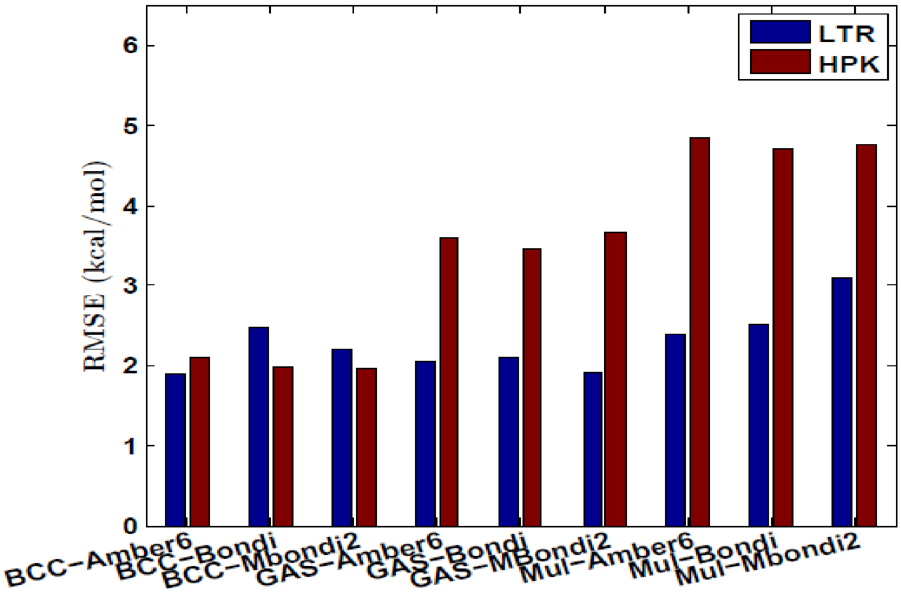


Figure 7.7: Illustration of prediction RMSEs obtained with different molecular parametrizations by the LTR and HPK models for SAMPL2 test set.

The SAMPL3 test set, which contains 36 molecules, is relatively ease for blind prediction. The structures of SAMPL3 molecules are relatively simple, and most molecules in this set are chlorinated hydrocarbon molecules [87]. The best prediction in the literature offers an RMSE of 1.29 kcal/mol [134]. Our earlier HPK model achieved the most accurate solvation free energy prediction for the whole set, especially when Gasteiger charge is used for parametrization. It is found that Gasteiger charges do give a good charge description for the chlorinated hydrocarbon molecules. Figure 7.8 depicts the RMSEs of the predictions by LTR and HPK models. It is easy to see that two methods give almost the same optimal prediction: the LTR prediction has an optimal RMSE of 0.87 kcal/mol, while the HPK prediction has optimal RMSE of 0.82 kcal/mol. The RMSEs of LTR predictions from different parametrizations vary over a small range of 0.61 kcal/mol (i.e., from 0.87 to 1.48 kcal/mol), which further verifies the robustness of the LTR solvation model.

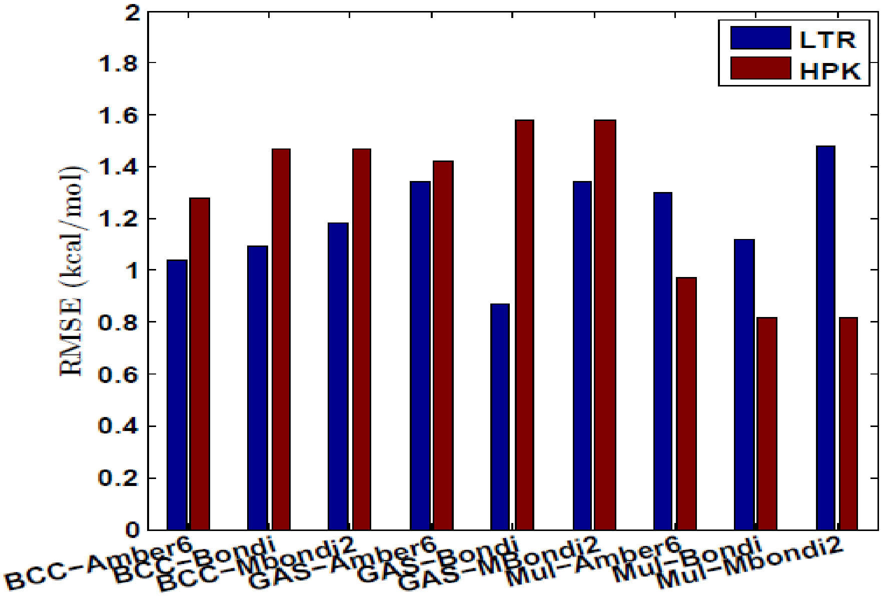


Figure 7.8: Illustration of prediction RMSEs obtained with different molecular parametrizations by the LTR and HPK models for SAMPL3 test set.

Finally, we consider the SAMPL4 test set, which is a very popular one. Many explicit, implicit, integral equation and quantum approaches have been applied to this set [177]. Our previous test using the HPK model shows an extremely accurate and robust prediction with the optimal RMSE of 1.03 kcal/mol. The optimal prediction of the LTR model has an RMSE of 1.01 kcal/mol. The variation in prediction RMSEs is as small as 0.24 kcal/mol (i.e., from 1.01 to 1.35 kcal/mol) over 9 different parametrizations. For a comparison, we depict the prediction RMSEs of both the LTR and HPK models in Fig. 7.9. It is seen that for SAMPL4, both models give the same level of accuracy and robustness in the solvation free energy prediction.

Table 7.5 provides a summary of the LRT RMSEs and MEs for all SAMPL0-SAMPL4 test sets. These results indicate that overall, the LTR framework is more accurate in solvation predictions than our earlier HPK model for SAMPL test sets. This is especially true for

Table 7.5: The RMSEs and MEs of the solvation free energy predictions with different parametrizations. The numbers inside and outside parentheses are the results calculated by the HPK and LTR models, respectively. All errors are with unit kcal/mol.

Test set	Radius	Error	BCC	Mulliken	Gasteiger
SAMPL0	Amber 6	RMSE	0.76 (1.26)	1.17 (1.25)	1.25 (1.20)
		ME	0.30 (0.76)	0.46 (-0.12)	0.10 (-0.13)
	Amber Bondi	RMSE	0.86 (1.37)	1.15 (1.27)	1.38 (1.27)
		ME	0.15 (0.86)	0.29 (-0.20)	-0.07 (-0.24)
	Amber MBondi2	RMSE	0.76 (1.37)	1.05 (1.32)	1.19 (1.29)
		ME	0.29 (0.88)	0.42 (0.21)	0.34 (-0.25)
SAMPL1	Amber 6	RMSE	2.81 (3.27)	2.69 (4.77)	2.70 (4.96)
		ME	-1.28 (0.88)	-0.43 (-2.28)	-0.33 (-1.12)
	Amber Bondi	RMSE	2.14 (3.07)	2.26 (4.68)	2.27 (5.55)
		ME	-0.97 (0.99)	-0.42 (-2.28)	-0.12 (-1.26)
	Amber MBondi2	RMSE	2.38 (3.30)	2.20 (5.41)	2.74 (4.82)
		ME	-1.15 (1.22)	-0.42 (-2.39)	-0.37 (-0.67)
SAMPL2	Amber 6	RMSE	1.90 (2.11)	2.04 (3.59)	2.39 (4.86)
		ME	0.71 (-0.65)	1.64 (1.65)	1.40 (2.65)
	Amber Bondi	RMSE	2.48 (1.97)	2.09 (3.47)	2.52 (4.72)
		ME	1.10 (-0.26)	1.41 (1.69)	1.53 (2.61)
	Amber MBondi2	RMSE	2.21 (1.96)	1.92 (3.66)	3.10 (4.76)
		ME	1.14 (-0.46)	1.31 (2.62)	2.19 (1.79)
SAMPL3	Amber 6	RMSE	1.04 (1.28)	1.34 (1.42)	1.30 (0.97)
		ME	0.04 (0.38)	-0.24 (0.72)	-0.15 (-0.16)
	Amber Bondi	RMSE	1.09 (1.47)	0.87 (1.58)	1.12 (0.82)
		ME	0.12 (-0.56)	-0.05 (0.85)	0.04 (-0.09)
	Amber MBondi2	RMSE	1.18 (1.47)	1.34 (1.58)	1.48 (0.82)
		ME	0.09 (-0.56)	0.24 (0.85)	0.00 (-0.09)
SAMPL4	Amber 6	RMSE	1.06 (1.28)	1.30 (1.20)	1.22 (1.08)
		ME	0.27 (-0.14)	0.23 (0.11)	0.34 (0.18)
	Amber Bondi	RMSE	1.01 (1.12)	1.35 (1.41)	1.28 (1.10)
		ME	0.02 (0.06)	-0.03 (0.31)	0.08 (0.33)
	Amber MBondi2	RMSE	1.10 (1.09)	1.21 (1.33)	1.10 (1.03)
		ME	0.16 (0.15)	0.23 (0.14)	0.37 (-0.08)

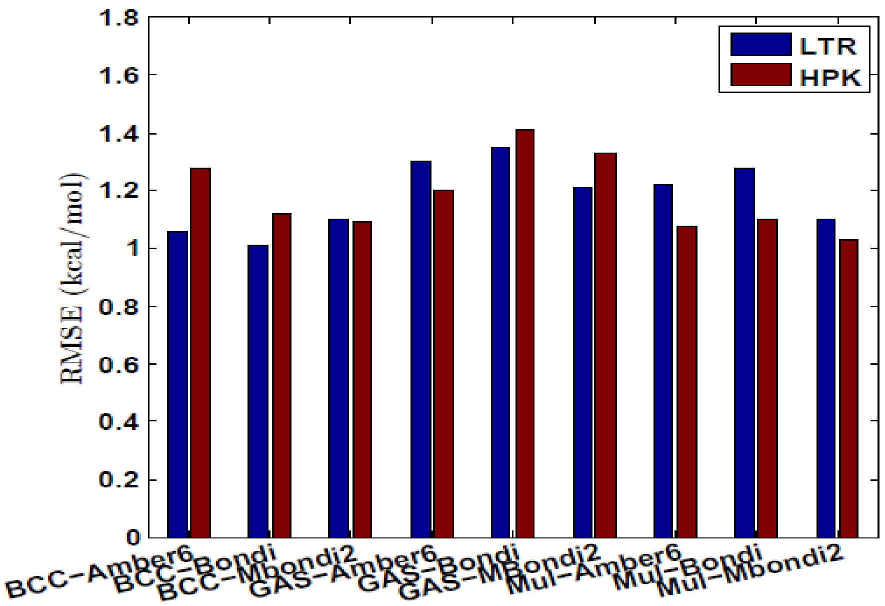


Figure 7.9: Illustration of prediction RMSEs obtained with different molecular parametrizations by the LTR and HPK models for SAMPL4 test set.

complex and challenging molecules in SAMPL1 and SAMPL2 test sets. Furthermore, the present LTR solvation prediction model is much less sensitive to different charge and radius parametrizations. Nevertheless, contrary to the small MAEs found in the leave-one-out tests, these errors amplify a lot in the blind prediction of SAMPLx test sets, particularly for SAMPL1 and SAMPL2 test sets. Possible explanations for this phenomenon are the complexity of molecules and the lack of physically and chemically similar molecules in our database. We also point out that in the LTR predictions, large RMSEs and MEs occur simultaneously, which indicates that large RMSEs might come from biased predictions. This phenomenon is under our further investigation.

7.4 Concluding Remarks

This work proposed a Learning-To-Rank (LTR) model for solvation free energy prediction. More precisely, this approach combines an LTR based nearest neighbor searching methodology with a local hyperplane learning procedure for solvation free energy prediction. The proposed model is inspired by our previous Hybrid Physical and Knowledge (HPK) model [256] on the nearest neighbor parametrization for solvation free energy prediction, in which the nearest neighbor was detected according to the simple cosine similarity between the molecular atomic features. Our previous attempt on the nearest neighbor approach motivates a basic assumption utilized in this work, i.e., similar molecules have similar solvation free energies. In machine learning terminology, our previous nearest neighbor search method can be regarded as an unsupervised learned method. With the similar molecules having similar solvation free energies, the nearest neighbor searching problem can be cast into a supervised learning problem. As a result, the nearest neighbor quality can be improved dramatically, which further improves the accuracy of solvation free energy prediction. The present LTR method can be considered as an enhanced nearest neighbor searching method.

To implement our new supervised LTR model, we first partition molecules into several groups according to their chemical compositions. Each group is regarded as a query in the LTR terminology. The query construction is of fundamental importance to make our model practical, since during the molecular ranking, we utilize the assumption that close solvation free energies imply similar molecules, which is generally not true without a proper query construction. Another fundamental assumption of this method is that there exists a feature vector that can uniquely characterize and distinguish one molecule from another. Obviously, the construction of the feature vector is of crucial importance to the performance of the present model. We utilize atomic features, such as atomic charge, dipole, reaction field energy, etc., evaluated by quantum mechanics, polarization theory and Poisson-Boltzmann theory. A state-of-the-art listwise LTR algorithm, LambdaMART, is adopted for training the LTR model. By using this algorithm, the quality of the nearest neighbor search improves significantly, which is supported from the fact that the difference between the solvation free energies of a target molecule and its neighbors decreased dramatically. Furthermore, we assume that molecular solvation free energy is a function of molecular features. Based on this assumption, we develop a regularized least square based local hyperplane learning algorithm for solvation free energy prediction. Highly accurate solvation free energy prediction is confirmed by both the leave-one-out test over 668 solvation molecules and blind prediction of five SAMPL test sets, namely, SAMPL0, SAMPL1, SAMPL2, SAMPL3 and SAMPL4.

A part from providing the state-of-the-art solvation free energy predictions, the proposed LTR model is much less parametrization dependent. That is, the LTR model is less sensitive to the parametrization of charges and radii in the solvation free energy prediction. This robust property is especially evident from the accurate and robust solvation free energy prediction of complex molecules, for which our previous HPK predictions can differ as large as 10 kcal/mol with different charges and radius parameterizations. A major reason leads to this robust in LTR based solvation free energy prediction is that in our basic assumption, solvation free energy is modeled as a unity, instead of isolated polar and nonpolar components. This treatment makes the present model robust to the molecule parametrization, and avoids the error propagation from polar modeling to nonpolar modeling. In our earlier HPK model, inappropriate atomic charge or radii assignments can lead to a huge electrostatic error, which propagates to the nonpolar solvation free energy prediction.

This work is our first attempt in developing an advanced machine learning based model for solvation free energy prediction. This model can be improved in a number of ways. One improvement is about query construction based on molecular element types. We believe that a more sophisticated query construction can further improve the accuracy of the nearest neighbor searching. Another potential improvement is a better feature selection. For example, one can select features according to their local correlations with the solvation free energies in a given query. The other improvement can be achieved through better feature design and more accurate feature evaluations. Many atomic features were computed via DFT in the present work. We believe that some other advanced quantum methodologies for atomic charge, dipole, and quadrupole calculations will significantly improve our prediction. The advantage of the DFT based polarable Poisson model has been noticed in our previous work [256]. Therefore, some improvements in the reaction field energy can be valuable as well. Overall, we believe that with a better set of molecule descriptors, molecular parametrization, and molecular partition, the proposed LTR based solvation free energy prediction can be further improved. The application of the proposed approach to protein-ligand binding is under our consideration.

Chapter 8

Protein Ligand Binding Free Energy Modeling

8.1 Introduction

Designing efficient drugs for curing diseases is of essential importance for the new century's life science. Indeed, one of the ultimate goals of molecular biology is to understand the molecular mechanism of human diseases and to develop efficient side-effect-free drugs for disease curing. Nevertheless, the drug discovery procedure is extremely complicated, and involves many scientific disciplines and technologies. As a brief summary, the drug discovering contains the following seven major steps [22], namely, i) Disease identification; ii) Target hypothesis, i.e., the activation or inhibition of drug targets (usually proteins within the cell) is thought to alter the disease state; iii) Screening potential principle compounds that will bind to the target; iv) Optimizing the identified compounds with respect to their structural characteristics in the context of the target binding site; v) Preclinical test, both *in vitro* and *in vivo* tests will be performed; vi) Clinical trials to determine their bioavailability and therapeutic potential; and vii) Optimizing chemical's efficacy, toxicity, and pharmacokinetics properties. Typically, the whole cost of a new drug development is estimated to be more than one billion dollars with more than ten years' group efforts [283]. This large amount of cost mostly comes from unsuitable chemical compounds that are used in the preclinical and clinical testing [2]. In terms of economical drug design, sophisticated and accurate computer aided compound screening methods become extremely important. Virtual screening (VS) methodologies focus on detecting a small set of highly promising candidates for further experimental testing [215]. Docking is one of the most important VS methodologies and is widely used in the Computer Aided Drug Design (CADD). It is a two-stage protocol [10]. The first step is sampling the ligand binding conformations, which determines the pose, orientation, and conformation of a molecule as docked to the target's binding site [28]. The second stage is protein-ligand binding affinity scoring. With the development of Molecular Dynamics (MD), Monte Carlo (MC), and Genetic Algorithm (GA) for pose generation, the sampling problem is relatively well resolved [265, 146, 187]. A major remaining challenge in achieving accurate docking is the development of accurate scoring functions for diverse protein ligand complexes. One of the most important open problems in computational bioscience is the accurate prediction of the binding affinities of a large set of diverse protein-ligand complexes [10]. A desirable goal is to achieve less than 1 kcal/mol Root Mean Square Error (RMSE) in the prediction.

