Development and Test of PBSA Solvation Models for Drug Design

by

Taoyu Niu

Bachelor of Science, China Pharmaceutical University, 2021

Submitted to the Graduate Faculty of the

School of Pharmacy in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

SCHOOL OF PHARMACY

This thesis or dissertation was presented

by

Taoyu Niu

It was defended on

April 10, 2024

and approved by

Junmei Wang, Associate Professor, School of Pharmacy

Robert Gibbs, Professor, School of Pharmacy

David Koes, Associate Professor, School of Medicine

Thesis Advisor/Dissertation Director: Junmei Wang, Associate Professor, School of Pharmacy

Development and Test of PBSA Solvation Models for Drug Design

Taoyu Niu, B.S.

University of Pittsburgh, 2024

The Poisson-Boltzmann Surface Area (PBSA) model was extensively used to predict solvation free energy (SFE) and protein-ligand binding free energies, as well as to study protein folding. In addition, partition coefficient (logP), which is an important physicochemical property that determines the distribution of a drug in vivo, can be derived directly from transfer free energies. Within the Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) 9 challenge, we applied the Poisson-Boltzmann (PB) surface area (SA) approach to predict toluene/water transfer free energy and partition coefficient ($logP_{tol/wat}$) from SFEs. PB calculation directly adopts our previously optimized boundary definition - a set of general AMBER force field 2 (GAFF2) atom-type based sphere radii for solute atoms. For the non-polar SA model, we newly developed the solvent-related molecular surface tension parameters γ and offset b for toluene and cyclohexane targeting experimental SFEs. This approach yielded the highest predictive accuracy in terms of root mean squared error (RMSE) of 1.52 kcal/mol in transfer free energy for 16 small drug molecules among all 18 submissions in SAMPL9 challenge. The re-evaluation of the challenge set using multi-conformation strategies based on molecular dynamic (MD) simulations further reduces the prediction RMSE to 1.33 kcal/mol. At the same time, an additional evaluation of our PBSA method on SAMPL5 cyclohexane/water distribution coefficient ($logD_{cyc/wat}$) prediction revealed that our model outperformed COSMO-RS, the best submission model with $RMSE_{PBSA}$ = 1.88 versus $RMSE_{COSMO\text{-}RS}$ = 2.11 log units. Two external $logP_{tol/wat}$ and $logP_{cyc/wat}$ datasets that contain 110 and 87 data points, respectively, are collected for extra validation and

provide in-depth insight of the error source of PBSA method. Finally, to identify the best set of radius parameters which define the solute-solvent boundary, we adopted the following strategies: (1) the nonpolar term is fixed; (2) a genetic algorithm is applied to conquer the couplings between the radius parameters; (3) the new nonpolar term is reoptimized. The above three steps will be repeated until there is no further improvement on the model performance. Encouragingly, the newly tuned radii parameters conjugated with the ABCG2 charge model outperformed many widely used models and our previous results.

Table of Contents

List of Tables

List of Figures

List of Equations

1.0 Ligand-based and Structure-based Drug Design

At each phase of drug development, computational methods are implemented, which can substantially reduce the time and expense required for the design, screening, and optimization of novel drugs. The two primary strategies utilized in computer-aided drug design are ligand-based drug design (LBDD) and structure-based drug design (SBDD). SBDD involves the characterization of the topology and stereochemistry of the binding site, prediction of ligand-receptor binding poses, calculation of binding affinity and interpretation of key interactions that enhance affinity, and identification of residues that contribute favorably to the binding in the presence of a known target structure. LBDD, on the other hand, focuses on ligands that interact with the target of interest. In the absence of a target structure, the ligand-based approaches reveal the functional groups, topology, and physicochemical properties of the ligands served for pharmacological activity.

1.1 Ligand-based Drug Design

Quantitative structure-activity relationships (QSAR) and pharmacophore modeling are the commonly used approaches in LBDD. In order to develop a QSAR model, it is necessary to have a collection of ligands that possess experimental bioactivity data. The correlation between molecular descriptors derived from these compounds and bioactivities is established by creating suitable relationships. These descriptors can be either structural descriptors or descriptors of ligands' physicochemical properties.

The Hansch-Fujita method is a traditional two dimensional (2D) QSAR method that models the correlation among the electronic, hydrophobic and steric features of a molecule to its bioactivities using a succinct functional equation:[1]

$$\log\left(\frac{1}{C}\right) = k_1\pi - k_2\pi^2 + k_3\sigma + k_4E_s + k_5$$

Equation 1 Function form used in Hansch-Fujita method

where C is the effective concentration of the compound to produce pharmacological activity, π quantitatively describes the hydrophobic effect of the ligand (i.e. partition coefficient), σ is the Hammett electronic substituent constant and E_s is the steric substituent constant.

In addition to 2D QSAR modeling, three dimensional (3D) QSAR modeling constructs relationships between molecular descriptors and bioactivities from spatial information of ligands. This category of methods typically takes the bond orientation and electrostatic potential around molecules into account. Comparative Molecular Field Analysis[2] (CoMFA) is a kind of traditional 3D QSAR modeling method. This method uses a hypothetical molecular probe with sp^3 carbon van der Waals properties and a unit positive charge to capture electrostatic and van der Waals interactions at the lattice around the ligand molecule. From these interaction energies and the bioactivity data of the ligands, a matrix is constructed.

Subsequent principal component analysis (PCA) of this matrix allows the identification of interactions that contribute to biological activities and their spatial arrangement. However, there are some shortcomings to this approach. Firstly, using the natural conformation of the model compound as a template will incorrectly estimate the interaction strengths and regions in the bonded state, since the bound conformation is not necessarily its natural conformation. In addition, the interaction energies in this method are all calculated in the gas phase, ignoring solvent effects

during the binding process, including desolvation energies and electrostatic screening effects from highly dielectric solvents such as water.

1.2 Structure-based Drug Design

Molecular docking is a widely used technique in SBDD. Molecular docking enables the concurrent exploration of ligand binding poses and the prediction of binding affinity. Search algorithms for predicting binding poses are very diverse and can be multi-nested, but all search algorithms necessitate the use of a scoring function to evaluate the binding mode during the search process. Scoring functions used in molecular docking can be categorized into three types: (1) force-field-based functions, (2) empirical functions, and (3) knowledge-based functions.

Force-field-based scoring functions have a definite physical meaning by calculating bonded and non-bonded interactions for docking poses via potential energy functions from the molecular mechanics force field. For example, in the DOCK[3] scoring function, the AMBER force field[4, 5] parameters were used to evaluate van der Waals and electrostatic interactions, and a dielectric term was added to the electrostatic interaction function to take solvent effects into account. The empirical scoring functions empirically decompose the energy of the binding process and construct relationships between all energy terms and the overall binding energies from available experimental data. The knowledge-based scoring functions are based on the fact that in thermodynamic ensembles, the probability of an atom being in a particular energy level follows the Boltzmann distribution. And the Helmholtz free energy of interaction between pairs of atoms can be obtained from the following equation:

$$A_{ij}(r) = -k_{\mathrm{B}}T \ln g_{ij}(r)$$

Equation 2 Helmholtz free energy related to distribution function

where $g_{\mathrm{ij}}(r)$ is distribution function for pairs of atoms i and j. $g_{\mathrm{ij}}(r)$ can be obtained from a number of crystal structures with existing ligand binding poses by calculating the distance of each atom pairs.[6, 7]

Molecular dynamics (MD) simulations are also widely used in the assessment of binding affinities. One of them, alchemical method (also known as pathway method), is more theoretically rigorous, including thermodynamic integration (TI) and free energy perturbations (FEP). The difference between TI and FEP is mainly the method of obtaining the free energy difference. The theory of TI and FEP are elaborated below.

Given the free energy difference between two states (X, Y):

$$\Delta A = A_Y - A_X = -k_B T \ln \frac{Q_Y}{Q_X}$$

Equation 3 Helmholtz free energy difference between two state X and Y

where Q_X and Q_Y are partition functions of state X and Y, respectively. k_B is Boltzmann constant and T is thermodynamic temperature.

Equation 3 can be simplified as:

$$\Delta A = -k_{\mathrm{B}}T \left\langle \exp\left[-\frac{E_Y - E_X}{k_{\mathrm{B}}T}\right] \right\rangle$$

Equation 4 Averaged helmholtz free energy difference between two state X and Y

where $E_{\boldsymbol{Y}}$, $E_{\boldsymbol{X}}$ are total energies of state X and Y, respectively. In order to more accurately sample the difference in free energy for the transition from state X to Y, a series of intermediate states between X and Y are introduced:

$$\Delta A = -k_{\mathrm{B}} T \ln\left[\frac{Q_Y}{Q_N} \cdot \frac{Q_N}{Q_{N-1}} \cdot \frac{Q_{N-1}}{Q_{N-2}} \ldots \frac{Q_2}{Q_1} \cdot \frac{Q_1}{Q_X}\right]$$

Equation 5 Helmholtz free energy difference between two state X and Y between two state X and Y with multiple introduced intermediate states

In calculating the ligand-receptor binding free energies, the free energy difference of the binding process can be obtained with Equation 5 by setting state X to the presence of the ligand at the binding site and state Y to the complete "disappearance" of the ligand from the binding site. The FEP method obtains the final ΔA from the free energy difference of the change at each step of the pathway, while the TI method obtains ΔA from the following integrals:

$$\Delta A = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial E(\lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda$$

Equation 6 Helmholtz free energy difference between two state X and Y in TI

In addition to pathway methods, end-point free energy methods are also widely used in the prediction of ligand-receptor binding free energies because of efficiency. One of the representative methods is the molecular mechanics Poisson-Boltzmann Surface Area (MM/PBSA), the theory and details of this method will be discussed in detail in the next chapter.

2.0 The Poisson-Boltzmann surface area model

Electrostatic interactions govern the biological process and biomolecular recognitions. The important role of implicit solvent model is to describe solvent electrostatics and eliminate the degree of freedom of explicit solvent molecules, which will consume most of computational resources during quantum mechanics (QM) and MM based simulations. When developing and applying implicit solvent models, solvation free energy (SFE) is a critical property because it quantitively describes the solvation effects. Rigorously, the SFE of a solute is the reversible work to create a neutral cavity for the solute. This reversible work involves electrostatic polarization and van der Waals dispersion between solute and solvent molecules:[8]

$$\Delta G = \Delta G_{\text{cavity}} + \Delta G_{\text{vdW}} + \Delta G_{\text{elec}}$$

Equation 7 Decomposition of solvation free energy

where ΔG_{elec} is total electrostatic contribution to the SFE, which is usually denoted by polar contribution. The sum of $\Delta G_{\text{cavity}} + \Delta G_{\text{vdW}}$ are counted as all non-polar contributions raised from cavitation and van der Waals dispersion. The detailed theory of PBSA for electrostatic interactions and SFE calculations will be elaborated in the following subsection.

2.1 Theory of Poisson-Boltzmann equation

In the framework of the implicit solvent model, the solvent is modeled as a structureless continuous dielectric medium, while the solute is described as point charges located at the center of atoms and its surface. The interactions between solute atoms are usually calculated by molecular

mechanics force fields, while the solute-solvent electrostatic interactions can be derived from the Poisson equation of classical electrostatics:

$$\nabla \cdot \varepsilon(\mathbf{r})\nabla\varphi(\mathbf{r}) = -4\pi\rho(\mathbf{r})$$

Equation 8 Poisson equation

where $\varepsilon(\mathbf{r})$ is the spatial dielectric constant, $\varphi(\mathbf{r})$ is the total electrostatic potential, and $\rho(\mathbf{r})$ is the charge density from solute. In an electrolyte solution, free ions obey the Boltzmann distribution, and the nonlinear Poisson-Boltzmann equation is obtained by adding the Boltzmann term for the ions to the Poisson equation:

$$\nabla \cdot [\varepsilon(\mathbf{r})\nabla\varphi(\mathbf{r})] - \kappa(\mathbf{r})^2 \sinh\big(\varphi(\mathbf{r})\big) = -4\pi\rho(\mathbf{r})$$

Equation 9 Non-linear Poisson-Boltzmann equation

where $\kappa(\mathbf{r})^2$ is Debye-Huckel parameter:

$$\kappa^2 = \frac{8\pi e^2 I}{\varepsilon_{\mathrm{sol}} k_{\mathrm{B}} T}$$

Equation 10 Debye-Huckel parameter

where e is proton charge, I is ionic strength, $\varepsilon_{\mathrm{sol}}$ is solvent dielectric constant, k_{B} is Boltzmann constant, and T is thermodynamic temperature.

At very small $\varphi(\mathbf{r})$, the above nonlinear Poisson-Boltzmann equation can be further linearized into the following form:

$$\nabla \cdot [\varepsilon(\mathbf{r})\nabla\varphi(\mathbf{r})] - \kappa^2\varphi(\mathbf{r}) = -4\pi\rho(\mathbf{r})$$

Equation 11 Linear Poisson-Boltzmann equation

Linear PBE has an analytical solution when the molecule has a regular geometry (such as sphere), but in practice, regular molecular geometry is almost impossible, so numerical methods are often used to solve PBE. A variety of numerical methods have been developed for solving PBE, including the finite-element method[9-12], boundary element method[13-22] and widely used

finite-difference method[23-25]. Given that the finite difference method was primarily used to solve the PBE in this work (implement in DELPHI[26-29]), I will mainly discuss the details of this method. Before solving the PBE using the finite-difference method, a series of initial setups need to be performed first. First the spatial region is gridded, and partial charges are mapped at the finite-difference grid points, and the solute-solvent boundary is constructed by spheres determined by the atomic radius, which simultaneously determines the boundaries of the different dielectric regions. Afterwards the electrostatic potential is assigned outside the solute-solvent boundary by a Debye-Huckle expression to determine the boundary conditions:

$$\varphi_i = \Sigma\left(q_j e^{-\kappa r_{ij}}\right)/\varepsilon_{\text{sol}} r_{ij}$$

Equation 12 Debye-Huckle expression for bounday conditions

where q_j is the charge at the jth lattice point, r_{ij} is the distance of the jth charge from the ith lattice point, and the boundary electrostatic potentials are kept constant during the iteration to ensure that the calculation can converge.