Since the pioneer work in the 1980s and 1990s, the study of the scoring function and sampling techniques has been blooming in the CADD community [143, 63, 97, 131]. In a recent review paper, Liu and Wang classify the existing popular scoring functions into four categories [158], namely, i) Force-field based or physical based scoring functions; ii) Empirical or regression based scoring functions; iii) Potential of the Mean Force (PMF) or knowledge based scoring functions; and iv) Machine learning based scoring functions. Physics based scoring functions provide some of the most accurate and detailed description of the protein and ligand molecules in the solvent environment. Typical models that belong to this category are Molecular Mechanics Poisson-Boltzmann Surface Area (MM PBSA) and Molecular Mechanics Generalized-Born Surface Area (MM GBSA) [139, 93] with a given force field parametrization of both solvent and solute molecules, like AMBER or CHARMM force fields [258, 164, 274]. In this framework, the binding free energy is modeled as a superposition of four parts: van der Waals (vdW), electrostatics interactions between protein and ligand, the hydrogen bonding, and solvation effects. In addition to MM PBSA and MM GBSA, several other prestigious scoring functions also belong to this group, including COMBINE [193] and MedusaScore [277]. Physical based scoring functions are a class of dynamically improved methods, the VS can become more accurate with the further development of more advanced and comprehensive molecular mechanics force fields. Plenty of improvements has already been done for improving the accuracy of these scoring functions, such as QM/MM multiscale coupling [229] and polarizable force fields [198]. Empirical or regression based scoring functions, usually also called Multiple Linear Regression (MLR) scoring functions, typically model the protein-ligand binding affinity contributed from vdW interaction, hydrogen bonding, desolvation, and metal chelation [287]. Several parameters are introduced in each of the above term, the scoring function is obtained by using the existing protein-ligand binding information to train these parameters in the given binding affinity function. Many other existing scoring functions also belong to this category, e.g., PLP [245], ChemScore [74], and X-Score [264], etc.

A recent study on a congeneric series of thrombin inhibitors concludes that free energy contributions to protein-ligand binding are non-additive, showing some theoretical deficiencies of the MLR based scoring functions [16]. Machine learning algorithms do not explicitly require a given form of the binding affinity to its related items, and thus do not require the additive assumption of energetic terms. Many machine learning based scoring functions are proposed in the past few decades. These methods apply Quantitative Structure-Activity

Relation (QSAR) principles to the prediction of the protein-ligand binding affinity. Representative work along this line is the Random Forest (RF) based scoring function, RF-Score [152]. In RF-Score, the random forest is selected as the basic regressor instead of the classical MLR which is restricted to the pre-defined linear form of the binding affinity function. By utilization of the features calculated from the existing scoring functions, it achieves highly accurate five-fold cross validation results on the PDBBind v2013 refined set. Prediction results on the PDBBind v2007 core set further confirms the accuracy of the RF-Score [152]. Many other machine learning tools are utilized as the main skeletons of the scoring functions, like Support Vector Regression (SVR) [135], multivariate adaptive regression (MARS), k-Nearest Neighbors (kNN), Boosted Regression Trees (BRT), etc [4]. The blooming of the big data approaches and more accurate descriptors characterization of the protein-ligand binding effects have made machine learning type of scoring function full of vitality in CADD. Machine learning based scoring functions can make continuous improvement through both advance in physical protein-ligand binding descriptors and discovery of new machine learning techniques.

Another important class of scoring functions is PMF based. This category of scoring functions is based on the simplified statistical mechanics theory in which the protein ligand binding affinity is modeled as the sum of pairwise statistical potentials between protein and ligand atoms. The major merit of the PMF type of scoring functions is their simplicity in both concept and computation. This simplified physical model captures major physical principles behind the protein ligand binding. In Knowledge-based and Empirical Combined Scoring Algorithm (KECSA), the binding affinity between protein and ligand are modeled by 49 pairwise modified Lennard-Jones types of potentials between different types of atoms [288]. Through a large number of training instances, the functional form of all these pairwise interaction potentials can be determined. Effective ligand binding conformation sampling procedure can also be incorporated into this theoretical framework [289]. There are many other interesting developments in the PMF based scoring functions, e.g., PMF [180], DrugScore [243], and IT-Score [116].

Essentially, the major purpose of the scoring function is to find the relative order of binding affinities of candidate chemicals to the target binding site. These ranked results are further used for the preclinical test in real drug design procedure. From this point of view, the scoring function development turns out to be the development of ranking methods. Many existing scoring functions are also developed from this perspective. For example, Learning To Rank (LTR) algorithms have been used to develop various scoring functions, including PTRank, RankNet, RankNet, RankBoost, ListNet, and AdaRank [283, 267, 2, 247]. Compared to other machine learning or simple MLR based scoring functions, the advantages of ranking based scoring functions are two-fold. First, they are applicable to identifying compounds on novel protein binding sites where no sufficient data available for other machine learning algorithms. Second, they are suitable for the case that binding affinities are measured in different platforms since ranking can be more focused on relative order [283].

In this work, we propose a Feature Functional Theory-Binding Predictor (FFT-BP) for the blind prediction of binding affinity. The FFT-BP is constructed based on three assumptions, i.e., i) representability assumption: there exists a microscopic feature vector that can uniquely characterize and distinguish one molecular complex from another; ii) feature-function relationship assumption: the macroscopic features, including binding free energy, of a molecule or complex is a functional of microscopic feature vectors; and iii) similarity assumption: molecules with similar microscopic features have similar macroscopic features, such as binding free energies. FFT-BP has three distinguishing traits. A major trait of the proposed FFT-BP is its use of microscopic features derived from physical models, including Poisson Boltzmann (PB) theory [218, 219, 208, 112, 92, 43, 248], nonpolar solvation models [226, 84, 83, 50, 246, 270, 46], and components in MM PBSA [139]. As such, electrostatic solvation free energy, electrostatic binding affinity, atomic reaction field energies, and Coulombic interactions are utilized to represent the electrostatic effects of protein-ligand binding. Atomic pairwise van der Waals interactions are employed to model the dispersion interactions between the protein and ligand. We also make use of atomic surface areas and molecular volume in our FFT-BP to describe hydrophobic and entropy effects of the protein-ligand binding process. Another trait of the present FFT-BP is its feature-function relationship assumption, which avoids the use of additive modeling of the total binding affinity by the direct sum of various energy components. The machine learning algorithm automatically rank the relative importance of each feature to the binding affinity. By utilizing the boosted regression tree type of algorithms for the ranking, our model can capture the nonlinear dependence of the binding affinity to each feature. The other trait of FFT-BP is its use of advanced LTR algorithm, the multiple additive regression tree (MART), for ranking the nearest neighbors via microscopic features. This approach allows us to further improve our method by incorporating the state-of-the-art machine learning techniques.

This chapter is structured as follows. In Section 8.2, we present the theoretical background of FFT-BP, which consists of four parts, basic assumptions, microscopic feature selection, MART algorithm and binding affinity function. In Section 8.3, we verify the accuracy and robustness of our FFT-BP by a validation set, a training set and three standard test sets involving a variety of diverse protein-ligand complexes. We show that FFT-BP delivers some of the best binding affinity predictions.

8.2 Theory and algorithm

In this section, we present FFT for binding free energy prediction. First, we discuss the basic FFT assumptions. Additionally, feature selections are based on physical models. Moreover, protein-ligand complexes are ranked from a machine learning algorithms, i.e., the MART ranking algorithm. Finally, we describe a linear regression algorithm for approximating the binding free energy based on features from nearest neighbors ranked by the MART algorithm.

8.2.1 Basic assumptions

Our FFT is based on three assumptions, including representability, feature-functional relationship and similarity. These assumptions are described below.

8.2.1.1 Representability assumption

Without lost of generality, we consider a total of N molecules or complexes $\{M_i\}_{i=1}^N$ with known names and geometric structures from related databases. One of FFT basic assumptions is that there exists an n-dimensional microscopic feature vector, denoted as $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i1}, \cdots, \mathbf{x}_{in})$ to uniquely characterize and distinguish the ith molecule or complex. Here the vector components include various microscopic features, such as atomic types and numbers, atomic charges, atomic dipoles, atomic quadrupole, atomic reaction field energies, electrostatic solvation or electrostatic binding free energies, atomic surface areas, pairwise atomic van der Waals interactions, etc.

For ith molecule or complex, apart from its n microscopic features, there are l macroscopic features, or physical observable $\mathbf{o}_i = (\mathbf{o}_{i1}, \mathbf{o}_{i1}, \cdots, \mathbf{o}_{il})$, such as density, pressure, boiling point, enthalpy of formation, heat of combustion, solvation free energy, pKa, pH, viscosity, permittivity, electrical conductivity, binding free energy, etc. We combine the microscopic and macroscopic feature vectors to construct an extended feature vector $\mathbf{v}_i = (\mathbf{x}_i, \mathbf{o}_i)$ for the ith molecule.

Extended feature vectors $\{\mathbf{v}_i\}_{i=1}^{N}$ span a vector space $\mathcal{V}$, which satisfies commonly required eight axioms for addition and multiplication, such as associativity, commutativity, identity element, and inverse elements of addition, compatibility of scalar multiplication with field multiplication, etc. Unlike the usual L_p space, the extended feature space does not have the notion of nearness, angles or distances. We therefore need additional techniques, namely, machine learning algorithms to study the nearness and distance between feature vectors. The selection of microscopic features depends on what physical or chemical prediction is interested. In our approach, we utilize microscopic features form related physical models. For example, for solvation and binding free energy prediction, we select features that are derived from implicit solvent models and quantum mechanics.

Based on our assumption, microscopic features along are able to characterize and distinguish molecules. In contrast, macroscopic features are used as the label in learning and ranking molecules for a given purpose. Therefore, for a given task, say binding free energy prediction, we do not include all the macroscopic features in the feature vector $\mathbf{o}_i$. We only select $\mathbf{o}_i = (\mathbf{o}_{i1}) = \Delta G_i, \forall i = 1, \cdots, N$, where $\{\Delta G_i\}$ are known binding free energies from databases. The resulting extended vector is used for the binding free energy prediction.

8.2.1.2 Feature-function relationship assumption

In FFT, a general feature-function relationship is assumed for the jth physical observable $\mathbf{o}_j$ of target molecule A

$$\mathbf{o}_{\mathrm{A}j} = f_j(\mathbf{x}_{\mathrm{A}}, \mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N), \tag{8.2.1}$$

where f_j is an unknown function modeling the jth physical observable of molecule A and $\mathbf{x}_{\mathrm{A}}$ is the microscopic feature vector of the target molecule A. This relation applies to the prediction of various physical and chemical properties. In the present application, we are interested in the prediction of binding free energies for a set of diverse protein-ligand complexes. We construct a feature space for the training set and the binding free energy of target molecular complex AB can be given as a functional of extended feature vectors

$$\Delta G_{\mathrm{AB}} = f_{\mathrm{binding}}(\mathbf{x}_{\mathrm{AB}}, \mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N) \tag{8.2.2}$$

where ΔG_{AB} is the binding free energy of molecular complex AB, and f_{binding} is an unknown functional for modeling the relationship between binding free energy and extended features. Obviously, the determination of f_{binding} is a major task of the present work.

8.2.1.3 Similarity assumption

In the FFT, we assume that molecules with similar microscopic features have similar macroscopic features, or physical observable. In the present application, we assume that protein complexes with similar microscopic features will have similar binding free energies. This assumption provides the basis for utilizing supervised machine learning algorithms to rank protein-ligand complexes.

In our earlier HPK model, we assume that molecules with similar features having the same set of parameters in a physical model. As a result, solvation or binding free energies are still computed based on a physical model, while a machine learning algorithm is used to find out the nearest neighbors for modeling the physical parameters. In the present FFT, the binding free energy is not modeled by a physical model directly. However, the microscopic features are constructed from physical models.

8.2.2 Microscopic features

In physical models, such as MM PBSA and MM GBSA, the protein ligand binding affinity is given by the combination of molecular mechanics energy, solvation free energy, and entropy term

$$\Delta G = \Delta E_{\text{MM}} + \Delta G_{\text{solv}} - T\Delta S, \tag{8.2.3}$$

where ΔE_{MM}, ΔG_{solv}, and $T\Delta S$ are the molecular mechanics energy, solvation free energy, and entropy terms, respectively. Further, the molecular mechanics energy can be decomposed as E_{Covalent}, which is the sum of bond, angle, and torsion energy terms, and $E_{\text{Noncovalent}}$, which includes the van der Waals term and a Coulombic term E_{Coul} [104]. Equation (8.2.3) is used as a guidance for the feature selection in our FFT-Score model.

8.2.2.1 Reaction field features

Molecular electrostatics is of fundamental importance in the protein solvation and binding processes [219, 112, 92]. In this work, we use a classical implicit solvent model, the PB theory, for modeling the molecular electrostatics in the solvent environment. This model is used for two purposes. On the one hand, the solvation effects during the protein ligand binding will be modeled via this theory. On the other hand, the electrostatic contribution to the protein ligand binding affinity is computed based on this model, as well.

For simplicity, we consider the linearized PB model in the pure water solvent, which is formulated as the following elliptic interface problem in mathematical terminology. The governing equation is given by

$$-\nabla \cdot (\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})) = \sum_{i=1}^{N_m} Q_i \delta(\mathbf{r} - \mathbf{r}_i), \tag{8.2.4}$$

with the interface conditions

$$[\phi]|_\Gamma = 0, \tag{8.2.5}$$

and

$$[\epsilon\phi_{\mathbf{n}}]|_\Gamma = 0, \tag{8.2.6}$$

where ϕ is the electrostatics potential over the whole solvent solute domain, Q_i is the partial charge located at $\mathbf{r}_i$ and $\delta(\mathbf{r}-\mathbf{r}_i)$ is the delta function at point $\mathbf{r}_i$. The permittivity function $\epsilon(\mathbf{r})$ is given by

$$\epsilon(\mathbf{r}) = \begin{cases} \epsilon_{\mathrm{m}} = 1, & \mathbf{r} \in \Omega^{\mathrm{m}} \\ \epsilon_{\mathrm{s}} = 80, & \mathbf{r} \in \Omega^{\mathrm{s}} \end{cases} \tag{8.2.7}$$

where Ω^{m} and Ω^{s} are solute and solvent domains, respectively. The two domains are separated by the molecular surface Γ.

The following Debye-Hückel type of boundary condition is imposed to make the PB model well posed

$$\phi(\mathbf{r}) = \sum_{i=1}^{N_m} \frac{Q_i}{4\pi\epsilon_{\mathrm{s}}|\mathbf{r}-\mathbf{r}_i|}, \text{ if } \mathbf{r} \in \partial\Omega, \tag{8.2.8}$$

where $\Omega = \Omega^{\mathrm{m}} \bigcup \Omega^{\mathrm{s}}$.

Molecular reaction field energy is computed by the following formula

$$\Delta G_{\mathrm{RF}} = \sum_{i=1}^{N_m} \Delta G_{\mathrm{RF}i} \tag{8.2.9}$$

where the ith atomic reaction field energy $\Delta G_{\mathrm{RF}i}$ is given by

$$\Delta G_{\mathrm{RF}i} = \frac{1}{2} Q_i (\phi(\mathbf{r}_i) - \phi_{\mathrm{home}}(\mathbf{r}_i)) \tag{8.2.10}$$

where ϕ_{home} is obtained through solving the PB model with $\epsilon(\mathbf{r}) = 1$ in the whole computational domain Ω. Note that atomic reaction field energies $\Delta G_{\mathrm{RF}i}$ are used as features in our FFT based solvation model.

Here the reaction field energy gives a good description of the solvation free energy. In our earlier study on the solvation model, we found that reaction field energy related molecular descriptor provides a very accurate characterization of the solvation effects. The study of a large amount of small solute molecules demonstrates that by using these microscopic features in the solvation model, the predicted solvation free energy is in an excellent agreement with the experimental solvation free energy. For example, the RMSE of our leave-one-out test for a large database of 668 molecules is around 1 kcal/mol [250].

Note that in Eq. (8.2.9), the whole reaction field energy is regarded as the sum of atomic reaction field energies. In the PB calculation, the solute molecule is usually assumed to be a homogeneous dielectric continuum with a uniform dielectric constant, which is an inappropriate assumption, since atoms in different environments should have different dielectric properties [265]. For this reason, we select the atomic reaction field energy as a microscopic feature and let the machine learning algorithm to automatically take care the possible difference in dielectric constants.

8.2.2.2 Electrostatic binding features

By using the PB model, we can further obtain the electrostatics contribution to the protein-ligand binding affinity. The electrostatics binding free energy is calculated by

$$\Delta G_{\text{el}} = (\Delta G_{\text{RF}})_{\text{Com}} - (\Delta G_{\text{RF}})_{\text{Pro}} - (\Delta G_{\text{RF}})_{\text{Lig}} + \Delta G_{\text{Coul}}, \tag{8.2.11}$$

where ΔG_{el} is the electrostatics binding free energy between protein and ligand, $(\Delta G_{\text{RF}})_{\text{Pro}}$ and $(\Delta G_{\text{RF}})_{\text{Lig}}$ are the reaction field energies of the protein and ligand, respectively. Here ΔG_{Coul} is the Coulombic interaction between the two parts in the vacuum environment, which is computed as

$$\Delta G_{\text{Coul}} = \sum_{i,j} \frac{Q_i Q_j}{r_{ij}}, \tag{8.2.12}$$

where r_{ij} is the distance between two specific charges, and indexes i and j run over all the atoms in the protein and ligand molecules, respectively. The PB model is solved by our in-house software, MIBPB [295, 278, 90, 39], which is shown to be grid size independent. Its relative ranking orders of reaction field energy and binding free energy calculated with different grid sizes are consistent [252]. This numerical accuracy guarantees the preserving of relative ranking orders, which in turn avoids the influence on the prediction from numerical errors.

8.2.2.3 Atomic Coulombic interaction

Coulombic energy plays an important role in the molecular mechanics energy [167, 139, 104]. Coulombic energy calculation also depends on the dielectric medium. To this end, we considered the atomic Coulombic interactions in vacuum environment. Specifically, for the

ith atom in the protein molecule, we select the microscopic feature from atomic Coulombic energy as

$$(\Delta G_{\text{Coul}})_i = \sum_j \frac{Q_i Q_j}{r_{ij}}, \tag{8.2.13}$$

where the summation index j runs over all the atoms in the ligand molecule. The Coulombic energy associated with the atoms in the ligand molecules can be defined analogously.

8.2.2.4 Atomic van der Waals interaction

It was shown that van der Waals interactions play an important role in solvation analysis [83, 50, 246, 46, 248]. We expect that van der Waals interactions are essential to binding process as well. In this work, we consider the 6-12 Lennard Jones (LJ) interaction potential for modeling the van der Waals interactions

$$u_{ij}(\mathbf{r}_i, \mathbf{r}_j) = \epsilon_{ij} \left[\left(\frac{r_i + r_j}{||\mathbf{r}_i - \mathbf{r}_j||} \right)^{12} - 2 \left(\frac{r_i + r_j}{||\mathbf{r}_i - \mathbf{r}_j||} \right)^{6} \right], \tag{8.2.14}$$

where r_i and r_j are atomic radii of the ith and jth atoms, respectively. ϵ_{ij} measures the depth of the attractive well at $||\mathbf{r}_i - \mathbf{r}_j|| = r_i + r_j$. For features related to the van der Waals interactions, we select pairwise particles interactions as microscopic features for describing the van der Waals interactions between the protein and ligand. In these features, each atom type is collected together, well-depth parameters ϵ_{ij} are left as training parameters in the subsequent machine learning algorithm.