Afterwards the linear PBE (Equation 11) is integrated in the space determined by the boundary conditions:

$$\iiint \vec{\nabla} \cdot \left(\varepsilon(\mathbf{r})\vec{\nabla}\varphi(\mathbf{r})\right) \mathrm{d}^3x - \iiint \kappa^2\varphi(\mathbf{r})\mathrm{d}^3x - 4\pi \iiint \rho(\mathbf{r})\mathrm{d}^3x = 0$$

Equation 13 Triple integration of Poisson-Boltzmann equation

The first integration is:

$$\iiint \vec{\nabla} \cdot \left(\varepsilon(\mathbf{r})\vec{\nabla}\varphi(\mathbf{r})\right) \mathrm{d}^3x = \Sigma\varepsilon_i(\varphi_i - \varphi_0)h$$

Equation 14 Integration of Poisson term

The second integration is:

$$\iiint \kappa^2 \varphi(\mathbf{r}) \mathrm{d}^3x = {\kappa_0}^2 \varphi_0 h^3$$

Equation 15 Integration of Boltzmann term

The third term integration is:

$$4\pi \iiint \rho(\mathbf{r}) \mathrm{d}^3x = 4\pi q_0$$

Equation 16 Integration of charge density term

where κ_0 is Debye-Huckel parameter at grid points, φ_0 is electrostatic potential at grid points, q_0 is charge at grid points. Finally, the Equation 11 becomes:

$$\varphi_0 = \left[\frac{(\sum_{i=1}^{6} \varepsilon_i \varphi_i) + 4\pi q_0/h}{(\sum_{i=1}^{6} \varepsilon_i) + (\kappa_0 h)^2} \right]$$

Equation 17 Linear Poisson-Boltzmann equation under finite difference framework

This equation can be expressed in linear form:

$$\mathbf{A}\varphi = \mathbf{b}$$

Equation 18 Linear form of linear Poisson-Boltzmann equation under finite difference framework

where the coefficient matrix **A** includes the dielectric constant and ion-dependent Boltzmann terms, **b** is the charge distribution matrix, and φ is the electrostatic potential to be solved.

A variety of commonly used numerical methods can be applied to solve the linear equation, including Jacobi relaxation[30], Gauss-Seidel[26], successive over-relaxation[28] (SOR), conjugate gradient[31] (CG).

In the content of this paper, we pay more attention to the SFE. In the framework of the finite-difference method, the induced charge at the solute surface can be obtained by the electrostatic potential at the solvent-solute boundary via Gauss's law. The reaction field energy is derived from the reversible work to induce surface charges. This energy is regarded as the electrostatic contribution to the solvation process of the solute.[32]

In addition to the electrostatic contribution, the SFE includes contributions from cavitation and van der Waals dispersion. The sum of these two contributions is proportional to the solvent accessible surface area of the solute:[33]

$$\Delta G_{SASA} = \gamma SASA + b$$

Equation 19 Solvent accessible surface area model for non-polar contribution in solvation free energy

2.2 Molecular mechanics Poisson-Boltzmann method for binding affinity prediction

In SBDD, the protein-ligand binding affinity prediction has been the focus of research in computer-aided drug design (CADD), as well as a major application scenario for molecular simulation methods. In order to accurately predict the binding affinity, researchers have developed a large number of empirical, physical, and machine-learning based computational approaches. The most widely used of these methods is molecular docking, an approach that predicts the ligand-protein binding pose along with the corresponding binding affinity. Affinity prediction methodologies includes empirical, knowledge-based, and molecular mechanics based scoring functions.[34] Although molecular docking methods are computationally efficient, the binding free energy prediction accuracy of this method is insufficient. With the development of high-performance graphics processing units (GPUs), pathway free energy methods have been widely used recently, including TI[35, 36] and FEP[37, 38]. Pathway methods are more theoretically rigorous, but require large amounts of computational resources, and need run longer simulations to achieve adequate sampling.

In addition to pathway methods, there are also end-point free energy methods. The most representative of this method is the MM/PBSA. Due to the computational efficiency and accuracy

of MM/PBSA, this method has been widely tested and applied in the last decade to predict the free energy of small molecule ligand-protein binding[39, 40] , protein-protein interactions[41-43] and nucleic acid complex[44-47].

2.2.1 Methodology of MM/PBSA for ligand-receptor binding free energy prediction

In the framework of MM/PBSA, the ligand-receptor binding free energy has definition as below:

$$\Delta G_{\text{bind}} = G_{\text{RL}} - G_{\text{R}} - G_{\text{L}}$$

Equation 20 Definition of bingding free energy

where G_{RL} is free energy of ligand-receptor complex, G_{R} and G_{L} is free energy of receptor and ligand, respectively. The binding free energy has the decomposition:

$$\Delta G_{\text{bind}} = \Delta H - T\Delta S = \Delta E_{\text{MM}} + \Delta G_{\text{sol}} - T\Delta S$$

Equation 21 Decomposition of binding free energy

where

$$\Delta E_{\text{MM}} = \Delta E_{\text{int}} + \Delta E_{\text{elec}} + \Delta E_{\text{vdW}}$$

Equation 22 Decomposition of molecular mechanics energy

$$\Delta G_{\text{sol}} = \Delta G_{\text{PB}} + \Delta G_{\text{SASA}}$$

Equation 23 Decomposition of solvation free energy

$$\Delta G_{\text{SASA}} = \gamma SASA + b$$

Equation 24 Non-polar contribution in solvation free energy

The molecular mechanical energy (ΔE_{MM}) can be further decomposed into internal energies ΔE_{int} (bond, angle, and dihedral energies), electrostatic energies ΔE_{elec}, and the van der Waals energies ΔE_{vdW}, whereas the SFE (ΔG_{sol}) can be divided into polar (ΔG_{PB}) and nonpolar terms

(ΔG_{SASA}), which depends on solvent accessible surface area SASA, the calculation of the polar terms of the solvation free energies defined here is detailed in Chapter 2.1. The conformational entropy of the binding process is usually obtained by normal mode analysis (NMA), which determines the entropy by constructing the partition function from the eigenvalues of the Hessian matrix, but NMA is very time-consuming, and usually only residues within about 10 Å around the ligand are intercepted for analysis[48, 49]. Except for NMA, there are some alternative methods such as weighted solvent accessible surface area[50] (WSAS) model, interaction entropy[51, 52] method to derive the entropy during ligand-receptor binding.

MM/PBSA calculations of binding free energies require MD simulations of the ligand-receptor complex and the sampling of a series of conformations. ΔE_{MM} can be obtained directly from MD simulations, whereas ΔG_{sol} requires the use of a sampled series of conformations to calculate solvation free energies. MD simulations typically use an explicit water model to obtain more accurate conformations and energies, however these conformations sampled from explicit water simulations may have inconsistencies in the description of energies when evaluated using PBSA.

There are two commonly used protocols for MD simulations in MM/PBSA calculations. The first protocol performs separate MD simulations for the receptor, ligand, and their complex and calculates G_{RL}, G_{R}, and G_{L}, respectively, while the second protocol performs a single MD simulation using the ligand-receptor complex and extracts the receptor and ligand from it.[53] The second protocol assumes that the ligand and receptor do not undergo significant conformational changes upon binding, but it avoids the large energy fluctuations associated with simulating ligands or proteins alone, resulting in more stable predictions.

2.2.2 Factors influence the performance of MM/PBSA

The performance of MM/PBSA can be improved in several aspects. The application of the PB equation to biological systems requires consideration of the effect of ionic concentration on the potential due to ion enrichment in highly charged regions on the surface of the molecule. Solving the nonlinear PB equation provides a more accurate description of the salt effects. The solute and the solvent are distinguished in PB calculations by boundaries with different dielectric constants inner and outer, with the solute dielectric constant usually set to 1 and the solvent dielectric constant using experimental values. Setting the solute dielectric constant to 1 is suitable for small molecules, but due to the presence of highly polarized residues in proteins and nucleic acids, assigning different dielectric constants to different residues or regions can improve the overall prediction performance of MM/PBSA.[54, 55]

In PB calculations, the charge method of describing the solute has a significant effect on the results, Xu et al. explored the effects of four different charge models on the prediction results of MM/PBSA and MM/GBSA, where RESP charge showed the best prediction accuracy on both MM/PBSA and MM/GBSA.[56] In addition, they tested the effects of molecular mechanics force field and different simulation time scale on the results and found that the best results were obtained using the AMBER ff03 force field, while a simulation duration of 2-4 ns gave more reliable results.[56] Su et al. tested the prediction performance of MM/PBSA and MM/GBSA using different sets of atomic radii, and their results showed that the MM/PBSA method using Bondi's radii had the best performance.[57]

Although a range of radius sets were tested, in implicit solvent models including PB/GB, atomic radii are treated as adjustable parameters. The MM/PBSA radius set used by the AMBER community is element-based, unparameterized radii. Therefore, to improve the accuracy of the

PBSA method, Sun et al.[40, 58] used a TI method to extract the electrostatic term in the SFE and then used it for the parameterization of the PB atomic radii. Such a parameterization strategy is due to the fact that the electrostatic and non-electrostatic terms of the solvation energy are difficult to be measured directly from experiments, so the more accurate TI method is used to directly parameterize the PB, and then the parameters of the non-polar terms are fitted using the parameterized $\Delta G_{\text{expt}} - \Delta G_{\text{PB}}$ as the non-polar term contributions to obtain the final PB model.

3.0 Development of new PBSA model for water and common organic solvents

3.1 Introduction

When using *in silico* simulations to study complex biomolecules, in addition to accurately modeling the biomolecule itself, how to model the solvent significantly affects the simulation results of the biomolecules. In conventional molecular dynamics simulations, water molecules are explicitly modeled, i.e., all water molecules have atomic-level details, but explicit water model consumes many computational resources for sampling the trajectories of water molecules. Even if, long-range electrostatic interactions are still approximated by summation-Ewald summation when dealing with larger systems. Therefore, implicit solvent models that treat the solvent as a homogenized dielectric medium were developed to minimize solvent degrees of freedom and quantitatively describe electrostatic interactions. This approximation also avoids numerical fluctuations arising from mean forces from the trajectories of explicit water molecules.

The precision of continuum models relies on the parameters employed to derive the solute charges, the dielectric constant of the solvent and solute, and the atomic radii that delineate the dielectric boundary. However, how to define solute-solvent boundary is a critical point in implicit solvent model. It is reliable to parameterize an implicit solvent model through experimental SFE, since SFE provides quantitative description of solvent effect.

In this chapter, we first constructed solvent models for two organic solvents using different non-polar parameters. These solvent models coupled with an earlier version of ABCG2 charge model[59] and previously tuned radii parameters[40]. Then I participated in Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) 9 challenge to predict partition coefficient, logP.

In addition to parameterization of non-polar model. I adopted an iteration process to further optimize the atomic radii targeting experimental HFEs based on newly developed ABCG2 charge model. The iteration process is shown in Figure 1: (1) the nonpolar term is fixed first; (2) a genetic algorithm (GA) is applied to conquer the couplings between the radius parameters; (3) the new nonpolar term is reoptimized.

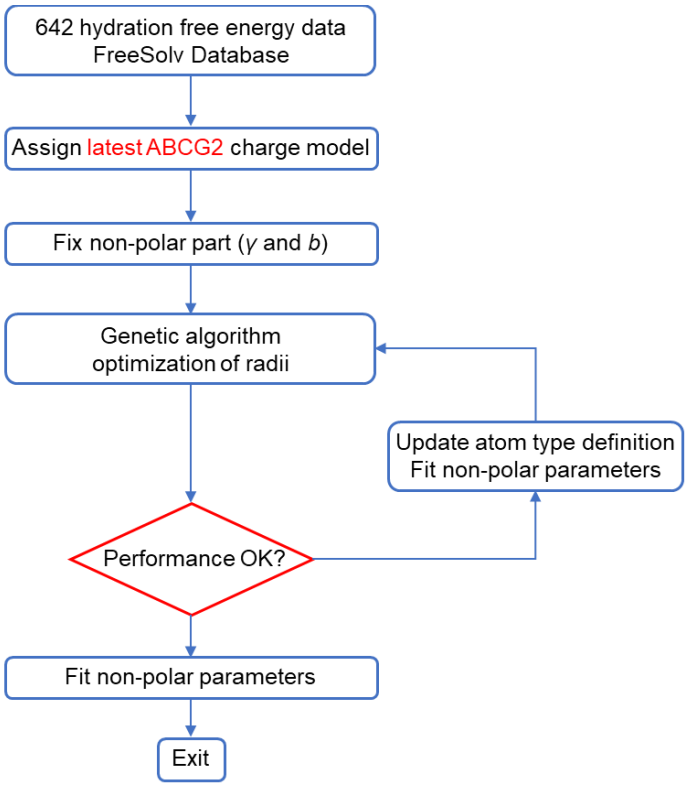

Figure 1 Iterative PBSA parameterization workflow

The above three steps will be repeated until there is no further improvement on the model performance. After several iterations, I tested the performance of the new set of parameters on SAMPL9 toluene/water $\log P_{tol/wat}$ and SAMPL5 cyclohexane/water distribution coefficient $\log D_{cyc/wat}$ dataset, respectively.

In SAMPL9 challenge, the organizers provided the simplified molecular-input line-entry system (SMILES) strings of 16 drug molecules as shown in Figure 2 and solicited blind prediction of $\log P_{tol/wat}$ on this set of molecules.[60] Unlike the logD predictions of the previous SAPML

challenge,[61, 62] the logP predictions do not require to account for the ionization state and the tautomer of the solute molecules. Therefore, it is unnecessary to re-model or introduce external empirical corrections for the charges. This also reduces the difficulty of making predictions based on the PBSA method in this study.