8.2.2.5 Atomic solvent excluded surface area and molecular volume

Molecular surface area and surface enclosed volume are usually employed in scaled-particle theory (SPT) to model the nonpolar solvation free energy [226, 197, 163] and/or entropy contribution to the protein ligand binding affinity. In our FFT-BP, the solvent excluded surface is employed for conformation modeling of the solvated molecule. The molecular surface area associated with each atom type and molecular volume are used as microscopic features. These features are also computed by our in-house software, ESES [157], in which a second order convergent scheme based on the level set theory and third order volume schemes are implemented. In ESES, the molecular surface area is partitioned into atomic surface areas based on the power diagram theory.

8.2.2.6 Summary of microscopic features

We consider microscopic features of a protein-ligand complex. For the protein molecule, microscopic features are selected from following types of atoms, i.e., C, N, O, and S. For the ligand molecule, atomic features are collected from C, N, O, S, P, F, Cl, Br, and I. Here we drop features from hydrogen atoms (H) since the positions of these atoms are not typically given in original X-ray crystallography data, their information itself may not be accurate. Coincidentally, this selection of representative effective atoms is consistent with that of some other existing scoring functions, e.g., Cyscore [33], AutoDock Vina [240], and RF-Score [9]. In our model, we collect atomic reaction field energies, molecular reaction field energy, atomic van der Waals and Coulombic interactions, atomic surface areas, and molecular volume as the building block of feature space. Due to the fact that binding is a dynamical process, the change of the atomic reaction field energies, atomic surface areas, and molecular volumes between the bounded and unbounded states are selected as microscopic features as well.

8.2.3 MART ranking algorithm

In this subsection, we briefly introduce the MART algorithm, and describe the application of this algorithm to protein ligand binding affinity scoring. MART is a list-wise LTR algorithm, for a given training set with feature vectors and associated ranking order (here we simply using the protein-ligand binding affinity as this label value), it trains a function that optimally simulates the relation between features and labels. When applied to a protein-ligand complex in the test set, this trained function acts on the corresponding features and gives a predicted value. The predicted value reflects the binding affinity of the complex in the test set. In the web-search community, LambdaMART is one of the state-of-the-art LTR algorithms, here LambdaMART is a coupling of Lambda and MART. Compared to the classical MLR model for training functions that link features and labels, MART can capture the nonlinear relationship. Furthermore, compared to most neuron network based algorithms, it is more efficient. MART also named GBDT (gradient boosting decision tree) is a very efficient ensemble method for regression. Meanwhile, due to the boosting of the weaker learners (usually quite simple models like decision tree), the over-fitting problem can be avoided effectively. The principles of the GBDT are summarized as following:

- For the training set, GBDT successively learns the weak learners, and each weak learner is a regression tree with quite a few levels for fitting the residual of the previous forest compared to the training set. This procedure starts from a regression tree for fitting the training set, and the regression tree is added into the forest gradually. Each succeed regression tree is used for fitting the residual of the previous forest.

- Instead of counting the whole contribution from each regression tree, shrinkage is adopted, which is a weight of the regression tree. This weight is obtained through solving an optimization problem via the simple line searching algorithm.

- Weighted contributions from the whole regression trees are presented in the final scoring function, which is the boosting of simple regression trees. Due to the simplicity of each regression tree, the over-fitting problem can be bypassed efficiently.

In summary, the MART itself learns a function between features and the binding free energy through the training set. In the testing step, this function assigns a predicted binding affinity to each sample in the testing set, and the ranking position of a given sample is determined through the obtained score. This ranking method is significantly different from the classical pairwise approaches, e.g., RankSVM [128, 129, 144], where ranking is based on the pairwise comparison between all sample pairs in the training set. The major drawback of these approaches is that they assumes the same penalty for all pairs. In contrast, we only care about a few top ranking results for a given query in most applications. For more comprehensive and mathematical description of the MART, reader is referred to the literature [26, 78]. Many other LTR algorithms can be used in our framework as well, like LambdaMART [26, 78], ListNet [34], etc.

8.2.4 Method for binding affinity prediction

In this subsection, we discuss the FFT prediction of the binding free energy of a given target protein-ligand complex AB. Based on our assumption that binding free energy is a functional of feature vectors, we construct a feature function around the target molecular complex and use it to predict the binding free energy. Even though the exact form of the function between feature and binding affinity is unknown, locally it can be approximated by a linear function. In other words, locally we assume the binding affinity is a linear function of the microscopic feature vector.

The importance of various features can be ranked automatically during the machine learning procedure, and thus the number of influential features (n) can be reduced by selecting features of top importance to represent the binding affinity. We assume that target molecular complex AB is characterized by its feature vector $\mathbf{x}_{\mathrm{AB}} = (\mathbf{x}_{\mathrm{AB}1}, \mathbf{x}_{\mathrm{AB}2}, \cdots, \mathbf{x}_{\mathrm{AB}n})$, where n is the dimension of the microscopic feature space, i.e., the space of all microscopic feature vectors. We also assume that by using the LTR algorithm, we can find top m nearest neighbors from the training set. The extended feature vectors of these nearest neighbor complexes are given by $\{\mathbf{v}_i = (\mathbf{x}_i, \Delta G_i)\}_{i=1}^m$. In general, the dimension of the feature space is much larger than the number of nearest neighbors used, i.e., $m \ll n$. Therefore, the direct least square approach may lead to over-fitting. To avoid over-fitting, we utilize a Tikhonov regularization based least square algorithm for training the binding affinity function. From the extended feature vectors, we can set up the following set of equations

$$\begin{pmatrix} \Delta G_1 \\ \Delta G_2 \\ \vdots \\ \Delta G_m \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} + \begin{pmatrix} b \\ b \\ \vdots \\ b \end{pmatrix}, \tag{8.2.15}$$

where $w_i = w_i(\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_m)$ and $b = b(\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_m)$ define the function for ΔG_i. By the similarity assumption, the same functional form can be used for target complex AB. For further derivation, we rewrite Eq. (8.2.15) as

$$\Delta \mathbf{G} = \mathbf{x}\mathbf{w} + b\mathbf{1}, \tag{8.2.16}$$

where $\Delta\mathbf{G} = (\Delta G_1, \Delta G_2, \cdots, \Delta G_m)^T$, $\mathbf{w} = (w_1, w_2, \cdots, w_n)^T$, $\mathbf{1}$ is an m-dimensional column vector with all elements equaling 1, and matrix $\mathbf{x}$ is given by

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}.$$

To avoid over-fitting, we add an L_2 penalty to the weight vector $\mathbf{w}$, and solve Eq. (8.2.16) as an optimization problem

$$\min_{\mathbf{w},b} ||\Delta\mathbf{G} - \mathbf{x}\mathbf{w} - b\mathbf{1}||_2^2 + \lambda||\mathbf{w}||_2^2 := \min_{\mathbf{w},b} \mathbf{F}, \tag{8.2.17}$$

where λ is a regularization parameter and is set to 10 in this work, and $|| * ||_2$ denotes the L_2 norm of the quantity $*$.

By solving $\frac{\partial \mathbf{F}}{\partial \mathbf{w}} = 0$, we have

$$\mathbf{w} = \left(\mathbf{x}^T\mathbf{x} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{x}^T\Delta\mathbf{G} - \mathbf{x}^T(b\mathbf{1})\right), \tag{8.2.18}$$

where I is an $m \times m$ identity matrix.

To determine b from Eq. (8.2.17), we relax $b\mathbf{1}$ to an arbitrary vector such that $\mathbf{b} = (b_1, b_2, \cdots, b_m)^T$. By solving $\frac{\partial \mathbf{F}}{\partial \mathbf{b}} = 0$, we have

$$\mathbf{b} = \Delta\mathbf{G} - \mathbf{x}\mathbf{w}. \tag{8.2.19}$$

An unbiased estimation of b is given by

$$b = \frac{\sum_{i=1}^{m} (\Delta\mathbf{G} - \mathbf{x}\mathbf{w})_i}{m}, \tag{8.2.20}$$

where $(\Delta\mathbf{G} - \mathbf{x}\mathbf{w})_i$ is the ith component of the vector $\Delta\mathbf{G} - \mathbf{x}\mathbf{w}$.

The optimization problem in Eq. (8.2.17) is solved by alternately iterating Eqs. (8.2.18) and (8.2.20), which is essentially an Expectation - Maximization (EM) algorithm.

After obtaining optimized weights $\mathbf{w}$ for the feature vector $\mathbf{x}$ and hyperplane height b, the binding free energy of target molecular complex AB can be predicted as

$$\Delta G_{\rm AB} = b + \sum_{i=1}^{n} w_i \mathbf{x}_{{\rm AB}i}. \tag{8.2.21}$$

Equation (8.2.21) can be regarded as a linear approximation of the binding free energy functional $\Delta G_{\rm AB} = f(\mathbf{x}_{\rm AB}, \mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_m)$.

Alternatively, we can also directly obtain the binding affinity of the target complex AB from the LTR ranking value if the ranking algorithm attempts to fit the target value. For general LTR algorithms, especially pairwise ranking algorithms, the direct use of the ranking score as a predicted binding affinity is not appropriate. However, the proposed protocol also applies to this scenario. These two approaches are compared in this present work.

8.3 Numerical results

In this section, we explore the validity, demonstrate the performance, and examine the limitation of the proposed FFT-BP. First, we describe datasets used in this work. Then, we examine whether FFT-BP's performance depends on protein clusters, where each cluster contains one specific protein and tens or hundreds of ligands. Our test on a validation set of 1322 protein-ligand complexes from 7 clusters indicates that the performance of the proposed FFT-BP does not depend on protein clusters. By using the same test set, we also study the impact of cut-off distance to FFT-BP prediction. Here cut-off distance refers to protein feature evaluation truncation distance. Protein atoms within the cut-off distance are allowed to contribute the atomic feature selection and calculation. To further benchmark the accuracy of the present FFT-BP, we carry out a five-fold cross validation on training set ($N = 3589$), which is derived from the PDBBind v2015 refined set [160]. Finally, we provide blind predictions on a benchmark set of 100 protein-ligand complexes [265], the PDBBind v2007 core set ($N = 195$) [9], and the PDBBind v2015 core set ($N = 195$) [160].

8.3.1 Dataset perparation

All data sets used in the present work are obtained from the PDBBind database [160], in which the PDBBind v2015 refined set of 3,706 entries was selected from a general set of 14,620 protein-ligand complexes with good quality, filtered over binding data, crystal structures, as well as the nature of the complexes [160]. Due to the feacture extraction, a pre-processing of data is required in the present method.

8.3.1.1 Datasets

This work utilizes one validation set ($N = 1322$), one training set ($N = 3589$), and three test sets ($N = 195, N = 195$ and $N = 100$).

8.3.1.1.1 Validation set ($N = 1322$) To explore the cluster dependence or independence and the optimal cut-off distance of the present FFT-BP, we select a subset of the

PDBBind v2015 refined set with 1322 complexes in 7 different clusters. Each cluster contains one protein and a large number, ranging from 93 to 333, of small ligand molecules.

8.3.1.1.2 Training set ($N = 3589$) We carry our FFT microscopic feature extraction of the PDBBind v2015 refined set via appropriate force field parametrization described below, which leads to a parametrized set of 3589 protein-ligand complexes. Whenever a test set is employed, its entries are carefully excluded from the training set of 3589 complexes.

8.3.1.1.3 Test sets Three test sets are standard ones described in the literature. PDB IDs of the training set and the validation set are given in the Supporting material.

The PDBBind v2015 core set of 195 benchmark-quality complexes is employed as a test. According to the literature, the PDBBind v2015 core set was selected with an emphasis on the diversity in structures and binding data. It contains 65 representative clusters from the refined set. For each cluster, it must have at least five protein-ligand complexes and three complexes, one with the highest binding constant, another with the lowest binding constant, and the other with a medium binding constant were selected for the PDBBind v2015 core set [160].

We also consider two additional test sets, *the PDBBind v2007 core set* of 195 complexes [48] and *the benchmark set* of 100 complexes [265] to benchmark the proposed FFT-BP against a large number of scoring functions.

When the training set ($N = 3589$) is applied to a test set, we first exclude all the overlapping entries between the training and the given test and re-train the training set for the specific test.

8.3.1.2 Data pre-processing

FFT-BP utilizes microscopic features, which requires appropriate feature extraction from the data set. Before the feature generation, structure optimization and force field assignment are carried out. Protein structures with corresponding ligand are prepared with the protein preparation wizard utility of the Schrödinger 2015-2 Suite [79, 213] with default parameters except filling the missing side chains. The protonation states for ligands are generated using Epik state penalties and the H-bond networks for the complex are further optimized using PROPKA at pH 7.0 [210, 189]. The restrained minimization on heavy atoms for the complex structures are finally performed with OPLS 2005 force field [132]. The atomic radii and charges for the complexes are parameterized by Amber tool14 [35]. For ligand molecules, charges are calculated by the antechamber module with AM1-BCC semi-empirical charge method and the atomic radii are assigned by using the mbondi2 radii set [122]. For protein molecules, radii and charges of each atom are parameterized by the Amber ff14SB general force field with tleap module [35].

Protein features are extracted with a cut-off distance. Specifically, we first find a tight bounding box containing the ligand, then extend feature generation domain along all directions around the box to a cut-off distance. We provide all the data involved in this work in the Supporting material, in which some protein-ligand structures that needs specific treatments are emphasized.

In the PDBBind database, the protein ligand binding affinity is provided in term of pK_d. We convert all the energy unit in the PDBBind database to kcal/mol. To derive the unit convert formula, one notes that

$$\Delta G = RT \ln k_d = -RT \ln K_{eq},$$

where ΔG is the Gibbs free energy, k_d is the disassociation constant, and R is the gas constant. Since $pK_d = -\log_{10} K_d$, then at the room temperature, $T = 298.15K$, one has the following relation between these two units

$$\Delta G = -1.3633 \, pK_d. \tag{8.3.1}$$

8.3.2 Validation

In this section, we explore the properties of FFT-BP and validate its performance. The following two important issues are examined in several existing scoring functions. The first issue is related to the protein-ligand binding affinity prediction of diverse multiple clusters, especially clusters with limited experimental data. Another issue is that a scoring method should be optimized with a cut-off distance in the feature extraction to maintain sufficient accuracy and avoid unnecessary feature calculations. In the existing work, the LTR based scoring functions can predict cross-cluster binding affinity well [283]. For the random forest and some other machine learning algorithms, one typically selects a cut-off distance of 12 Å, in the protein feature calculation [10].

In this work, we demonstrate the capability of the FFT-BP for the accurate cross-cluster binding affinity prediction. Additionally, we explore the optimal cut-off distance for FFT-BP feature extraction. Finally, since the accuracy of the FFT-BP predictions depends on the numbers of the nearest neighbors and top features, we investigate robustness of the proposed FFT-BP with respect to choice of the nearest neighbors and top features.

Two sets of protein-ligand complexes, i.e., the validation set ($N = 1322$) and the training set ($N = 3589$), are employed in this validation study.

8.3.2.1 Validation on the validation set ($N = 1322$)

Table 8.1: The RMSEs (kcal/mol) for the five-fold validation on the 7 clusters of the validation set and on the whole validation set ($N = 1322$) with 10 different cut-off distances in the feature extraction.

Test set	Group	5 Å	10 Å	15 Å	20 Å	25 Å	30 Å	35 Å	40 Å
Cluster 1	Group1	1.90	1.86	1.76	1.73	1.81	1.90	1.81	1.82
	Group2	2.07	2.15	2.38	2.23	2.35	2.25	2.21	2.24
	Group3	2.31	1.98	2.04	1.95	1.85	1.87	1.87	1.89
	Group4	1.89	1.75	1.58	1.63	1.63	1.67	1.62	1.66
	Group5	2.35	2.22	2.09	2.05	2.14	1.67	2.10	2.12
	Average	2.11	2.01	1.99	1.93	1.97	2.13	1.93	1.96
Cluster 2	Group1	1.39	1.33	1.31	1.32	1.39	1.43	1.46	1.42
	Group2	1.66	1.24	1.31	1.23	1.19	1.14	1.15	1.19
	Group3	1.39	1.28	1.14	1.21	1.28	1.31	1.37	1.37
	Group4	1.44	1.33	1.35	1.36	1.37	1.38	1.35	1.40
	Group5	1.53	1.44	1.38	1.49	1.36	1.37	1.38	1.33
	Average	1.49	1.33	1.34	1.32	1.32	1.33	1.34	1.35
Cluster 3	Group1	2.56	2.40	2.65	2.41	2.53	2.62	2.61	2.60
	Group2	2.07	2.13	2.08	2.10	2.11	2.11	2.11	2.09
	Group3	1.54	1.53	1.52	1.57	1.55	1.51	1.52	1.50
	Group4	1.82	1.75	1.70	1.71	1.64	1.68	1.70	1.72
	Group5	2.14	2.23	2.20	2.15	2.18	2.26	2.26	2.27
	Average	2.05	2.03	2.07	2.01	2.03	2.08	2.08	2.08

Table 8.1 (cont'd)

Test set	Group	5 Å	10 Å	15 Å	20 Å	25 Å	30 Å	35 Å	40 Å
Cluster 4	Group1	1.59	1.78	1.80	1.88	1.76	1.68	1.72	1.72
	Group2	1.41	1.47	1.53	1.25	1.34	1.39	1.37	1.34
	Group3	1.58	1.46	1.50	1.59	1.56	1.52	1.55	1.55
	Group4	1.91	1.76	1.76	1.87	1.83	1.84	1.80	1.78
	Group5	1.57	1.54	1.61	1.73	1.81	1.84	1.74	1.67
	Average	1.62	1.61	1.64	1.68	1.67	1.67	1.65	1.62
Cluster 5	Group1	2.01	2.43	1.83	1.64	1.60	1.65	1.67	1.69
	Group2	2.15	2.08	1.89	1.88	1.92	1.86	1.94	1.88
	Group3	2.52	2.26	2.54	2.42	2.41	2.37	2.39	2.40
	Group4	1.65	1.70	1.30	1.37	1.25	1.33	1.35	1.36
	Group5	3.18	2.87	2.89	2.49	2.59	2.54	2.56	2.67
	Average	2.39	2.31	2.18	2.03	2.03	2.02	2.05	2.07
Cluster 6	Group1	3.17	2.99	3.03	2.95	3.00	3.02	2.90	2.92
	Group2	2.09	1.83	1.83	1.82	1.91	1.83	1.88	1.86
	Group3	1.68	1.71	1.55	1.65	1.69	1.55	1.63	1.58
	Group4	1.73	1.69	1.55	1.60	1.60	1.51	1.58	1.58
	Group5	2.30	1.97	2.04	2.13	2.03	2.06	2.05	2.05
	Average	2.26	2.09	2.08	2.09	2.10	2.07	2.06	2.06
Cluster 7	Group1	1.83	1.97	2.16	1.93	1.68	1.66	2.01	1.90
	Group2	1.92	1.99	1.97	1.97	2.00	1.93	2.09	2.06
	Group3	1.68	1.69	1.45	1.39	1.35	1.39	1.44	1.51

Table 8.1 (cont'd)

Test set	Group	5 Å	10 Å	15 Å	20 Å	25 Å	30 Å	35 Å	40 Å
	Group4	2.27	2.11	2.13	1.91	2.14	2.14	2.39	2.36
	Group5	1.76	1.40	1.29	1.32	1.41	1.35	1.38	1.39
	Average	1.90	1.83	1.81	1.71	1.73	1.71	1.88	1.86
Average		1.90	1.88	1.87	1.83	1.84	1.84	1.85	1.85
Whole set	Group1	1.81	1.55	1.62	1.57	1.67	1.69	1.66	1.55
	Group2	1.63	1.76	1.62	1.69	1.64	1.67	1.55	1.71
	Group3	1.71	1.58	1.65	1.65	1.65	1.55	1.55	1.63
	Group4	1.73	1.62	1.65	1.57	1.56	1.53	1.78	1.57
	Group5	1.64	1.65	1.59	1.64	1.65	1.68	1.60	1.63
	Average	1.70	1.63	1.63	1.63	1.64	1.63	1.64	1.63

We validate the proposed FFT-BP on the validation set of 1322 complexes. We utilize the five-fold cross validation strategy to test the model and determine optimal cut-off distance. In this strategy, the validation set of 1322 complexes is randomly partitioned into five essentially equal sized subsets. Of the five subsets, a single subset is retained as the test set for testing the FFT-BP, and the remaining four subsets are used as training data. First, we run a coarse test with cut-off distance from 5 to 50 Å using 5 Å as the step size, which helps to determine the rough optimal cut-off distance. Second, we carry a refined search for the optimal the cut-off distance based on coarse test results with a step of size 1 Å. At a given cut-off size, we do the five-fold cross validation on the validation set of 1322 complexes, together with the

Table 8.2: The RMSEs (kcal/mol) for the five-fold test on the validation set ($N = 1322$) with FFT-BP calculated at different cut off distances.