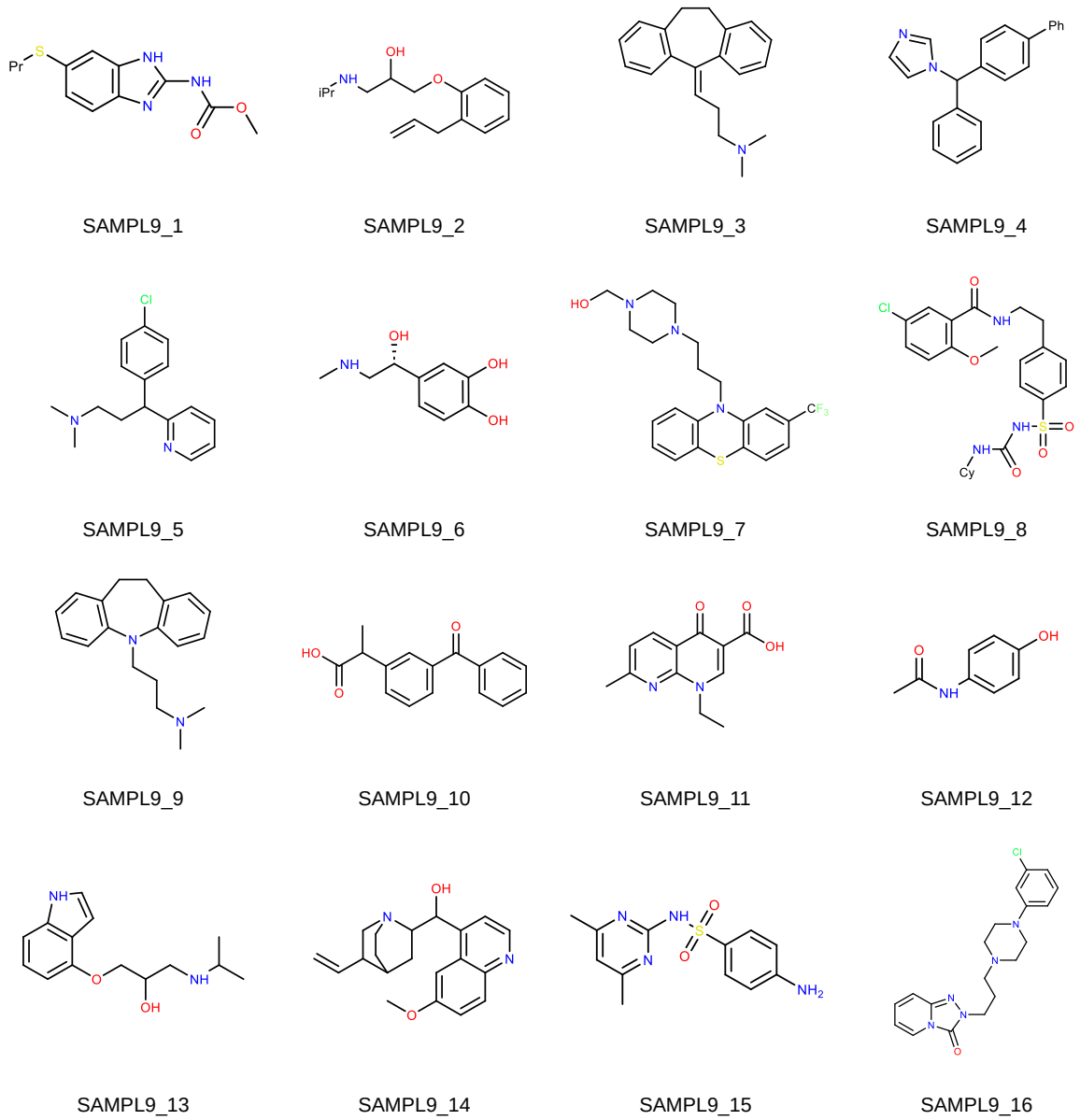

Figure 2 Structures of the 16 molecules involved in the SAMPL9 partition coefficient challenge

In most cases, $\log P_{i/j}$ is proportional to the transfer free energy of the solute molecule from solvent j to solvent i:

$$logP_{i/j} = \frac{-\Delta G_{j\to i}}{\mathrm{R}Tln10}$$

Equation 25 Partition coefficient definition definition from solvent j to solvent i

where i, j are two immiscible solvents, R is gas constant (8.314 $J{\cdot}mol^{-1}{\cdot}K^{-1}$), and T is thermodynamic temperature.

Transfer free energy can be derived from the difference between the SFE of the solute in these two solvents:

$$\Delta G_{j\to i} = \Delta G_i - \Delta G_j$$

Equation 26 Transfer free energy definition from solvent j to solvent i

In PBSA-based SFE predictions, electrostatic interactions are usually derived from PBE, and the free energy associated with cavitation and dispersion is usually described by SASA model.[33]

The solute-solvent boundary has uncertainty in implicit solvent models that include the PB method. This is due to the homogenization approximation of the solvent by implicit solvent models and the fact that the solute-solvent boundaries cannot be fully defined by atomic radii based on atomic number. This also implies that it is necessary to clarify the coupled charge method when discussing the definition of solute-solvent boundaries. Moreover, the separating measurements of the electrostatic and non-electrostatic contributions to the solvation effect are typically not available, hence it is difficult to optimize the electrostatic and non-electrostatic contributions individually.[63] Modeling the solvent effect as a whole may lead to overfitting and the unbalanced contributions of the two types of solvent effect.

Therefore, recently a series of studies were conducted on the development of high accurate PBSA model for SFE prediction,[40, 58] which were combined with the general AMBER force field 2 (GAFF2) and earlier developed ABCG2 charge model[59]. In this PBSA model, a set of atom radii for PB calculation were developed targeting the electrostatic (polar) contribution from thermodynamic integration (TI) calculations of hydration free energy (HFE); then the non-electrostatic (non-polar part) term was fitted targeting experimental values of HFE or SFE. This new PBSA parameters obtained a root mean square error (RMSE) of 1.05 kcal/mol on HFEs of 544 molecules.[40] Extending this method to solvent n-octanol yielded a prediction error of RMSE = 0.91 log units on $\log P_{oct/wat}$ calculations of 707 drug molecules in the ZINC database.[58] Note that the PB atomic radii optimized from HFE were directly utilized for SFE calculation in organic solvent, by this way only non-polar ΔG_{SASA} model needs to be redeveloped for individual organic solvents. In this study, I used the previously developed PB boundary definitions,[40, 59] and derived the solvent dependent parameters γ and b for toluene and cyclohexane solvents. The parameterization of γ and b targeted to fit experimental SFEs and using multiple conformations to avoid overfitting. In addition to blind testing on the SAMPL9 dataset, we collected 110 molecules of toluene/water logP for additional testing. Furthermore, we tested this PBSA model for cyclohexane using both the SAMPL5 $\log D_{cyc/wat}$ dataset (110 solutes) and an additional $\log P_{cyc/wat}$ dataset (87 solutes) compiled by us. In addition to parameterization of non-polar model, the new PB radii parameters tuned from GA demonstrate slightly better prediction performance on SAMPL9 and SAMPL5 dataset.

3.2 Method

3.2.1 Data Preparation

In training sets, all the experimental data of SFE in organic solvents, in this work toluene and cyclohexane, were taken from the Minnesota Solvation Database v2012,[64] and the experimental data of HFE were taken from the FreeSolv v0.52 database.[65] All the initial structures from Minnesota Solvation Database v2012 are in xyz format, and all initial structures from FreeSolv v0.52 database are in mol2 format. All the structures were imported to Schrödinger Maestro v11.2[66] for visual inspection and were saved in mol2 files for further processing. In total 47 molecules have both HFE and SFE in toluene, and 83 molecules have both HFE and SFE in cyclohexane.

The initial structures of SAMPL9 molecules are converted from SMILES strings to mol2 files by Open Babel 3.1.0 with the "-gen3d" option.[67] The additional logP test set data were taken from the works done by Leo *et al*,[68] Shalaeva *et al*,[69] and Byrne *et al*,[70] and the structures were downloaded from PubChem as sdf files and converted to mol2 files by Open Babel 3.1.0.[67]

The modified module of ANTECHAMBER[71] in AMBER Tools was utilized to assign GAFF2 topologies and ABCG2 charges.

3.2.2 Molecular Dynamic Simulation

Selected solute molecules were solvated in explicit water molecules with at least 15 Å distance from any solute atom to the edges of cubic simulation box. The solute molecules were treated with the GAFF2 force field parameters.[72] The adopted water model was TIP3P. The

periodic boundary condition and the NPT ensemble were applied with P = 1.0 atm and T = 298.15 K. The time step was set to 1.0 fs and the total simulation time was 10.0 ns for each system. The software AMBER18[73] was utilized for MD simulations.