Group	Cut-off distance										
	5 Å	6 Å	7 Å	8 Å	9 Å	10 Å	11 Å	12 Å	13 Å	14 Å	15 Å
Group1	1.81	1.68	1.80	1.58	1.61	1.55	1.49	1.62	1.50	1.60	1.62
Group2	1.63	1.67	1.61	1.79	1.67	1.76	1.63	1.65	1.76	1.72	1.62
Group3	1.71	1.68	1.57	1.65	1.61	1.58	1.80	1.62	1.56	1.71	1.65
Group4	1.73	1.56	1.46	1.58	1.64	1.62	1.74	1.58	1.55	1.70	1.65
Group5	1.64	1.57	1.82	1.57	1.60	1.65	1.46	1.56	1.59	1.59	1.59
Average	1.70	1.64	1.66	1.64	1.63	1.63	1.63	1.60	1.60	1.66	1.63

five-fold cross validation on each of 7 clusters. Table 8.1 lists the RMSEs on all the five-fold cross validation with cut-off distance 5 to 50 Å and step size 5 Å.

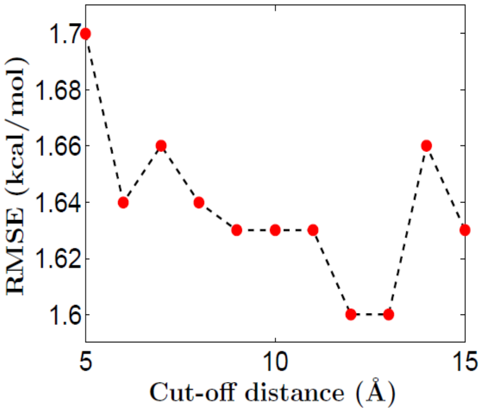


Figure 8.1: The prediction RMSE vs the cut-off distance.

Results in Table 8.1 indicate that: 1) Overall, prediction over the whole set of 1322 complexes gives better results than predictions on individual clusters. Therefore, the proposed method favors blind cross-cluster predictions. 2) According the results from the whole validation set tests, feature cuf-off distance at 10 Å is has reached its optimal value. This distance is actually consistent with the explicit solvent modeling in which a 10 Å cut-off distance is designed to account for long range electrostatic interactions. To better estimate the optimal cut-off distance, we carry out a more accurate searching in the range of 5 to 15 Å distance with a step size of 1 Å. Table 8.2 lists the RMSEs of the five-fold cross validation

Table 8.3: The RMSEs (kcal/mol) for the validation set ($N = 1322$) with different numbers of nearest neighbors and top features.

Number of	Number of top features									
nearest neighbors	5	10	15	20	25	30	35	40	45	50
1	1.60	1.60	1.60	1.61	1.61	1.61	1.61	1.61	1.61	1.62
2	1.60	1.60	1.61	1.61	1.61	1.61	1.61	1.62	1.62	1.62
3	1.60	1.59	1.60	1.70	1.66	1.68	1.71	1.70	1.69	1.70
4	1.61	1.57	1.62	1.71	1.70	1.72	1.73	1.70	1.85	1.83
5	1.61	1.60	1.67	1.74	1.75	1.74	1.75	1.73	1.78	1.77
6	1.62	1.61	1.68	1.79	1.80	1.81	1.81	1.88	1.85	1.85
7	1.61	1.61	1.65	1.78	1.77	1.78	1.78	1.81	1.82	1.82
8	1.62	1.62	1.65	1.74	1.76	1.76	1.77	1.78	1.80	1.81
9	1.62	1.61	1.65	1.74	1.75	1.76	1.76	1.76	1.78	1.77
10	1.62	1.62	1.73	1.74	1.79	1.80	1.82	1.82	1.88	1.90

on the whole validation set of 1322 complexes. These results show that 12 Å is the optimal cut-off distance in the searched solution space, which is consistent with that used in the RF-Score [10]. We plot the relation between the cut-off distance and prediction error in Fig. 8.1. In the rest of this work, the cut-off distance of 12 Å is utilized.

Finally, all the above predictions are based on the LTR ranking results. Alternatively, we can also carry out the prediction by using nearest neighbors and their associated features. We are interested to see the difference between these two approaches. To this end, we compute the binding affinities of five-fold results with different numbers of nearest neighbors and top features involved. Here top features are ranked by the LTR algorithm automatically according to their importance during the complex ranking. We list the top 50 important features to the protein ligand binding for the validation set in the Supporting material. We noted that the most important five features are the volume change, atomic Coulombic interaction of S atoms, area change of the C atoms in the protein and complex parts, and electrostatic binding free energy.

The RMSEs of the tests with different numbers of top features and nearest neighbors

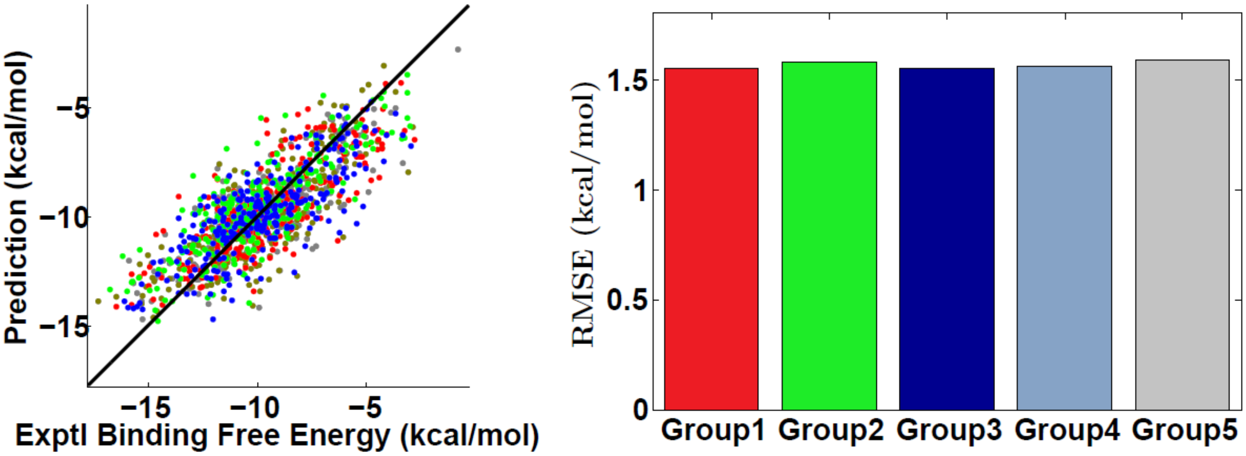


Figure 8.2: Five-fold cross validation on the validation set ($N = 1322$). Left chart: correlation between experimental binding affinities and FFT-BP predictions. Right chart: RMSEs for five groups. Here, RMSEs are 1.55, 1.58, 1.55, 1.56, and 1.59 kcal/mol for five groups, respectively. Overall Pearson correlation to the experimental binding affinities is 0.80.

involved are presented in Table 8.3. The optimal result is obtained when four nearest neighbor and 10 top features are utilized, with RMSE 1.57 kcal/mol. It is seen that when less than or equal to 10 top features are employed the prediction is quite accurate. However, with more features and more neighbors involved, the prediction become slightly worse. One possible reason is the reduced quality of the nearest neighbors involved for the prediction. Indeed, the neighbors that are not very close to the target molecule complex may make a large difference to the prediction of the target complex. This problem also motivates us to seek a better set of features for protein-ligand binding analysis.

Figure 8.2 depicts the optimal prediction results (Left chart) and RMSEs for each group (Right chart). It is seen that the RMSEs for all groups are almost the same, indicating the unbiased nature of five-fold cross-validation. The success of proposed FFT-BP is implied by the small RMSEs ($1.55 \sim 1.59$ kcal/mol) and the high overall Pearson correlation of 0.80.

Table 8.4: The RMSEs (kcal/mol) for the five-fold cross validation on the training set ($N = 3589$) with different number of nearest neighbors and top features.

Number of	Number of top features									
nearest neighbors	5	10	15	20	25	30	35	40	45	50
1	2.00	1.99	2.00	2.00	2.00	2.01	2.01	2.01	2.01	2.02
2	2.00	1.99	2.00	1.99	2.00	2.01	2.01	2.01	2.01	2.01
3	2.01	2.00	2.00	2.00	2.00	2.00	2.02	2.01	2.01	2.01
4	2.00	2.01	2.00	1.99	2.00	2.01	2.01	2.01	2.02	2.01
5	2.01	2.00	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.02
6	2.00	1.99	1.99	2.00	2.00	2.01	2.01	2.01	2.01	2.01
7	2.00	2.00	2.00	2.00	2.01	2.01	2.02	2.02	2.02	2.02
8	2.00	1.99	1.98	1.99	1.99	2.00	2.00	2.00	2.01	2.00
9	2.00	2.00	2.00	2.01	2.02	2.05	2.05	2.05	2.05	2.04
10	1.99	2.00	2.00	2.03	2.04	2.07	2.08	2.08	2.08	2.08

8.3.2.2 Validation on the training set ($N = 3589$)

We also consider the five-fold cross validation on our training set of 3589 complexes. We randomly divide this data set into five groups with 717, 718, 718, 718, and 718 complexes, respectively. In the five-fold cross validation, each time we regard one group of molecules as the test set without binding affinity data, and using the remaining four groups to predict the binding affinities of the selected test set.

Directly using the ranking score as the predicted binding affinity leads to RMSE 2.00 kcal/mol. Alternatively, we can predict binding affinities using the nearest neighbors and top features.

Table 8.4 shows the RMSEs for the five-fold cross validation test on the training set ($N = 3589$). The number of nearest neighbors is varied from 1 to 10 and the number of to features is changed from 5 to 50. The most important 50 features indicated from the LTR algorithm are provided in the Supporting material. Five top important features are volume change, electrostatics binding free energy, and van der Waals interactions between C-S, C-O and C-N pairs, respectively. The optimal prediction is achieved when 8 nearest

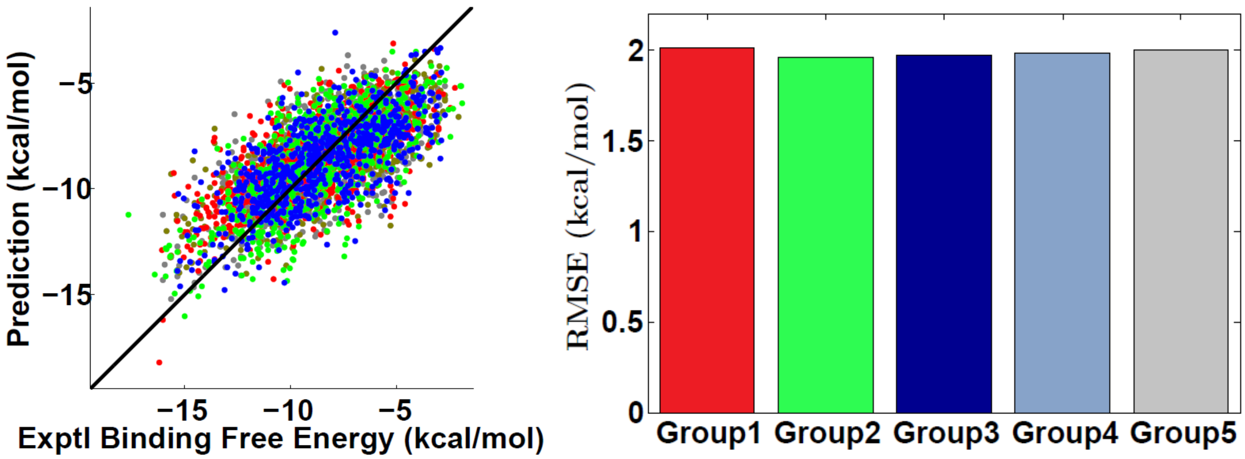


Figure 8.3: Five-fold cross validation on the training set (3589 complexes). Left chart: correlation between experimental binding affinities and FFT-BP predictions. Right chart: RMSEs for five groups. Here, RMSEs are 2.01, 1.96, 1.97, 1.98, and 2.00 kcal/mol for five groups, respectively. Overall Pearson correlation to the experimental is 0.70.

neighbors and top 10 features are used for binding affinity prediction, with the RMSE being 1.98 kcal/mol. Different numbers of nearest neighbors and top features basically give very consistent predictions. Compared to the five-fold test on the 1322 protein ligand complexes, the prediction errors on this set are much larger, which is partially due to the fact that structures in this test set is more complexes. For example, binding-site metal effects are presented without an appropriate treatment. We believe a better treatment of metal effects and a classification of ligand molecules would improve the FFT-BP prediction.

Figure 8.3 depicts the optimal prediction results (Left chart) and RMSEs for each group (Right chart). These tests demonstrate the following two facts. First, five-fold cross validation prediction is unbiased. The prediction results do not depends on the data itself and the RMSEs for all groups are almost at the same level. Second, when the protein-ligand complexes become diverse, the prediction becomes slightly worse due to the lack of similar complexes for certain clusters.

Table 8.5: The RMSEs (kcal/mol) of the FFT-BP for the benchmark test set ($N = 100$) with different numbers of nearest neighbors and top features.

Number of	Number of top features									
nearest neighbors	5	10	15	20	25	30	35	40	45	50
1	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
2	2.01	2.01	1.99	1.99	2.01	2.00	2.01	2.01	2.01	2.01
3	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.01	2.01	2.01
4	2.01	2.01	2.01	2.00	2.00	2.00	2.01	2.01	2.01	2.01
5	2.01	2.01	2.01	2.01	2.01	2.01	2.00	2.01	2.01	2.01
6	2.02	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.01
7	2.02	2.02	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.01
8	2.01	2.01	2.01	2.01	2.01	2.01	2.02	2.02	2.02	2.02
9	2.01	2.01	2.01	2.01	2.01	2.01	2.02	2.01	2.01	2.01
10	2.01	2.01	2.01	2.02	2.01	2.02	2.01	2.02	2.02	2.02

8.3.3 Blind predictions on three test sets

To further verify the accuracy of the FFT-BP, we perform the blind prediction on three benchmark test sets. The training set ($N = 3589$) that is processed from the PDBBind v2015 refined set is utilized for the training in all blind predictions. Due to the LTR algorithm used in our FFT-BP, the RMSE and correlation of our FFT-BP prediction would be around 0 kcal/mol and 1, respectively, had we include all the test set complexes in our training set. Therefore, in each blind prediction, we carefully exclude the overlapping test set complexes from the training set and re-train the training set with a reduced number of complexes.

8.3.3.1 Prediction on the benchmark set ($N = 100$)

First of all, we consider a popular benchmark set originally used by Wang *et al* [265]. This set contains 100 protein ligand complexes which involves a large variety of protein receptors. Originally this test set was used to test the performance of a large amount of well-known scoring functions and docking algorithms [265]. Recently, Zheng *et al* have utilized this test set to demonstrate the superb performance of their KECSA method [288]. In this work, we examine the accuracy and robustness of our FFT-BP on this benchmark test set.

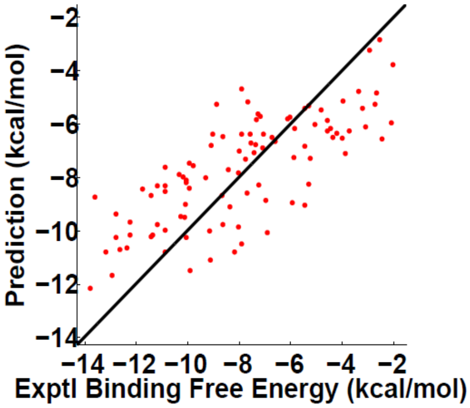


Figure 8.4: The correlation between experimental binding feree energies and FFT-BP predictions on the benchmark test set ($N = 100$) with the RMSE of 1.99 kcal/mol and the Pearson correlation of 0.75.

Directly using the ranking score as the predicted binding affinity leads to the RMSE of 2.01 kcal/mol and Pearson correlation coefficient of 0.75. Alternatively, we examine FFT-BP predictions using different numbers nearest neighbors and top features. Table 8.5 lists the predicted RMSEs for the benchmark set ($N = 100$). The numbers of nearest neighbors and tops features vary from 1 to 10 and 5 to 50, respectively. The most important 50 features indicated by the LTR algorithm are provided in the Supporting material. Five top important features are volume change, electrostatics binding free energy, van der Waals interaction between C-S and C-C pairs, and the complex's area change. The optimal prediction is reached when 2 nearest neighbors and top 15 or 20 features are used for binding prediction. The corresponding RMSEs and correlations for both cases are 1.99 kcal/mol and 0.75, respectively. Different numbers of nearest neighbors and top features basically give rise to very consistent predictions. We also note that the prediction errors for this 100 test set are very similar to those of the five-fold cross validation tests on our training set ($N = 3589$). This consistency indicates the robustness of the proposed FFT-BP in binding affinity predictions.

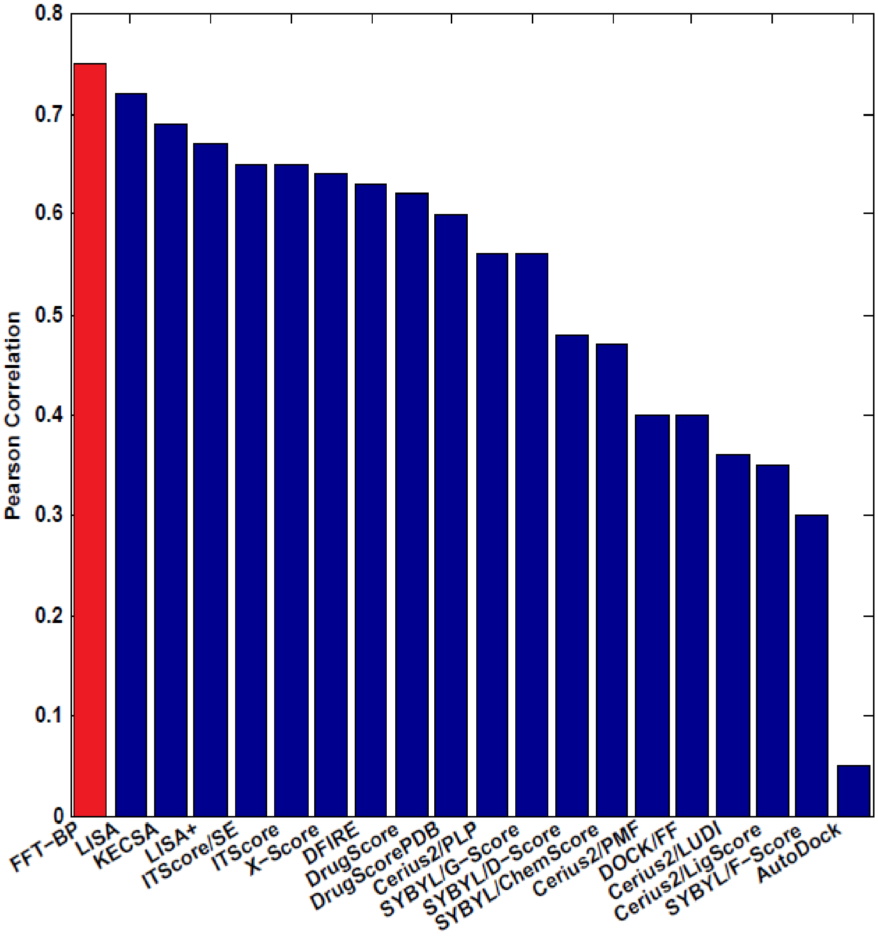


Figure 8.5: Performance comparison between different scoring functions on the benchmark test set ($N = 100$). The binding affinity comparisons was done for FFT-BP, and 19 well-known scoring functions, namely LISA, KECSA, LISA+ [288, 287], ITScore/SE [117], ITScore [116], X-Score [264], DFIRE [282], DrugScoreCSD [244], DrugScorePDB [96], Cerius2/PLP [88], SYBYL/G-Score [130], SYBYL/D-Score [171], SYBYL [74], Cerius2/PMF [180], DOCK/FF [171], Cerius2/LUDI [23], Cerius2[1], SYBYL/F-Score [201], and AutoDock [179].

Figure 8.4 illustrates the optimal prediction results compared to the experimental data. The RMSE and Pearson correlation coefficient are 1.99 kcal/mol and 0.75, respectively. This test set is a critical test set with diverse protein-ligand complexes and a wide range of experimental binding free energies. In our prediction, most predictions are quite appealing with less than 2 kcal/mol RMSE compared to experimental results.

Many outstanding scoring functions have been tested on this test set as summarized by

Table 8.6: The RMSEs (kcal/mol) of the FFT-BP for the PDBBind Core 2007 test set ($N = 195$) with different numbers of nearest neighbors and top features.

Number of	Number of top features									
nearest neighbors	5	10	15	20	25	30	35	40	45	50
1	2.10	2.10	2.10	2.10	2.10	2.10	2.10	2.10	2.10	2.10
2	2.09	2.09	2.09	2.09	2.09	2.09	2.10	2.10	2.10	2.10
3	2.10	2.10	2.10	2.09	2.09	2.09	2.09	2.10	2.09	2.09
4	2.09	2.09	2.09	2.09	2.09	2.10	2.09	2.09	2.09	2.10
5	2.09	2.10	2.09	2.10	2.11	2.10	2.11	2.11	2.11	2.10
6	2.10	2.08	2.09	2.08	2.09	2.08	2.10	2.09	2.10	2.10
7	2.09	2.09	2.10	2.10	2.11	2.12	2.12	2.12	2.12	2.12
8	2.10	2.10	2.11	2.10	2.10	2.10	2.10	2.10	2.10	2.10
9	2.09	2.10	2.10	2.10	2.10	2.10	2.11	2.11	2.11	2.11
10	2.10	2.10	2.09	2.10	2.10	2.10	2.11	2.11	2.11	2.11

Zheng *etr al* [288]. Here we also add our prediction to this list. As shown in Fig. 8.5, the performance of our FFT-BP is highlighted with red color. The performance of other 19 scoring functions are due to the courtesy of Ref. [288].

8.3.3.2 Prediction on the PDBBind v2007 core set ($N = 195$)

PDBBind v2007 core set ($N = 195$) which contains high quality data mainly aims for testing the performance of scoring functions [160]. It has been employed to study and compare many excellent scoring functions [151, 9, 150, 8, 48]. Here we also examine our FFT-BP test set. If we regard the ranking score itself as the predicted binding affinity, the RMSE is 2.09 kcal/mol for this test set, which is slightly larger than that of the five-fold test on the training set ($N = 3589$) and that of the benchmark test set ($N = 100$).

Table 8.6 shows the prediction RMSEs for PDBBind v2007 core set. We have varied the numbers of nearest neighbors and tops features from 1 to 10 and 5 to 50, respectively. The most important 50 features indicated by the LTR algorithm are provided in the Supporting material. The top five important features are volume change, electrostatics binding free energy, van der Waals interaction between C-O, the complex's area change, and van der Waals interaction between C-C. These features are basically consistent with those of all the previous test cases. The optimal FFT-BP prediction is found when 6 nearest neighbors and top 15 or 30 features are used for binding prediction, with RMSEs for both cases being 2.08 kcal/mol. Pearson correlation coefficient of 0.76 is consistent with the earlier finding from the test set of 100 complexes. It is found that for this test set, the predictions based on different numbers of nearest neighbors and top features do not differ much from each other.

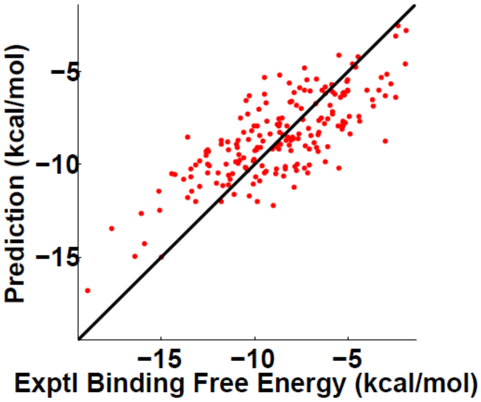


Figure 8.6: The correlation of between experimental binding feree energies and FFT-BP predictions on the PDBBind core 2007 ($N = 195$). The RMSE and Pearson correlation coefficient are 2.08 kcal/mol and 0.76, respectively.

Figure 8.6 illustrates the correlation between experimental binding free energies and the best predictions obtained by the FFT-BP. Obviously, there is a bias in the predicted binding affinities, which will be addressed in our future work.

Li *et al* have given a comparison of tests on the PDBBind v2007 core set ($N = 195$) using many outstanding scoring functions [151]. In this content, we also plot the performance of our FFT-BP in terms of Pearson correlation coefficient in Fig. 8.7. The FFT-BP correlation coeffiicent of 0.76 is highlighted with red color.

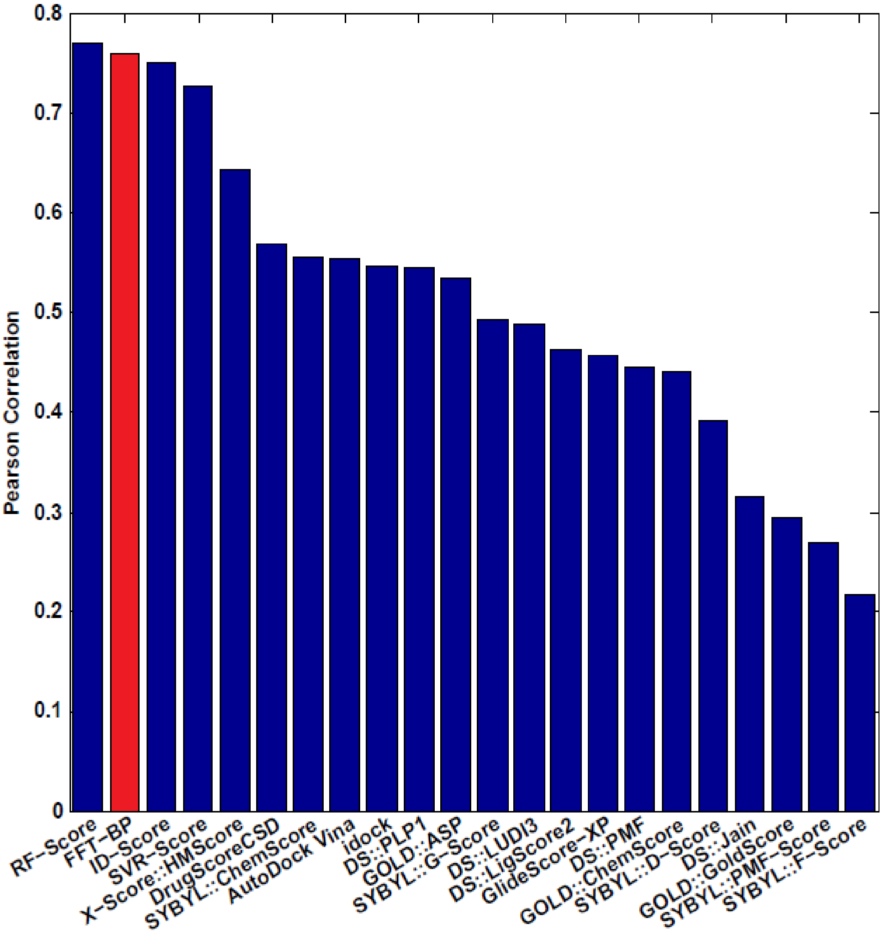


Figure 8.7: Performance comparison between different scoring functions on the PDBBind v2007 core set ($N = 195$). The performances of the other scoring function are adopted from the literature [151, 9, 150, 8, 48].

.

8.3.3.3 Prediction on the PDBBind v2015 core set ($N = 195$)

Finally, we perform a test on the PDBBind v2015 core set ($N = 195$), which contains high quality experimental data. This test set is also quite challenging due its diversity of 65 protein-ligand clusters and a wide binding affinity range. In a similar routine, we first consider the FFT-BP prediction with different numbers of neighbors and top features. Table 8.7 shows the RMSEs of FFT-BP for PDBBind v2015 core set ($N = 195$). The top 50 features are also listed in the Supporting material. The most important features are similar

Table 8.7: The prediction RMSEs (kcal/mol) for the PDBBind v2015 core set ($N = 195$) with different numbers of nearest neighbors and top features.

Number of	Number of top features									
nearest neighbors	5	10	15	20	25	30	35	40	45	50
1	1.95	1.95	1.95	1.95	1.95	1.95	1.95	1.95	1.95	1.95
2	1.95	1.94	1.95	1.95	1.95	1.95	1.95	1.95	1.96	1.96
3	1.94	1.94	1.95	1.95	1.95	1.95	1.95	1.95	1.95	1.95
4	1.94	1.94	1.93	1.93	1.94	1.94	1.94	1.95	1.95	1.95
5	1.95	1.95	1.92	1.94	1.95	1.95	1.96	1.95	1.95	1.95
6	1.95	1.95	1.95	1.95	1.96	1.96	1.95	1.96	1.95	1.94
7	1.95	1.93	1.93	1.94	1.95	1.97	1.95	1.95	1.95	1.95
8	1.95	1.95	1.95	1.96	1.96	1.97	1.97	1.96	1.95	1.95
9	1.94	1.94	1.94	1.94	1.94	1.95	1.95	1.94	1.94	1.94
10	1.95	1.95	1.95	1.95	1.95	1.94	1.94	1.94	1.94	1.94

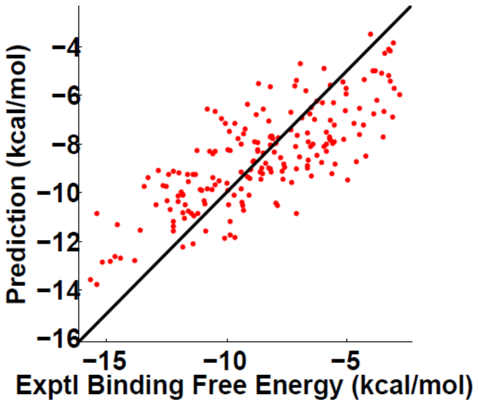


Figure 8.8: The correlation between experimental binding free energies and FFT-BP predictions on the PDBBind v2015 core set ($N = 195$). The RMSE and Pearson correlation coefficient are 1.92 kcal/mol and 0.78, respectively.

to those in previous tests, which indicates that the volume change, electrostatic binding free energy and van der Waals interactions are of fundamental importance to the protein-ligand binding. It is worth noting that the RMSEs of FFT-BP predictions are lower than those from earlier test sets. A possible reason is this data set is consistent with the training set as both obtained from the PDBBind 2015 refined set. Additionally, a better data quality might also contribute our better predictions. Our optimal prediction has the RMSE of 1.92 kcal/mol and Pearson correlation coefficient of 0.78, when 5 nearest neighbors and 15 features are used for the prediction.

Figure 8.8 plots the correlation between experimental binding free energies and FFT-BP predictions on the PDBBind v2015 core set ($N = 195$). Compared to the earlier two blind predictions, the prediction on this set is more accurate. However, similar to the behavior in two other test sets, the present prediction is biased. This issue will be studied in our future work.

8.4 Concluding remarks

In this work, we propose a new scoring function, feature functional theory - binding predictor (FFT-BP). FFT-BP is constructed based on three fundamental assumptions, namely, representability, feature-function relationship, and similarity assumptions. A validation set of 1322 complexes, a training set of 3589 complexes, and three test sets with 100, 195 and 195 complexes are considered in the present work to validate the proposed method, explore its utility, demonstrate its performance and reveal its deficiency. Extensive numerical experiments indicate that FFT-BP delivers some of the most accurate blind predictions in the field with the root-mean-square error around 2 kcal/mol and Pearson correlation coefficient around 0.76.

A major advantage of FFT-BP is that it extracts microscopic features from conventional implicit solvent models so that the validity of these physical models for binding analysis and prediction can be systematically examined. Consequently, the proposed FFT-BP can be improved via the improvement of our understanding on physical models. Another advantage of FFT-BP is that it provides a framework to systematically incorporates and continuously absorb advanced machine learning algorithms to improve its predictive power. The other advantage of FFT-BP is that it becomes more and more accurate as the existing binding database becomes larger and larger.

This work is our first attempt in exploring the mathematical modeling of the protein-ligand binding affinity. Our model can be further improved in several aspects. First, we have employed a very crude force filed parameterized of the Poisson model. More accurate Poisson-Boltzmann (PB) modeling, such as polarizable PB model, and feature extraction from more accurate quantum mechanics/molecular mechanics (QM/MM) models will improve the present FFT-BP. Additionally, we employ the MART algorithm for the molecules ranking. More sophisticated machine learning algorithms, such as deep learning, can potentially improve FFT-BP prediction, and eliminate the current prediction bias in test sets. Finally, a deficiency of the current model is that it neglects the metal effect on protein-ligand binding affinity. The incorporation of this effect into our model is under our investigation.

Chapter 9

Dissertation Contribution

In this chapter, we will summarize the contribution of this dissertation.

- In chapter 3, we proposed a novel parametrization method for the differential geometry based implicit solvent model, this work was published in [248]. The original solvation model was proposed by Wei[270], and first implemented by Chen et al [43]. In the work [43], parameters selection was considered. Nevertheless, the parametrization cannot give optimal prediction for polar molecule. I proposed a systematic parametrization scheme which incorporates the PDE analysis and convex optimization techniques. I implemented the parametrization scheme to their framework.

- In chapter 4, we presented an Eulerian solvent excluded surface. This is a collaborative work with Liu et al [157]. In this work, I focused on the surface validation.

- In chapter 5, we proposed an electrostatics potential interpolation scheme and implemented the Eulerian solvent excluded surface to the MIBPB Poisson Boltzmann software which is contributed from many people [39]. The software is shown to be of second order convergence in the numerical solution to the Poisson Boltzmann model. However, there are some deficiencies on this software in terms of robustness and highly accurate reaction field energy calculation. My work makes this software more robust and provides grid spacing almost independent reaction field energy calculation.