3.2.3 PBSA Calculation

All PB calculations were performed using Delphi V4 release 1.1.[29, 74] The salt concentration was set to 0 mol/L; the grid spacing was set to 1.2 grids/Å; the percentage of the object longest linear dimension to the lattice linear dimension was set to 80%; and the boundary condition was set as coulombic boundary. The probe radius was 1.4 Å. The internal dielectric constant was always set to 1.00, and the dielectric constant of solvent was set to 80.00 for water, 2.3741 for toluene, and 2.0165 for cyclohexane, respectively. Calculation mode was set as reaction field energy, which is regarded as the electrostatic component of SFE ΔG_{PB}. The radii from Sun et al. were listed in Appendix A, and the radii tuned from GA were listed in Appendix B. The solvent accessible surface area $SASA$ was generated by an internal program called MS[50] using Bondi's van der Waals radii[75] and water probe (radius of 1.4 Å). This program is also available upon request. SASA was used to derive non-electrostatic term ΔG_{SASA} using Equation 19.[50]

3.2.4 Toluene and Cyclohexane Modeling

The same PB radius parameters derived using hydration free energies in our previous work[40, 58] are directly applied in toluene and cyclohexane, therefore, the only parameters of toluene and cyclohexane that differ from those of water are γ and b of Equation 19 in addition to the dielectric constant. The parameterization of γ and b can be obtained directly by linear regression

analysis (single data point per solute), but given the limited amount of data in organic solvents, we used the multi-conformation approach when conduct the linear regression process (multiple data points per solute). All conformations are generated by the "-conformer" option of the Open Babel software through genetic algorithm,[67] with the generation criterion being set to minimum energy and the maximum number of generated conformations being set to 20. The advantage of generating multiple conformations through Open Babel is that the number of conformations depends on the degree of freedom of the molecule. Therefore, the modeling of toluene and cyclohexane is the fitting of the following linear equations:

$$\Delta G_{SFE,M}^{expt} - \Delta G_{PB}^{calc}\left(\mathbf{R}_{M_k}\right) = \gamma_s SASA\left(\mathbf{R}_{M_k}\right) + b_s$$

Equation 27 Mathematical expression of the non-polar term to be fitted

where $\mathbf{R}_{M_k}$ is the *k*th conformation of molecule M, s is organic solvent, here represent for either toluene or cyclohexane.

3.2.5 Calculate logD from logP

Only one ionization state is considered for the logD calculation from logP. The modified Henderson-Hasselbalch equation is used.

$$\log D = \log P - \log(1 + 10^{pK_a - pH})$$

Equation 28 Modified Henderson-Hasselbalch equation for basic solutes logD calculation

$$\log D = \log P - \log(1 + 10^{pH - pK_a})$$

Equation 29 Modified Henderson-Hasselbalch equation for acidic solutes logD calculation

Equation 28 is used for basic solutes and Equation 29 is used for acidic solutes. For amphipathic molecules, acidic pK_a is adopted as the correction factor.

3.2.6 Thermodynamic Integration Simulation Protocol

We compared the PBSA method with TI method on SAMPL9 and SAMPL5 dataset, and the TI calculation details were elaborated in this section. The alchemical enhanced sampling (ACES) method,[76] proposed by Lee *et al* and implemented in the GPU version[77-79] of TI modules in AMBER22, was employed for HFE and SFE calculations.

The TLEAP module in AMBER22 was used to generate all solute-solvent boxes. For a solute molecule being solvated in water, the minimum distance between any solute atoms and an edge of the water box was set to 15 Å. Similarly, a solute molecule was solvated in the cubic box of toluene or cyclohexane utilizing TLEAP. Note that toluene solvent box which has a dimension of 33.623 Å and cyclohexane solvent box which has a dimension of 39.418 Å were first created following the standard procedure as detailed in our previous publication.[8]

The organic solute-solvent system was first subjected to an initial equilibration for 200 ps using the CPU-TI at $\lambda = 0.01592$. A 2 ns MD simulation was conducted for each of the 9 λ windows (0.01592, 0.08198, 0.19331, 0.33787, 0.5, 0.66213, 0.80669, 0.91802, 0.98408). For the first λ window ($\lambda = 0.01592$), the initial configurations were sampled from the CPU-TI, while the initial configurations for the other eight λ windows were obtained from the preceding λ window. Following the system setup, periodic boundary condition and the isothermal-isobaric NPT ensemble were produced in all simulations. Using Langevin dynamics to maintain the temperature at 298K, with the collision frequency (gamma_ln) set to 2.0 ps^{-1}. The pressure was kept at 1.01325 bar with Monte Carlo barostat and the pressure relaxation time being set to 5.0 ps. Disable the SHAKE constrains for solute and set time step to 1fs. It is pointed out that the purpose of running GPU-TI here was to provide an equilibrium system for the ACES simulation protocol. Specifically, we enlarged the simulation boxes for the organic solvents about 15-40% from the last snapshots

of the GPU TI runs for the λ = 0.5 window. The new simulation boxes have dimensions around 46.0 Å.

All the subsequent ACES simulations were based on the new simulation boxes following the same protocol of GPU-TI except that the van der Waals and electrostatic interactions were scaled by smoothstep soft-core potential[80, 81] with switching function $W(r_{ij})$:

$$r_{ij}^{VDW}(\lambda;\alpha^{VDW}) = \left[r_{ij}^{n} + W(r_{ij}) \cdot \alpha^{VDW} \cdot S_P(\lambda) \cdot \sigma_{ij}^{n}\right]^{1/n}$$

Equation 30 Smoothstep soft-core potential for van der Waals interactions

$$r_{ij}^{Elec}(\lambda;\alpha^{Elec}) = \left[r_{ij}^{m} + W(r_{ij}) \cdot \alpha^{Elec} \cdot S_P(\lambda) \cdot \sigma_{ij}^{m}\right]^{1/m}$$

Equation 31 Smoothstep soft-core potential for electrostatic interactions

The lower boundary of the switching function $W(r_{ij})$ was set to 8 Å and the upper boundary was set to 10 Å. Additionally, the internal VDW interactions scaling within soft-core region were disabled by setting the gti_add_sc to 5. Nine equally-spaced λ windows (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9) were applied to decouple the endpoint states. Hamiltonian exchange between different λ windows was performed every 100000 steps under the REMD[76] framework to achieve the enhanced sampling. It is pointed out that the above ACES protocol is same as that reported by Lee et al.[76] with an aim to achieve the consistent performance. The free energies were derived from unweighted integration of the alchemical pathway as below:

$$\Delta G = G(\lambda = 1) - G(\lambda = 0) = \int_0^1 \left\langle \frac{\partial V}{\partial \lambda} \right\rangle_\lambda \cdot d\lambda \approx \sum 0.1 \times \left\langle \frac{\partial V}{\partial \lambda} \right\rangle_i$$

Equation 32 Unweighted integration alone the alchemical pathway

Three independent ACES based GPU-TI runs were performed for each solute, with 2 ns MD simulations for each λ windows. For each MD run, the beginning 0.5 ns simulation was considered as the equilibration phase and excluded from the later free energy analysis. The final HFE and SFE were then derived from the arithmetic average of the three independent TI runs,

while the standard deviation of the three independent runs was calculated to measure the precision of the protocol. The corresponding logP was calculated from HFE and SFE using Equation 25, and the logD was calculated from logP using Equations 28 and 29.