- In chapter 6, we developed a hybrid physical and knowledge based solvation prediction paradigm. This work can also be found in [255]. One major part is the polarizable Poisson model, in which I coupled the Poisson dielectric model with the Kohn Sham density functional theory. This coupling has been investigated by several work [262, 42], however, the solvent solute interface conditions have not been explicitly implemented. I incorporated the interface of the SIESTA software used by Chen et al [45] to the MIBPB software, and developed several schemes for the communication between this two software. This coupling can be regarded as the interface method based polarizable continuum model. Another contribution of this chapter is developing a framework for solvation free energy prediction. Furthermore, we introduced the distributed multipole analysis [227] for characterizing the solute molecules.

- In chapter 7, we delivered a protocol that couples the learning to rank method and the implicit solvent model for solvation prediction. Compared to some classical multiscale methods, where the microscopic models provide parametrization for macroscopic models. The new coupling provides an implicit parametrization strategy, which is demonstrated by treating the atomic reaction field energy through the big data approach. Large amount of numerical results show state-of-the-art accuracy and robustness for blind solvation free energy prediction. This work can also be found in [249].

- In chapter 8, we studied the protein ligand docking problem. More specifically, we proposed a novel protein ligand binding scoring function. This scoring function can be regarded as an extension of our learning to rank based solvation model to protein ligand binding scenario. The testing on the several benchmark test sets verify the accuracy of the proposed scoring function. This work can also be found in [254].

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] CERIUS2 ligandfit user manual; accelrys, inc.: San deigo, ca. pages 3–48, 2000.

[2] Shivani Agarwal, Deepak Dugar, and Shiladitya Sengupta. Ranking chemical structure for drug discovery: A new machine learning approach. *Journal of Chemical Information and Model*, 50:716–731, 2010.

[3] Li Anbang. Performance comparison of poissonboltzmann equation solvers delphi and pbsa in calculation of electrostatic solvation energies. *Journal of Theoretical and Computational Chemistry*, 13(13):1450040, 2014.

[4] Hossam M. Ashtawy and Nihar R. Mahapatra. A comparative assessment of ranking accuracies of conventional and machine-learning-based scoring functions for protein-ligand binding affinity prediction. *IEEE/ACM Transactions on computational biology and bioinformatics*, 9(5):1301–1313, 2012.

[5] Franz Aurenhammer. Voronoi diagrams a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3):345–405, 1991.

[6] N. A. Baker. Improving implicit solvent simulations: a Poisson-centric view. *Current Opinion in Structural Biology*, 15(2):137–43, 2005.

[7] Nathan A. Baker, David Sept, Michael J. Holst, and J. Andrew Mccammon. The adaptive multilevel finite element solution of the Poisson-Boltzmann equation on massively parallel computers. *IBM Journal of Research and Development*, 45(3-4):427–438, 2001.

[8] Pedro J. Ballester. Machine learning scoring functions based on random forest and support vector regression. *Proceedings of the 7th IAPR international conference on Pattern Recognition in Bioinformatics*, pages 14–25, 2012.

[9] Pedro J. Ballester and John B. O. Mitchell. A machine learning approach to predicting proteinligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175, 2010.

[10] Pedro J. Ballester, Adrian Schreyer, and L. Blundell Tom. Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? *Journal of Chemical Information and Model*, 54:944–955, 2014.

[11] D. Bashford and D. A. Case. Generalized Born models of macromolecular solvation effects. *Annual Review of Physical Chemistry*, 51:129–152, 2000.

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

[13] P. W. Bates, G. W. Wei, and S. Zhao. The minimal molecular surface. *arXiv:q-bio/0610038v1*, [q-bio.BM], 2006.

[14] P. W. Bates, G. W. Wei, and S. Zhao. The minimal molecular surface. *Midwest Quantitative Biology Conference*, Mission Point Resort, Mackinac Island, MI:September 29 – October 1, 2006.

[15] P. W. Bates, G. W. Wei, and Shan Zhao. Minimal molecular surfaces and their applications. *Journal of Computational Chemistry*, 29(3):380–91, 2008.

[16] B. Baum, L. Muley, M. Smolinski, A. Heine, D. Hangauer, and G. Klebe. Non-additivity of functional group contributions in protein-ligand binding: a comprehensive study by crystallography and isothermal titration calorimetry. *J. Mol. Bio*, 397(4):1042–1054, 2010.

[17] J. T. Beale and A. T. Layton. On the accuracy of finite difference methods for elliptic problems with interfaces. *Comm. Appl. Math. Comp. Sci.,*, 1:91–119, 2006.

[18] Axel D. Becke. Density-functional thermochemistry. iii. the role of exact exchange. *Journal of Chemical Physics*, 98(7):5648, 1993.

[19] D. Beglov and B. Roux. Solvation of complex molecules in a polar liquid: an integral equation theory. *Journal of Chemical Physics*, 104(21):8678–8689, 1996.

[20] C. A. S. Bergstrom, M. Strafford, L. Lazorova, A. Avdeef, K. Luthman, and P. Artursson. Absorption classification of oral drugs based on molecular surface properties. *Journal of Medicinal Chemistry*, 46(4):558–570, 2003.

[21] J. Blinn. A generalization of algebraic surface drawing. *ACM Transactions on Graphics*, 1(3):235–256, 1982.

[22] Joel R. Bock and David A. Gough. A new method to estimate ligand-receptor energetics. *Molecular and Cellular Proteomics*, 1(11):904–910, 2002.

[23] H. J. Bohm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *Journal of Computer-Aided Molecular Design*, 8:234–256, 1994.

[24] A. H. Boschitsch and M. O. Fenley. A Fast and Robust Poisson-Boltzmann Solver Based on Adaptive Cartesian Grids. *Journal of Chemical Theory and Computation*, 7:1524–1540, 2011.

[25] C. J. C. Burges, R. Ragno, and Quoc Viet Le. Learning to rank with nonsmooth cost functions. *Advances in Neural Information Processing Systems*, 19:193–200, 2007.

[26] Christopher J.C. Burges. From RankNet to LambdaRank to LambdaMART: An overview. *Microsoft Research Technical Report*, 82, 2010.

[27] C.J.C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. *In Proc. of ICML*, pages 89–96, 2005.

[28] B. D. Bursulaya, M. Totrov, R. Abagyan, and C. L. Brooks. Comparative study of several algorithms for flexible ligand docking. *Journal of Computer-Aided Molecular Design*, 17:755–763, 2003.

[29] H. Butt, L. Graf, and M. Kappl. *Physics and Chemistry of Interfaces.* Weinheim, Germany: Wiley-VCH., 2006.

[30] S. Cabani, P. Gianni, V Mollica, and L Lepori. Group Contributions to the Thermodynamic Properties of Non-Ionic Organic Solutes in Dilute Aqueous Solution. *Journal of Solution Chemistry*, 10(8):563–595, 1981.

[31] Sergio Cabani, Paolo Gianni, Vincenzo Mollica, and Luciano Lepori. Group contributions to the thermodynamic properties of non-ionic organic solutes in dilute aqueous solution. *Journal of Solution Chemistry*, 10:563 –595, 1981.

[32] Michael L. Cannolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612):709–713, 1983.

[33] Y Cao and L Li. Improved protein-ligand binding affinity prediction by using a curvature-dependent surface-area model. *Bioinformatics*, 30(12):1674–1680, 2014.

[34] Z. Cao, T. Qin, T. Y. Liu, M. F. Tsai, and F. Li. Learning to rank: From pairwise approach to listwise approach. *ICML*, 2007.

[35] D. A. Case, J. T. Berryman, R. M. Betz, D. S. Cerutti, T. E. Cheatham III, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K. M. Merz, G. Monard, P. Needham, H. Nguyen, H. T. Nguyen, I. Omelyan, A. Onufriev, D. R. Roe, A. Roitberg, R. Salomon-Ferrer, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R.M. Wolf, X. Wu, D. M. York, and P. A. Kollman. Amber 2015. *University of California, San Francisco*, 2015.

[36] D. S. Cerutti, N. A. Baker, and J. A. McCammon. Solvent reaction field potential inside an uncharged globular protein: A bridge between implicit and explicit solvent models? *The Journal of Chemical Physics*, 127(15):155101, 2007.

[37] Arghya Chakravorty, Lin Li, and Emil Alexov. Sensitivity test of poisson-boltzmann modeling of electrostatic component of the binding free energy: Effect of energy minimization, force field parameters and modeling protocols. Preprint.

[38] D. Chen, G. W. Wei, X. Cong, and G. Wang. Computational methods for optical molecular imaging. *Communications in Numerical Methods in Engineering*, 25:1137–1161, 2009.

[39] Duan Chen, Zhan Chen, Changjun Chen, W. H. Geng, and G. W. Wei. MIBPB: A software package for electrostatic analysis. *J. Comput. Chem.*, 32:657 – 670, 2011.

[40] Duan Chen, Zhan Chen, and G. W. Wei. Quantum dynamics in continuum for proton transport II: Variational solvent-solute interface. *International Journal for Numerical Methods in Biomedical Engineering*, 28:25 – 51, 2012.

[41] Duan Chen and G. W. Wei. Quantum dynamics in continuum for proton transport—Generalized correlation. *J Chem. Phys.*, 136:134109, 2012.

[42] J. L. Chen, Louis Noodleman, D. A. Case, and D. Bashford. Incorporating solvation effects into density-functional electronic-structure calculations. *J. Phys. Chem.*, 98:11059–11068, 1994.

[43] Z. Chen, N. A. Baker, and G. W. Wei. Differential geometry based solvation models I: Eulerian formulation. *J. Comput. Phys.*, 229:8231–8258, 2010.

[44] Z. Chen, N. A. Baker, and G. W. Wei. Differential geometry based solvation models II: Lagrangian formulation. *J. Math. Biol.*, 63:1139– 1200, 2011.

[45] Z. Chen and G. W. Wei. Differential geometry based solvation models III: Quantum formulation. *J. Chem. Phys.*, 135:194108, 2011.

[46] Z. Chen, Shan Zhao, J. Chun, D. G. Thomas, N. A. Baker, P. B. Bates, and G. W. Wei. Variational approach for nonpolar solvation analysis. *Journal of Chemical Physics*, 137(084101), 2012.

[47] L. T. Cheng, Joachim Dzubiella, Andrew J. McCammon, and B. Li. Application of the level-set method to the implicit solvation of nonpolar molecules. *Journal of Chemical Physics*, 127(8), 2007.

[48] T. Cheng, X. Li, Y. Li, Z. Liu, and R. Wang. Comparative assesment of scoring functions on a diverse test set. *J. Chem. Inf. Model.*, 49:1079–1093, 2009.

[49] I. L. Chern, J.-G. Liu, and W.-C. Weng. Accurate evaluation of electrostatics for macromolecules in solution. *Methods and Applications of Analysis*, 10(2):309–28, 2003.

[50] Niharendu Choudhury and B Montgomery Pettitt. On the mechanism of hydrophobic association of nanoscopic solutes. *Journal of the American Chemical Society*, 127(10):3556–3567, 2005.

[51] M. L. Connolly. Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16(5):548–558, 1983.

[52] M. L. Connolly. Depth buffer algorithms for molecular modeling. *J. Mol. Graphics*, 3:19–24, 1985.

[53] R. B. Corey and L. Pauling. Molecular models of amino acids, peptides and proteins. *Rev. Sci. Instr.*, 24:621–627, 1953.

[54] Maurizio Cossi, Nadia Rega, Giovanni Scalmani, and Vincenzo Barone. Energies, structures, and electronic properties of molecules in solution with the c-pcm solvation model. *Journal of Computational Chemistry*, 24(6):669–681, 2003.

[55] C. J. Cramer. *Essentials of Computational Chemistry: Theories and Models*. John Wiley and Sons, 2013.

[56] C. J. Cramer and D. G. Truhlar. Implicit solvation models: equilibria, structure, spectra, and dynamics. *Chemical Reviews*, 99(8):2161–2200, 1999.

[57] Christopher J. Cramer and Donald G. Truhlar. A universal approach to solvation modeling. *Accounts of chemical research*, 41:760–768, 2008.

[58] Peter B. Crowley and Adel Golovin. Cation-pi interactions in protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 59(2):231–239, 2005.

[59] M. Daily, J. Chun, A. Heredia-Langner, G. W. Wei, and N. A. Baker. Origin of parameter degeneracy and molecular shape relationships in geometric-flow calculations of solvation free energies. *Journal of Chemical Physics,*, 139:204108, 2013.

[60] R. Daudel. Quantum theory of chemical reactivity. In *Quantum Theory of Chemical Reactivity*, 1973.

[61] L. David, R. Luo, and M. K. Gilson. Comparison of generalized Born and Poisson models: Energetics and dynamics of HIV protease. *Journal of Computational Chemistry*, 21(4):295–309, 2000.

[62] M. E. Davis and J. A. McCammon. Electrostatics in biomolecular structure and dynamics. *Chemical Reviews*, 94:509–21, 1990.

[63] R. L. DesJarlais, R. P. Sheridan, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan. Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.*, 29:2149–2153, 1986.

[64] B. N. Dominy and C. L. Brooks, III. Development of a generalized Born model parameterization for proteins and nucleic acids. *Journal of Physical Chemistry B*, 103(18):3765–3773, 1999.

[65] F. Dong, M. Vijaykumar, and H. X. Zhou. Comparison of calculation and experiment implicates significant electrostatic contributions to the binding stability of barnase and barstar. *Biophysical Journal*, 85(1):49–60, 2003.

[66] F. Dong and H. X. Zhou. Electrostatic contribution to the binding stability of protein-protein complexes. *Proteins*, 65(1):87–102, 2006.

[67] J. P. Donley, J. G. Curro, and J. D. McCoy. A density functional theory for pair correlation functions in molecular liquids. *The Journal of chemical physics*, 101:3205, 1994.

[68] Anatoly I. Dragan, Christopher M. Read, Elena N. Makeyeva, Ekaterina I. Milgotina, Mair E. Churchill, Colyn Crane-Robinson, and Peter L. Privalov. DNA binding and bending by HMG boxes: Energetic determinants of specificity. *Journal of Molecular Biology*, 343(2):371–393, 2004.

[69] J. Dzubiella and J.-P. Hansen. Competition of hydrophobic and coulombic interactions between nanosized solutes. *The Journal of Chemical Physics*, 121(11):5514–5530, September 2004.

[70] J. Dzubiella, J. M. J. Swanson, and J. A. McCammon. Coupling hydrophobicity, dispersion, and electrostatics in continuum solvent models. *Physical Review Letters*, 96:087802, 2006.

[71] B. Mennucci E. Cances and J. Tomasi. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic dielectrics. *Jornal of Chemical Physics*, 107:3032, 1997.

[72] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.

[73] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[74] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided. Mol. Des*, 11:425–445, 1997.

[75] M. Feig, W. Im, and C. L. Brooks III. Implicit solvation based on generalized Born theory in different dielectric environments. *Journal of Chemical Physics*, 120(2):903–911, 2004.

[76] F. Fogolari, A. Brigo, and H. Molinari. The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *Journal of Molecular Recognition*, 15(6):377–92, 2002.

[77] B. Fornberg. Calculation of weights in finite difference formulas. *SIAM Rev*, 40:685–691, 1998.

[78] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[79] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M Shelley, J. K. Perry JK, D. E. Shaw, P. Francis, and P. S. Shenkin. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.*, 47:1739, 2004.

[80] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, . Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian09 Revision E.01. Gaussian Inc. Wallingford CT 2009.

[81] J. Fu, Y. Liu, and J. Wu. Fast prediction of hydration free energies for sampl4 blind test from a classical density functional theory. *The Journal of Computer-Aided Molecular Design*, 28:299–304, 2014.

[82] E. Gallicchio, 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. *Journal of Physical Chemistry B*, 104(26):6271–6285, 2000.

[83] E. Gallicchio and R. M. Levy. AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *Journal of Computational Chemistry*, 25(4):479–499, 2004.

[84] E. Gallicchio, 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*, 23(5):517–29, 2002.

[85] J Gasteiger and M Marsili. Iterative partial equalization of orbital electronegativitya rapid access to atomic charges. *Tetrahedron*, 36:3219–3228, 1980.

[86] M. T. Geballe, A. G. Skillman, A. Nicholls, J. P. Guthrie, and J. P. Taylor. The sampl2 blind prediction challenge: introduction and overview. *Journal of Computer-Aided Molecular Design*, 24:259 –279, 2010.

[87] Matthew T. Geballe and J. P. Guthrie. The SAMPL3 blind prediction challenge: transfer energy overview. *Journal of Computer-Aided Molecular Design*, 26:489 –496, 2012.

[88] DK Gehlhaar, GM Verkhivker, PA Rejto, CJ Sherman, DB Fogel, LJ Fogel, and ST Freer. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem Biol.*, 2(5):317–324, 1995.

[89] Weihua Geng and G. W. Wei. Multiscale molecular dynamics using the matched interface and boundary method. *Journal of computational physics*, 230(2):435–457, 2011.

[90] Weihua Geng, Sining Yu, and G. W. Wei. Treatment of charge singularities in implicit solvent models. *Journal of Chemical Physics*, 127:114106, 2007.

[91] G. M. Giambasu, T. Luchko, D. Herschlag, D. M. York, and D. A. Case. Ion counting from explicit-solvent simulations and 3d-rism. *Biophysical Journal*, 106:883–894, 2014.

[92] M. K. Gilson, M. E. Davis, B. A. Luty, and J. A. McCammon. Computation of electrostatic forces on solvated molecules using the Poisson-Boltzmann equation. *Journal of Physical Chemistry*, 97(14):3591–3600, 1993.

[93] M. K. Gilson and Huan Xiang Zhou. Calculation of protein-ligand binding affinities. *Annual Review of Biophysics and Biomolecular Structur*, 36:21–42, 2007.

[94] Michael K. Gilson, Kim A. Sharp, and Barry H. Honig. Calculating the electrostatic potential of molecules in solution: Method and error assessment. *Jornal of Computational Chemistry*, 9:327–335, 1988.

[95] H. Gohlke and D. A. Case. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex ras-raf. *Journal of Computational Chemistry*, 25(2):238–250, 2004.

[96] H Gohlke, M Hendlich, and G Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol.*, 295(2):337–356, 2000.

[97] D. S. Goodsell and A. J. Olson. Automated docking of substrates to proteins by simulated annealing. *Protein Struct. Funct. Genet.*, 8:195–202, 1990.

[98] J. A. Grant and B. T. Pickup. A gaussian description of molecular shape. *Journal of Physical Chemistry*, 99:3503–3510, 1995.

[99] J. A. Grant, B. T. Pickup, M. T. Sykes, C. A. Kitchen, and A. Nicholls. The Gaussian Generalized Born model: application to small molecules. *Physical Chemistry Chemical Physics*, 9:4913–22, 2007.

[100] J. Andrew Grant, Barry T. Pickup, and Anthony Nicholls. A smooth permittivity function for Poisson-Boltzmann solvation methods. *Journal of Computational Chemistry*, 22(6):608–640, 2001.

[101] J.A. Grant, B. Pickup, and A. Nicholls. A smooth permittivity function for poisson-boltzmann solvation methods. *J Comput Chem*, 22:608, 2001.

[102] Michael Grant and Stephen Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. http://stanford.edu/ boyd/graph-dcp.html.

[103] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

[104] Paulette A. Greenidge, Christian Kramer, Jean-Christophe Mozziconacci, and Romain M. Wolf. MM/GBSA binding energy prediction on the PDBBind data set: Successes, failures, and directions for further improvement. *Journal of Chemical Information and Model*, 53:201–209, 2013.

[105] J. Peter Guthrie. A blind challenge for computational solvation free energies: Introduction and overview. *Journal of Physical Chemistry B*, 113:4501–4507, 2009.

[106] J. Peter Guthrie. Sampl4, a blind challenge for computational solvation free energies: the compounds considered. *J. Comput Aided Mol Des*, 28:151–168, 2014.

[107] Robert C. Harris, Aleander H. Boschitsch, and Marcia O. Fenley. Influence of grid spacing in Poisson-Boltzmann equation binding energy estimation. *Journal of Chemical Theory and Computation*, 9:3677–3685, 2013.

[108] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning: Data mining, inference, and prediction. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. Springer, 2009.

[109] Kenneth Haugland. Polynomial equation solver. *http://www.codeproject.com/Articles/552678/Polynomial-Equation-Solver*, 2013.

[110] F. Hirata and et al. Molecular theory of solvation. In F. Hirata, editor, *Molecular Theory of Solvation*. Springer, 2003.

[111] Christian Holm, Patrick Kekicheff, and Rudolf Podgornik. *Electrostatic effects in soft matter and biophysics; NATO Science Series.* Kluwer Academic Publishers, Boston, 2001.

[112] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–9, 1995.

[113] C. S. Hsu and D. Chandler. Rism calculation of the structure of liquid acetonitrile. *Molecular Physics*, 36(1):215–224, 1978.

[114] David M Huang and David Chandler. Temperature and length scale dependence of hydrophobic effects and their possible implications for protein folding. *Proceedings of the National Academy of Sciences*, 97(15):8324–8327, 2000.

[115] H. Huang and Z. L. Li. Convergence analysis of the immersed interface method. *IMA Journal of Numerical Analysis*, 19(4):583–608, 1999.

[116] S. Y. Huang and X. Zou. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. derivation of interaction potentials. *J. Comput. Chem.*, 27:1865–1875, 2006.

[117] Sheng-You Huang and Xiaoqin Zou. Inclusion of solvation and entropy in the knowledge-based scoring function for proteinligand interactions. *J. Chem. Inf. Model.*, 50(2):262–273, 2010.

[118] W. Humphrey, A. Dalke, and K. Schulten. VMD – visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[119] R. Ishizuka, S. H. Chong, and F. Hirata. An integral equation theory for inhomogeneous molecular fluids: The reference interaction site model approach. *Journal of Chemical Physics*, 128(3):34504–34504, 2008.

[120] Richard M. Jackson and Michael J. Sternberg. A continuum model for protein-protein interactions: Application to the docking problem. *Journal of Molecular Biology*, 250(2):258–275, 1995.

[121] Benedetta Mennucci Jacopo Tomasi and Roberto Cammi. Quantum mechanical contunuum solvation models. *Chem. Rev.*, 105:2999–3093, 2005.

[122] A. Jakalian, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. parameterization and validation. *Journal of Computational Chemistry*, 23(16):1623–1641, 2002.

[123] Araz Jakalian, Bruce L. Bush, David B. Jack, and Christopher I. Bayly. Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method. *Journal of Computational Chemistry*, 21(2):132–146, 2000.

[124] B. Jayaram, D. Sprous, and D. L. Beveridge. Solvation free energy of biomacromolecules: Parameters for a modified generalized Born model consistent with the AMBER force field. *Journal of Physical Chemistry B*, 102(47):9571–9576, 1998.

[125] F. Jensen. *Introduction to computational chemistry.* John Wiley and Sons, 2007.

[126] R. Jinnouchi and A. B. Anderson. Electronic structure calculations of liquid-solid interfaces: Combination of density functional theory and modified Poisson-Boltzmann theory. *PHYSICAL REVIEW B*, 77:245417, 2008.

[127] S. Jo, T. Kim, V.G. Iyer, and W. Im. Charmm-gui: a web-based graphical user interface for charmm. *J Comput Chem*, 29(11):1859–1865, 2008.

[128] T. Joachims. Optimizing search engines using clickthrough data. *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2002.

[129] T. Joachims. Training linear SVMs in linear time. *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

[130] Gareth Jones, Peter Willett, Robert C Glen, Andrew R Leach, and Robin Taylor. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 267(3):727–748, 1997.

[131] W. L. Jorgensen. Rusting of the lock and key model for protein-ligand binding. *Science*, 254:954–955, 1991.

[132] William L. Jorgensen and Julian. Tirado-Rives. The OPLS optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110(6):1657–1666, 1988.

[133] T. Kaczynski, K. Mischaikow, and M. Mrozek. *Computational homology.* Springer-Verlag, 2004.

[134] Charles W. Kehoe, Christopher J. Fennell, and Ken A. Dill. Testing the semi-explicit assembly solvation model in the sampl3 community blind test. *J Comput Aided Mol Des*, 26:563–568, 2012.

[135] Sarah L. Kinnings, Nina Liu, Peter J. Tonge, Richard M. Jackson, Lei Xie, and Philip E. Bourne. A machine learning based method to improve docking scoring functions and its application to drug repurposing. *Journal of Chemical Information and Model*, 51(2):408–419, 2011.

[136] J. G. Kirkwood. Theory of solution of molecules containing widely separated charges with special application to zwitterions. *J. Comput. Phys.*, 7:351 – 361, 1934.

[137] Pavel V. Klimovich and David L. Mobley. Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. *J. Comput Aided Mol Des.*, 24:307–316, 2010.

[138] P. Koehl. Electrostatics calculations: latest methodological advances. *Current Opinion in Structural Biology*, 16(2):142–51, 2006.

[139] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and III Cheatham, T. E. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of Chemical Research*, 33(12):889–97, 2000.

[140] M.M. Kreevoy and D.G. Truhlar. In investigation of rates and mechanisms of reactions, part i. In C.F. Bernasconi, editor, *In Investigation of Rates and Mechanisms of Reactions, Part I*, page 13. Wiley: New York, 1986.

[141] M. Krone, K. Bidmon, and T. Ertl. Interactive visualization of molecular surface dynamics. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1391–1398, Nov 2009.

[142] Leslie A. Kuhn, Michael A. Siani, Michael E. Pique, Cindy L. Fisher, Elizabeth D. Getzoff, and John A. Tainer. The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures. *Journal of Molecular Biology*, 228(1):13–22, 1992.

[143] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161:269–288, 1982.

[144] T.-M. Kuo, C.-P. Lee, and C.-J. Lin. Large-scale kernel RankSVM. *SIAM International Conference on Data Mining*, 2014.

[145] G. Lamm. The Poisson-Boltzmann equation. In K. B. Lipkowitz, R. Larter, and T. R. Cundari, editors, *Reviews in Computational Chemistry*, pages 147–366. John Wiley and Sons, Inc., Hoboken, N.J., 2003.

[146] A. R. Leach, B. K. Shoichet, and C. E. Peishoff. Prediction of protein-ligand interactions. docking and scoring: Successes and gaps. *J. Med. Chem.*, 49:5851–5855, 2006.

[147] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*, 55(3):379–400, 1971.

[148] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785, 1988.

[149] R. J. LeVeque and Z. L. Li. The immersed interface method for elliptic equations with discontinuous coefficients and singular sources. *SIAM J. Numer. Anal.*, 31:1019–1044, 1994.

[150] Guo-Bo Li, Ling-Ling Yang, Wen-Jing Wang, Lin-Li Li, and Sheng-Yong Yang. ID-Score: A new empirical scoring function based on a comprehensive set of descriptors related to proteinligand interactions. *J. Chem. Inf. Model.*, 53(3):592–600, 2013.

[151] H. Li, K.S. Leung, P.J. Ballester, and M. H. Wong. iStar: A web platform for large-scale protein-ligand docking. *Plos One*, 9(1), 2014.

[152] Hongjian Li, Kwong-Sak Leung, ManHon Wong, and Pedro J Ballester. Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: Cyscore as a case study. *BMC Bioinformatics*, 15(291), 2014.

[153] Jiabo Li, Tianhai Zhu, Gergory D. Hawkins, Paul Winget, Daniel A. Liotard, Christopher J. Cramer, and Donald G. Truhlar. Extension of the platform of applicability of the sm5.42r universal solvation model. *Theoretical Chemistry Accounts*, 103, 1999.

[154] Lin Li, Chuan Li, Subhra Sarkar, Jie Zhang, Shawn Witham, Zhe Zhang, Lin Wang, Nicholas Smith, Marharyta Petukh, and Emil Alexov. Delphi: a comprehensive suite for delphi software and associated resources. *BMC Biophysics*, 5:9:2046–1682, 2012.

[155] V. J. Licata and N. M. Allewell. Functionally linked hydration changes in escherichia coli aspartate transcarbamylase and its catalytic subunit. *Biochemistry*, 36(33):10161–10167, 1997.

[156] Jung-Hsin Lin, Nathan Andrew Baker, and J. Andrew McCammon. Bridging the implicit and explicit solvent approaches for membrane electrostatics. *Biophysical Journal*, 83(3):1374–1379, 2002.

[157] Beibei Liu, Bao Wang, Rundong Zhao, Yiying Tong, and Guo Wei Wei. ESES: software for Eulerian solvent excluded surface. *Preprint*, 2015.

[158] Jie Liu and Renxiao Wang. Clasification of current scoring functions. *Journal of Chemical Information and Model*, 55(3):475–482, 2015.

[159] T.T. Liu, M.X. Chen, and B.Z. Lu. Parameterization for molecular gaussian surface and a comparison study of surface mesh generation. *J. Molecular Modeling*, 21(5):113, 2015.

[160] Zhihai Liu, Yan Li, Li Han, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31(3):405–412, 2015.

[161] J. R. Livingstone, R. S. Spolar, and M. T. Record Jr. Contribution to the thermodynamics of protein folding from the reduction in water-accessible nonpolar surface area. *Biochemistry*, 30(17):4237–44, 1991.

[162] Benzhuo Lu, Xiaolin Cheng, Jingfang Huang, and J. Andrew McCammon. AFMPB: An Adaptive Fast Multipole Poisson-Boltzmann Solver for Calculating Electrostatics in Biomolecular Systems. *Comput. Phys. Commun.*, 184:2618–2619, 2013.

[163] K. Lum, D. Chandler, and J. D. Weeks. Hydrophobicity at small and large length scales. *Journal of Physical Chemistry B*, 103(22):4570–7, 1999.

[164] Jr. MacKerell, A. D., D. Bashford, M. Bellot, Jr. Dunbrack, R. L., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, III Reiher, W. E., B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B*, 102(18):3586–3616, 1998.

[165] Aleksandr V. Marenich, Christopher J. Cramer, , and Donald G. Truhlar. Performance of SM6, SM8, and SMD on the SAMPL1 test set for the prediction of small-molecule solvation free energies. *Journal of Physical Chemistry B*, 113:45384543, 2009.

[166] A. Marquina and S. Osher. Explicit algorithms for a new time dependent model based on level set motion for nonlinear deblurring and noise removal. *SIAM Journal on Scientific Computing*, 22(2):387–405, 2000.

[167] Irina Massova and Peter A Kollman. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspectives in drug discovery and design*, 18(1):113–135, 2000.

[168] A. Mayo. The fast solution of Poisson's and the biharmonic equations on irregular regions. *SIAM J. Numer. Anal.*, 21:285–299, 1984.

[169] Paolo Mazzatorta, Lin-Anh Tran, Benot Schilter, and Martin Grigorov. Integration of structure-activity relationship and artificial intelligence systems to improve in silico prediction of ames test mutagenicity. *J. Chem. Inf. Model.*, 47(1):34–38, 2007.

[170] A. McKenney, L. Greengard, and A. Mayo. A fast Poisson solver for complex geometries. *J. Comput. Phys.*, 118:348–355, 1995.

[171] Elaine C. Meng, Brian K. Shoichet, and Irwin D. Kuntz. Automated docking with grid-based energy evaluation. *Journal of Computational Chemistry*, 13:505–524, 1992.

[172] K. Mischaikow and V. Nanda. Morse theory for filtrations and efficient computation of persistent homology. *Discrete and Computational Geometry*, 50(2):330–353, 2013.

[173] David L. Mobley, Christopher I. Bayly, Matthew D. Cooper, and Ken A. Dill. Predictions of hydration free energies from all-atom molecular dynamics simulations. *J. Phys. Chem. B.*, 13:4533–4537, 2009.

[174] David L. Mobley, Ken A. Dill, and John D. Chodera. Treating entropy and conformational changes in implicit solvent simulations of small molecules. *J. Phys. Chem. B.*, 112:938–946, 2008.

[175] David L. Mobley and J. Peter Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of Computer-Aided Molecular Design*, 28:711–720, 2014.

[176] David L. Mobley, Shaui Liu, David S. Cerutti, William C. Swope, and Julia E. Rice. Alchemical prediction of hydration free energies for sampl. *J. Comput Aided Mol Des.*, 26:551–562, 2012.

[177] David L. Mobley, Karisa L. Wymer, Nathan M. Lim, and J. Peter Guthrie. Blind prediction of solvation free energies from the sampl4 challenge. *J. Comput Aided Mol Des*, 28:135–150, 2014.

[178] J. Mongan, C. Simmerling, J. A. McCammon, D. A. Case, and A. Onufriev. Generalized Born model with a simple, robust molecular volume correction. *Journal of Chemical Theory and Computation*, 3(1):159–69, 2007.

[179] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19:1639–62, 1998.

[180] I Muegge and YC Martin. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem.*, 42(5):791–804, 1999.

[181] Vidit Nanda. Perseus: the persistent homology software. Software available at http://www.sas.upenn.edu/ vnanda/perseus.

[182] R. R. Netz and H. Orland. Beyond Poisson-Boltzmann: Fluctuation effects and correlation functions. *European Physical Journal E*, 1(2-3):203–14, 2000.

[183] Anthony Nicholls, David L. Mobley, J. Peter Guthrie, John D. Chodera, Chridtopher I. Bayly, Matthew D. Cooper, and Vijay S. Pande. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. *J. Med. Chem.*, 51:769–799, 2008.

[184] Anthony Nicholls, David L Mobley, J Peter Guthrie, John D Chodera, Christopher I Bayly, Matthew D Cooper, and Vijay S Pande. Predicting small-molecule solvation free energies: an informal blind test for computational chemistry. *Journal of Medicinal Chemistry*, 51(4):769–779, 2008.

[185] Anthony Nicholls, David L. Mobley, Peter J. Guthrie, John D. Chodera, and Vijay S. Pande. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. *Journal of Medicinal Chemistry*, 51(4):769–79, 2008.

[186] M. Nina, W. Im, and B. Roux. Optimized atomic radii for protein continuum electrostatics solvation forces. *Biophysical Chemistry*, 78(1-2):89–96, 1999.

[187] F. N. Novikov, A. A. Zeifman, O. V. Stroganov, V. S. Stroylov, V. Kulkov, and G. G. Chilov. CSAR Scoring challenge reveals the need for new concepts in estimating protein-ligand binding affinity. *Journal of Chemical Information and Model*, 51:2090–2096, 2011.

[188] A. Okur, L. Wickstrom, M. Layten, R. Geney, K. Song, V. Hornak, and C. Simmerling. Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. *Journal of Chemical Theory and Computation*, 2(2):420–433, 2006.

[189] Mats H. M. Olsson, Chresten R. Sondergaard, Michal Rostkowski, and Jan H. Jensen. PROPKA3: consistent treatment of internal and surface residues in empirical pka predictions. *J. Chem. Theory Comput.*, 7(2):525–537, 2011.

[190] A. Onufriev, D. Bashford, and D. A. Case. Modification of the generalized Born model suitable for macromolecules. *Journal of Physical Chemistry B*, 104(15):3712–3720, 2000.

[191] A. Onufriev, D. A. Case, and D. Bashford. Effective Born radii in the generalized Born approximation: the importance of being perfect. *Journal of Computational Chemistry*, 23(14):1297–304, 2002.

[192] M. Orozco and F. J. Luque. Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem. Rev.*, 100:4187–4225, 2000.

[193] A. R. Ortiz, M. T. Pisabarro, F. Gago, and R. C. Wade. Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem*, 38:2681–2691, 1995.

[194] Stanley Osher and Ronald P. Fedkiw. Level set methods: An overview and some recent results. *J. Comput. Phys.*, 169(2):463–502, 2001.

[195] Insook Park, Yun Hee Jang, Sungu Hwang, and Doo Soo Chung. Poisson-boltzmann continuum solvation models for nonaqueous solvents i. 1-octanol. *Chemistry Letters*, 32:4, 2003.

[196] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera – a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

[197] R. A. Pierotti. A scaled particle theory of aqueous and nonaqeous solutions. *Chemical Reviews*, 76(6):717–726, 1976.

[198] J. W. Ponder, C. J. Wu, P. Y. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson, and T. Head-Gordon. Current status of the amoeba polarizable force field. *J. Phys. Chem. B*, 114:2549 – 2564, 2010.

[199] N. V. Prabhu, P. Zhu, and K. A. Sharp. Implementation and testing of stable, fast implicit solvation in molecular dynamics using the smooth-permittivity finite difference Poisson-Boltzmann method. *Journal of Computational Chemistry*, 25(16):2049–2064, 2004.

[200] Rosa Ramirez and Daniel Borgis. Density functional theory of solvation and its relation to implicit solvent models. *Journal of Physical Chemistry B*, 109:6754–6763, 2005.

[201] M Rarey, B Kramer, T Lengauer, and G Klebe. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol.*, 261(3):470–489, 1996.

[202] E. L. Ratkova, G. N. Chuev, V. P. Sergiievskyi, and M. V. Fedorov. An accurate prediction of hydration free energies by combination of molecular integral equations theory with structural descriptors. *J. Phys. Chem. B*, 114(37):12068–2079, 2010.