3.2.7 Ab initio logP Calculation

We used quantum mechanics (QM) based SMD model implemented in the Gaussian 16[82] software to derive the logP benchmark for our model. The principle of SMD derived logP is also based on the transfer free energy as Equation 25. Geometry optimization in the liquid phase at the B3LYP/6-31G* level of theory was first performed prior to SMD calculations, with the solvent specified directly by keywords; then the optimized geometries were read out to perform single point calculations in gas phase at the same level of theory. The energy difference between the liquid and gas phase is regarded as SFE.

3.2.8 Globally tune the PB parameters

GA is an efficient stochastic optimization method that has been widely applied to minimization problems because it is ideally suited for multiple-dimensional global search problems where the search space contains multiple local minima and the search variables may or may not be correlated.[83] All molecules were treated with GAFF2 force field parameters and new ABCG2 charge model. We started the search with the fixed non-polar term, and ΔG_{PB} was determined from $\Delta G_{expt} - \Delta G_{non-polar}$. The initial atom types were element-based, and atom types would be updated according to the molecules have larger errors. After each round of GA

search, the non-polar parameters would be re-fitted using the latest radii. Population size in GA search was varied based on the number of atom types, and the fitness function was RMSE.

3.3 Results and Discussion

3.3.1 Modeling of Toluene and Cyclohexane

With the multi-conformation strategy described above applied on the training sets, the descriptors (γ and b) of toluene and cyclohexane for SASA model were derived: $\gamma_{tol} = -0.023556$, $b_{toluene} = 4.40$ and $\gamma_{cyc} = -0.024237$, $b_{cyc} = 4.64$. $\Delta G_{SFE,M}^{expt} - \Delta G_{PB}^{calc}$.

3.3.2 SAMPL9 Toluene/Water logP Blind Prediction

As required by the SAMPL9 organizer, we submitted predicted transfer free energies $\Delta G_{tol/wat}$ of the 16 drug molecules before the deadline. Note that only a single conformation (with minimum energy) automatically generated by Open Babel for each drug molecule was used for the PBSA calculation of HFEs in water and SFEs in toluene. Based on the analysis result on all 18 submissions provided by the organizer (https://github.com/samplchallenges/SAMPL9/tree/main/logP/Analysis/prelim_analysis), our submission achieved the lowest overall RMSE of 1.52 kcal/mol. After the completion of this blind prediction contest, we also applied MD simulation conjugated with PBSA to re-calculate the transfer free energy $\Delta G_{tol/wat}$ for the 16 molecules and summarized the results in Table 1 and Figure 3. Table 2 reports the calculated HFE, SFE in toluene and the transfer free energy derived from

the difference between HFE and SFE. Figure 3 shows the correlation between experimental and calculated transfer free energies. The re-calculated transfer free energies achieved a better RMSE of 1.33 kcal/mol and the Pearson correlation coefficient (R) of 0.94.

In addition to the PBSA parameters and charge model that can affect the prediction accuracy of SFEs and corresponding transfer free energies, the adopted methodology and protocol for conformation generation is another factor affecting the prediction performance. The prediction error of Compound 8 significantly reduced after being treated by MD simulations compared to the value in our submission with single-conformation strategy. Also, Compound 8 has the maximum solvent accessible area, 709.35 $Å^2$ (B3LYP/6-31G* optimized geometry), and greater flexibility. Therefore, we focused on Compound 8 to investigate the conformational effect on the prediction accuracy of transfer free energies and illustrate the results in Figure 4. The error of the calculated transfer free energies from the experimental value were evaluated using 10, 20, 50 and 100 conformations. Conformations of Compound 8 were generated through three different ways: MD simulations, genetic algorithm using Open Babel,[67] and Omega using mmff94smod_NoEstat force field parameters.[84] The conformations generated by MD simulation yielded the lowest computational errors among the three methods, and demonstrated a trend that the error approached to zero as shown in the panel B of Figure 4 (from -0.76 kcal/mol on 10 conformations to -0.52 kcal/mol on 100 conformations). The magnitude of the computational error from the conformations generated by Omega also decreased as the number of conformations increases, just as the result from MD simulations, however, there was a much long way to go before the error could reduce to certain low threshold. In contrast, the computational error from the conformations generated by Open Babel fluctuated around -2.0 kcal/mol as the number of conformations changed, with the magnitude of error higher than that from MD simulation (around -0.6 kcal/mol) but lower

than that from Omega (from -6.4 kcal/mol on 10 conformations to -5.0 kcal/mol on 100 conformations).

Except for Compound 8, other compounds which have prediction errors close to 2 kcal/mol should also be noticed. The prediction error of Compounds 1, 6 and 11 most likely arose from the formation of intramolecular hydrogen bond. As reported by Shalaeva *et al*,[69] the difference between $logP_{oct/water}$ and $logP_{tol/water}$ is a potential descriptor to indicate the formation of intramolecular hydrogen bond. Molecular fragments that have the structural potential to form intramolecular hydrogen bonds in 6- or 7-membered rings are screened in a highly dielectric medium such as water ($\varepsilon = 80$) and form intermolecular hydrogen bonds with water molecules. Such molecule first undergoes desolvation during water-toluene phase transfer, and then, due to the jump in the dielectric environment, is more inclined to form intramolecular hydrogen bonds, thus decreasing the molecular polarity and increasing solubility. As such, Compounds 1 and 11 adopt different conformations in the two different solvents, and the large prediction errors of transfer free energies of the two compounds may be due to using the same set of conformations. Unfortunately, it is necessary to use the same set of conformations for the SFE calculation in two different solvents to achieve the best error cancellation.[85]

Since the TI method demonstrates high accuracy in free energy calculations, we also employed TI method to calculate the $logP_{tol/wat}$ for the 16 molecules in SAMPL9 dataset. The result of TI-calculated transfer free energies versus the experimental values was shown in Figure 5, and the detailed data were summarized in Appendix Table 3. The overall prediction error of TI in terms of RMSE was 2.11 kcal/mol, and the Pearson correlation coefficient of TI predictions was 0.92. Note that the prediction error of TI was slightly larger than that of the COSMO-RS method, but smaller than those of the other 11 submissions in this SAMPL9 challenge.