[203] Ekaterina L. Ratkova, Gennady N. Chuev, Volodynyr P. Sergiievskyi, and Maxim V. Fedorov. An accurate prediction of hydration free energies by combination of molecular integral equations theory with structural descriptors. *Journal of Physical Chemistry B*, 114:12068–12079, 2010.

[204] C. Reichardt. Solvents and solvent effects in organic chemistry. In *Solvents and Solvent Effects in Organic Chemistry*. VCH:New York, 1990.

[205] Jens Reinisch and Andreas Klamt. Prediction of free energies of hydration with cosmo-rs on the sampl4 data set. *J. Comput Aided Mol Des*, 28:169–173, 2014.

[206] F. M. Richards. Areas, volumes, packing, and protein structure. *Annual Review of Biophysics and Bioengineering*, 6(1):151–176, 1977.

[207] Robert C. Rizzo, Tiba Aynechi, David A. Case, and Irwin D. Kuntz. Estimation of absolute free energies of hydration using continuum methods: Accuracy of partial charge models and optimization of nonpolar contributions. *Journal of Chemical Theory and Computation*, 2:128–139, 2006.

[208] W. Rocchia, E. Alexov, and B. Honig. Extending the applicability of the nonlinear poisson-boltzmann equation: Multiple dielectric constants and multivalent ions. *J. Phys. Chem.*, 105:6507–6514, 2001.

[209] W. Rocchia, S. Sridharan, A. Nicholls, E Alexov, A Chiabrera, and B. Honig. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. *Journal of Computational Chemistry*, 23:128 – 137, 2002.

[210] Michal Rostkowski, Mats HM Olsson, Chresten R Sondergaard, and Jan H Jensen. Graphical analysis of pH-dependent properties of proteins predicted using PROPKA. *BMC Structural Biology*, 11(6), 2011.

[211] B. Roux and T. Simonson. Implicit solvent models. *Biophysical Chemistry*, 78(1-2):1–20, 1999.

[212] M. F. Sanner, A. J. Olson, and J. C. Spehner. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers*, 38:305–320, 1996.

[213] G. Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aid. Mol. Des.*, 27:221–234, 2013.

[214] A. Savelyev and G. A. Papoian. Inter-DNA electrostatics from explicit solvent molecular dynamics simulations. *Journal of the American Chemical Society*, 129(19):6060–1, 2007.

[215] BK Schichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, 2004.

[216] Tamar Schlick. *Innovations in Biomolecular Modeling and Simulations: Volume 1*. Royal Society of Chemistry, 2012.

[217] J. A. Sethian. *Level Set Methods and Fast Marching Methods*, volume 3 of *Monographs on Appl. Comput. Math.* Cambridge University Press, Cambridge, 2nd edition, 1999.

[218] K. A. Sharp and B. Honig. Calculating total electrostatic energies with the nonlinear Poisson-Boltzmann equation. *Journal of Physical Chemistry*, 94:7684–7692, 1990.

[219] K. A. Sharp and B. Honig. Electrostatic interactions in macromolecules - theory and applications. *Annual Review of Biophysics and Biophysical Chemistry*, 19:301–332, 1990.

[220] Devleena Shlvakumar, Joshua Willams, Yujie Wu, Wolfgang Damm, John Shelly, and Woody Sherman. Prediction of absolute solvation fre energies using molecular dynamics free energy perturbation and the opls force field. *Journal of Chemical Theory and Computation*, 6(5):1509–1519, 2010.

[221] Doree Sitoff, Nir Ben-Tal, and Barry Honig. Calculation of alkane to water solvation free energies using continuum solvent models. *J. Phys. Chem.*, 100:2744–2752, 1996.

[222] R. Skyner, J. L. McDonagh, C. R. Groom, T. van Mourik, and J. B. O. Mitchell. A review of methods for the calculation of solution free energies and the modelling of systems in solution. *Phys. Chem. Chem. Phys*, 17(9):6174, 2015.

[223] Peter Smereka. The numerical approximation of a delta function with application to level set methods. *J. Comput. Phys.*, 211(1):77–90, 2006.

[224] J. M. Soler, E. Artacho, J. D. Gale, A. Garca, J. Junquera, P. Ordejn, and D. Snchez-Portal. The siesta method for ab-initio order-n materials simulation. *J. Phys.: Condens. Matt.*, 14:2745–2779, 2002.

[225] R. S. Spolar, J. H. Ha, and M. T. Record Jr. Hydrophobic effect in protein folding and other noncovalent processes involving proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 86(21):8382–8385, 1989.

[226] F. H. Stillinger. Structure in aqueous solutions of nonpolar solutes from the standpoint of scaled-particle theory. *J. Solution Chem.*, 2:141 – 158, 1973.

[227] A. J. Stone. Distributed multipole analysis, or how to describe a molecular charge distribution. *Chemical Physics Letters*, 83(2):233–239, 1981.

[228] Joey W. Storer, David J. Giesen, Gregory D. Hawkins, Gillian C. Lynch, Christopher J. Cramer, Donald G. Truhlar, and Daniel A. Liotard. Solvation modeling in aqueous and nonaqueous solvent, new techniques and a reexamination of the claisen rearrangement. In C. J. Cramer and D. G. Truhlar, editors, *Structure, Energetics, and Reactivity in Aqueous Solution: Characterization of Chemical and Biological Systems*, 568, pages 24–49. American Chemical Society Symposium, 1994.

[229] Pin-Chih Su, Cheng-Chieh Tsai, Shahila Mehboob, Kirk E. Heveber, and Michael E. Johnson. Comparison of radii sets, entropy, qm methods, and sampling on MM-PBSA, MM-GBSA, and QM/MM-GBSA ligand binding energies of f. tularensis enoyl-acp reductase (fabl). *Journal of Computational Chemistry*, 36:1859–1873, 2015.

[230] J. M. J. Swanson, 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*, 86(1):67–74, 2004.

[231] J. M. J. Swanson, J. Mongan, and J. A. McCammon. Limitations of atom-centered dielectric functions in implicit solvent models. *Journal of Physical Chemistry B*, 109(31):14769–72, 2005.

[232] J. M. J. Swanson, J. A. Wagoner, N. A. Baker, and J. A. McCammon. Optimizing the poisson dielectric bounday with explicit solvent forces and energies: Lessons learned with atom-centered dielectric functions. *Journal of Chemical Theory and Computation*, 3(1):170–83, 2007.

[233] C. Tan, L. Yang, and R. Luo. How well does Poisson-Boltzmann implicit solvent agree with explicit solvent? A quantitative analysis. *Journal of Physical Chemistry B*, 110(37):18680–18687, 2006.

[234] Jian J. Tan, Wei Z. Chen, and Cun X. Wang. Investigating interactions between HIV-1 gp41 and inhibitors by molecular dynamics simulation and MM-PBSA/GBSA calculations. *Journal of Molecular Structure: Theochem.*, 766(2-3):77–82, 2006.

[235] D.G. Thomas, J. Chun, Z. Chen, G. W. Wei, and N. A. Baker. Parameterization of a geometric flow implicit solvation model. *J. Comput. Chem.*, 24:687–695, 2013.

[236] H. Tjong and H. X. Zhou. GBr6NL: A generalized Born method for accurately reproducing solvation energy of the nonlinear Poisson-Boltzmann equation. *Journal of Chemical Physics*, 126:195102, 2007.

[237] Jacopo Tomasi, Benedetta Mennucci, and Roberto Cammi. Quantum mechanical continuum solvation models. *Chem. Rev.*, 105:2999–3093, 2005.

[238] Jacopo Tomasi, Benedetta Mennucci, and Roberto Cammi. Quantum mechanical continuum solvent models. *Chem. Rev.*, 105:2999–3093, 2005.

[239] Jacopo Tomasi and Maurizio Persico. Molecular interactions in solution: an overview of methods based on continuous distributions of the solvent. *Chem. Rev.*, 94:2027–2094, 1994.

[240] O. Trott and A. J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Computat Chem*, 31(2):455–461, 2010.

[241] V. Tsui and D. A. Case. Molecular dynamics simulations of nucleic acids with a generalized Born solvation model. *Journal of the American Chemical Society*, 122(11):2489–2498, 2000.

[242] Weijo V1, Randrianarivony M, Harbrecht H, and Frediani L. Wavelet formulation of the polarizable continuum model. *Journal of Computational Chemistry*, 31(7):1469–1477, 2010.

[243] H. F. G. Velec, H. Gohlke, and G. Klebe. Knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem*, 48:6296–6303, 2005.

[244] HF Velec, H Gohlke, and G Klebe. DrugScore (CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J Med Chem.*, 48(20):6296–303, 2005.

[245] G. Verkhivker, K. Appelt, S. T. Freer, and J. E. Villafranca. Empirical free energy calculations of ligand-protein crystallographic complexes. i. knowledge based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus protease binding affinity. *Protein Eng*, 8:677–691, 1995.

[246] J. A. Wagoner and N. A. Baker. Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms. *Proceedings of the National Academy of Sciences of the United States of America*, 103(22):8331–6, 2006.

[247] N. Wale and G. Karypis. Target fishing for chemical compounds using target-ligand activity data and ranking based methods. *Journal of Chemical Information and Model*, 49(10):2190–201, 2009.

[248] B. Wang and G. W. Wei. Parameter optimization in differential geometry based solvation models. *Journal Chemical Physics*, 143:134119, 2015.

[249] Bao Wang, Chengzhang Wang, and Guowei Wei. Feature functional theory - solvation predictor (fft-sp) for the blind prediction of solvation free energy. *Preprint*.

[250] Bao Wang, Chengzhang Wang, and Guowei Wei. Learning to rank for solvation free energy prediction. *Preprint*, 2016.

[251] Bao Wang and G. W. Wei. Objective-oriented Persistent Homology. *ArXiv e-prints*, December 2014.

[252] Bao Wang and Guo-Wei Wei. Coarse grid poisson boltzmann solver without loss of accuracy. *Preprint*.

[253] Bao Wang and Guo-Wei Wei. Accurate, robust and reliable calculations of poisson-boltzmann solvation and binding energies. *Preprint*, 2016.

[254] Bao Wang, Zhixiong Zhao, and Guowei Wei. Feature functional theory - binding predictor (fft-bp) for the blind prediction of binding free energy. *Preprint*.

[255] Bao Wang, Zhixiong Zhao, and Guowei Wei. Hybrid physical and statistical models for the blind prediction of solvation free energies. *Preprint*.

[256] Bao Wang, Zhixiong Zhao, and Guowei Wei. Hybrid physical and statistical models for the blind prediction of solvation free energies. *preprint*, 2016.

[257] Changhao Wang, Peter H. Nguyen, Kevin Pham, Danielle Huynh, Thanh-Binh Nancy Le, Hongli Wang, Pengyu Ren, and Ray Luo. Calculating protein-ligand binding affinities with mmpbsa: Method and error analysis. Preprint.

[258] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general AMBER force field. *Journal of Computational Chemistry*, 25(9):1157–74, 2004.

[259] Jun Wang, Qin Cai, Ye Xiang, and Ray Luo. Reducing Grid Dependence in Finite Difference Poisson Boltzmann Calculations. *Journal of Chemical Theory and Computation*, 8:2741–2751, 2012.

[260] Junmei Wang, Wei Wang, Shuanghong Huo, Matthew Lee, and Peter A. Kollman. Solvation model based on weighted solvent accessible surface area. *Journal of Physical Chemistry B*, 105:5055–5067, 2001.

[261] Junmei Wang, Wei Wang, Shuanghong Huo, Matthew Les, and Peter A. Kollman. Solvation model based on weighted solvent ccessible surface area. *J. Phys. Chem. B*, 105:5055–5067, 2001.

[262] M. L. Wang and C. F. Wong. Calculation of solvation free energy from quantum mechanical charge density and continuum dielectric theory. *J. Phys. Chem. A*, 110:4873–4879, 2006.

[263] M. L. Wang, C. F. Wong, J. H. Liu, and P. X. Zhang. Efficient quantum mechanical calculation of solvation free energies based on density functional theory, numerical atomic orbitals and poisson-boltzmann equation. *Chemical Physics Letters*, 442:464–467, 2007.

[264] R. Wang, L. Lai, and S. Wang. Further development and validation of empirical scoring functions for structure based binding affinity prediction. *J. Comput. Aided. Mol. Des*, 16:11–26, 2002.

[265] Renxiao Wang, YiPin Lu, and Shaomeng Wang. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.*, 46:2287–2303, 2003.

[266] A. Warshel and A. Papazyan. Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Current Opinion in Structural Biology*, 8(2):211–217, 1998.

[267] A. M. Wassermann, H. Geppert, and J. R. Bajorath. Searching for target-selective compounds using different combinations of multiclass support vector machine ranking methods, kernel functions, and fingerprint descriptors. *Journal of Chemical Information and Model*, 49(3):582–92, 2009.

[268] Still WC, Tempczyk A, Hawley RC, and Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc*, 112(16):6127–6129, 1990.

[269] G. W. Wei. Generalized Perona-Malik equation for image restoration. *IEEE Signal Processing Lett.*, 6:165–167, 1999.

[270] G. W. Wei. Differential geometry based multiscale models. *Bulletin of Mathematical Biology*, 72:1562 – 1622, 2010.

[271] G. W. Wei, Y. H. Sun, Y. C. Zhou, and M. Feig. Molecular multiresolution surfaces. *arXiv:math-ph/0511001v1*, pages 1 – 11, 2005.

[272] Guo-Wei Wei. Multiscale, multiphysics and multidomain models I: Basic theory. *Journal of Theoretical and Computational Chemistry*, 12(8):1341006, 2013.

[273] Guo-Wei Wei, Qiong Zheng, Zhan Chen, and Kelin Xia. Variational multiscale models for charge transport. *SIAM Review*, 54(4):699 – 754, 2012.

[274] S. J. Weiner, P. A. Kollman, D. T. Nguyem, and D. A. Case. An all atom force-field for simulations of proteins and nucleic-acids. *J Comp Chem*, 7(2):230–252, 1986.

[275] T. J. Willmore. *Riemannian Geometry.* Oxford University Press, USA, 1997.

[276] J.M. Word and A. Nicholls. Application of the gaussian dielectric boundary in zap to the prediction of protein pka values. *Proteins*, 79(12):3400–3409, 2011.

[277] S. Yin, L. Biedermannova, J. Vondrasek, and N. V. Dokholyan. Medusascore: An accurate force field-based scoring function for virtual drug screening. *Journal of Chemical Information and Model*, 48:1656–1662, 2008.

[278] S. N. Yu, W. H. Geng, and G. W. Wei. Treatment of geometric singularities in implicit solvent models. *Journal of Chemical Physics*, 126:244108, 2007.

[279] S. N. Yu and G. W. Wei. Three-dimensional matched interface and boundary (MIB) method for treating geometric singularities. *J. Comput. Phys.*, 227:602–632, 2007.

[280] S. N. Yu, Y. C. Zhou, and G. W. Wei. Matched interface and boundary (MIB) method for elliptic problems with sharp-edged interfaces. *J. Comput. Phys.*, 224(2):729–756, 2007.

[281] Z. Y. Yu and C. Bajaj. Computational approaches for automatic structural analysis of large biomolecular complexes. *IEEE/ACM Trans Comput Biol Bioinform*, 5:568–582, 2008.

[282] C Zhang, S Liu, Q Zhu, and Y Zhou. A knowledge-based energy function for protein-ligand, protein-protein, and protein-dna complexes. *J Med Chem.*, 48(7):2325–35, 2005.

[283] Wei Zhang, Lijuan Ji, Yanan Chen, Kailin Tang, Haiping Wang, Ruixin Zhu, Wei Jia, Zhiwei Cao, and Qi Liu. When drug discovery meets web search: Learning to rank for ligand-based virtual screening. *Journal of Cheminformatics*, 7(5), 2015.

[284] Shan Zhao. Pseudo-time-coupled nonlinear models for biomolecular surface representation and solvation analysis. *International Journal for Numerical Methods in Biomedical Engineering*, 27:1964–1981, 2011.

[285] Shan Zhao. Operator splitting adi schemes for pseudo-time coupled nonlinear solvation simulations. *Journal of Computational Physics*, 257:1000 – 1021, 2014.

[286] Shan Zhao and G. W. Wei. High-order FDTD methods via derivative matching for Maxwell's equations with material interfaces. *J. Comput. Phys.*, 200(1):60–103, 2004.

[287] Zheng Zheng and Kenneth M. Merz Jr. Ligand identification scoring algorithm (LISA). *Journal of Chemical Information and Model*, 51:1296–1306, 2011.

[288] Zheng Zheng and Kenneth M. Merz Jr. Development of the knowledge-based and empirical combined scoring algorithm (KECSA) to score proteinligand interactions. *Journal of Chemical Information and Model*, 53:1073–1083, 2013.

[289] Zheng Zheng, Melek N. Ucisik, and Kenneth M. Merz Jr. The movable type method applied to proteinligand binding. *Journal of Chemical Theory and Computation*, 9:5526–5538, 2013.

[290] Zheng Zheng, Ting Wang, Pengfei Li, and Kenneth M. Merz Jr. KECSA-Movable type implicit solvation model (KMTISM). *Journal of Chemical Theory and Computation*, 11:667–682, 2015.

[291] S. G. Zhou, L. T. Cheng, H. Sun, J. W. Che, J. Dzubiella, B. Li, and J. A. McCammon. Ls-vism: A software package for analysis of biomolecular solvation. *Journal of Computational Chemistry*, 36:1047–1059, 2015.

[292] Y. C. Zhou. Matched interface and boundary (mib) method and its applications to implicit solvent modeling of biomolecules. *Ph.D. Thesis.*, Michigan State University, 2006.

[293] Y. C. Zhou, M. Feig, and G. W. Wei. Highly accurate biomolecular electrostatics in continuum dielectric environments. *Journal of Computational Chemistry*, 29:87–97, 2008.

[294] Y. C. Zhou and G. W. Wei. On the fictitious-domain and interpolation formulations of the matched interface and boundary (MIB) method. *J. Comput. Phys.*, 219(1):228–246, 2006.

[295] Y. C. Zhou, Shan Zhao, Michael Feig, and G. W. Wei. High order matched interface and boundary method for elliptic equations with discontinuous coefficients and singular sources. *J. Comput. Phys.*, 213(1):1–30, 2006.

[296] J. Zhu, E. Alexov, and B. Honig. Comparative study of generalized Born models: Born radii and peptide folding. *Journal of Physical Chemistry B*, 109(7):3008–22, 2005.