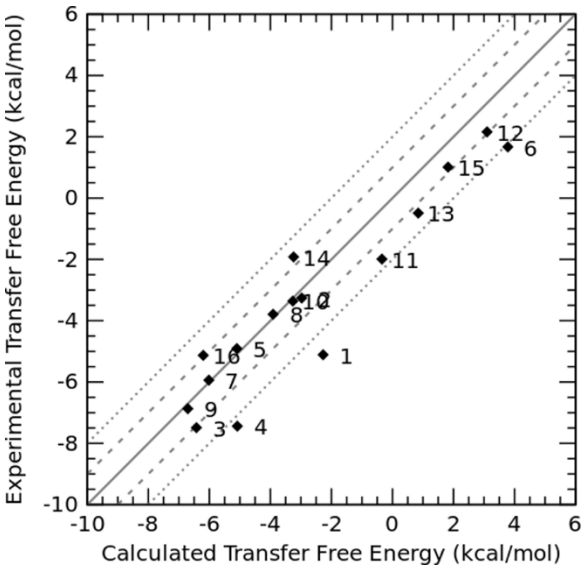

Figure 3 Experimental transfer free energy versus calculated transfer free energy using PBSA method for 16 drug molecules in SAMPL9 challenge

Table 1 Detailed experimental and calculated transfer free energies, calculated hydration free energies in water and solvation free energies in toluene using the PBSA method. The overall Pearson correlation coefficient (R), mean signed error (MSE), mean unsigned error (MUE) and root mean square error (RMSE) were listed for 16 SAMPL9 compounds

Compound	Experiment ΔG (kcal/mol)	Hydration ΔG (kcal/mol)	Solvation ΔG (kcal/mol)	Transfer ΔG (kcal/mol)
1	-5.11	-13.48	-15.75	-2.27
2	-3.26	-11.41	-14.38	-2.97
3	-7.49	-6.14	-12.57	-6.42
4	-7.44	-10.99	-16.07	-5.08
5	-4.91	-8.27	-13.37	-5.10
6	1.67	-18.37	-14.59	3.78
7	-5.94	-13.81	-19.83	-6.02
8	-3.79	-18.49	-22.40	-3.91
9	-6.87	-5.68	-12.39	-6.71
10	-3.36	-10.80	-14.07	-3.26
11	-1.99	-13.76	-14.10	-0.35
12	2.16	-15.42	-12.32	3.10
13	-0.49	-18.04	-17.19	0.85
14	-1.92	-14.16	-17.40	-3.24

15	1.01	-19.57	-17.75	1.82
16	-5.13	-12.52	-18.72	-6.20
R				0.94
MSE				0.68
MUE				1.03
RMSE				1.33

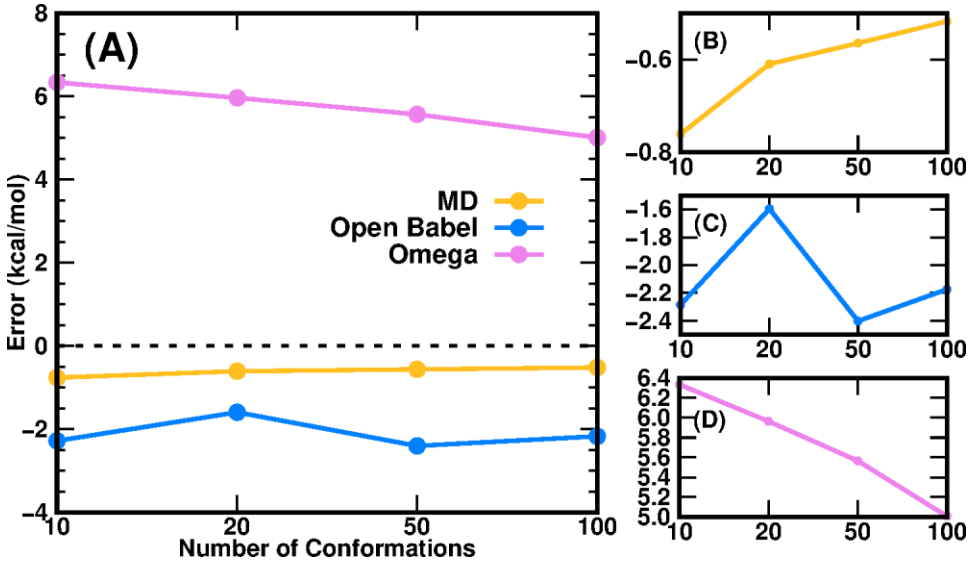

Figure 4 The relationship between the numbers of conformations and the prediction errors of the transfer free energies using the PBSA method. Figure 4A. Prediction errors of three conformation generation methods; Figure 4B 4C and 4D are re-ranged plots for individual methods.

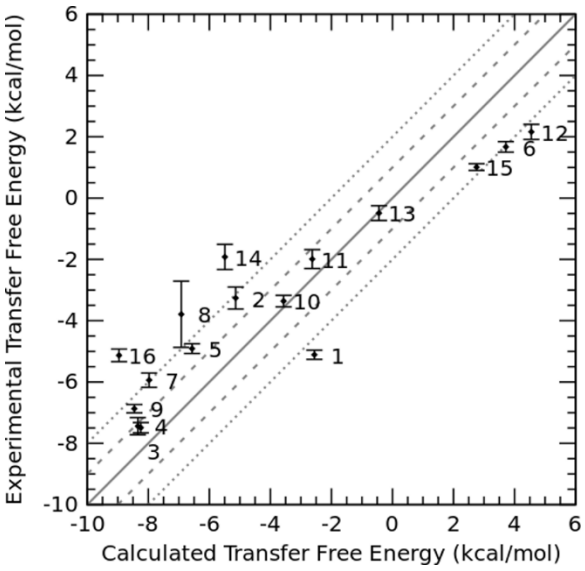

Figure 5 Experimental transfer free energy versus calculated transfer free energy using the TI method for 16 drug molecules in SAMPL9 challenge. The uncertainties of calculated transfer free energy were standard deviations derived from three independent TI runs

3.3.3 SAMPL5 Cyclohexane/Water logD Prediction

In addition to modeling toluene for the SAMPL9 challenge, we also modeled cyclohexane and tested the cyclohexane/water logD prediction for 53 organic molecules in SAMPL5 challenge as well as the cyclohexane/water logP prediction for 87 molecules we collected.[61] The prediction results of comparing our PBSA method with the best-ranked SAMPL5 submission from Klamt *et al* using COSMO-RS method[86] (hereafter referred to as COSMO-RS) were summarized in Figure 6 and Table 2. Panel A in Figure 6 shows the correlation between experimental logD and PBSA calculated logD, and panel B illustrates the correlation between experimental logD value and the initial submitted logD using COSMO-RS method by Klamt *et al*.[86] The overall RMSE prediction error of our PBSA method is 1.88 log units, which is smaller than that of COSMO-RS (RMSE = 2.11 log units). It is worth noting, however, that the logD values calculated by the PBSA method

were corrected from logP values using Equations 28 and 29, and the solutes' pK_a values were borrowed from Klamt *et al.* According to their report, the pK_a values were predicted using the *ab initio* COSMOtherm program.[87] In addition to the COSMOtherm, *ab initio* calculations using the Schrödinger Jaguar pK_a module[88] can yield comparable accurate predictions (RMSD within 0.2-0.5 pK_a units) for logD predictions. As shown in Figure 6, the yielded large prediction errors by the PBSA method were mainly for some neutral and basic molecules, among which Compounds 74 and 82 also had large prediction errors by the COSMO-RS method. Regarding to Compound 74, based on our experience in developing the PBSA method, the conformation of polyhydroxylated compounds represented by glycerol has a significant effect on the prediction accuracy, and the use of a multi-conformation approach sampled by MD simulations usually leads to a predicted SFE of such molecules closer to the experimental value. The prediction error for SAMPL5_083 raises from using a less dominate tautomer as reported by Klamt *et al*.[86] Similarly, we conducted TI calculations on the SAMPL5 $logD_{cyc/wat}$ dataset for comparison. We also adopted the predicted pK_a (summarized in Table 2) to correct the TI calculated logP to obtain logD. The performance of TI predictions was illustrated in Figure 7 and the detailed data were listed in Appendix Table 4. The overall prediction error of TI in terms of RMSE was 2.15 log units, which was comparable with the COSMO-RS method.

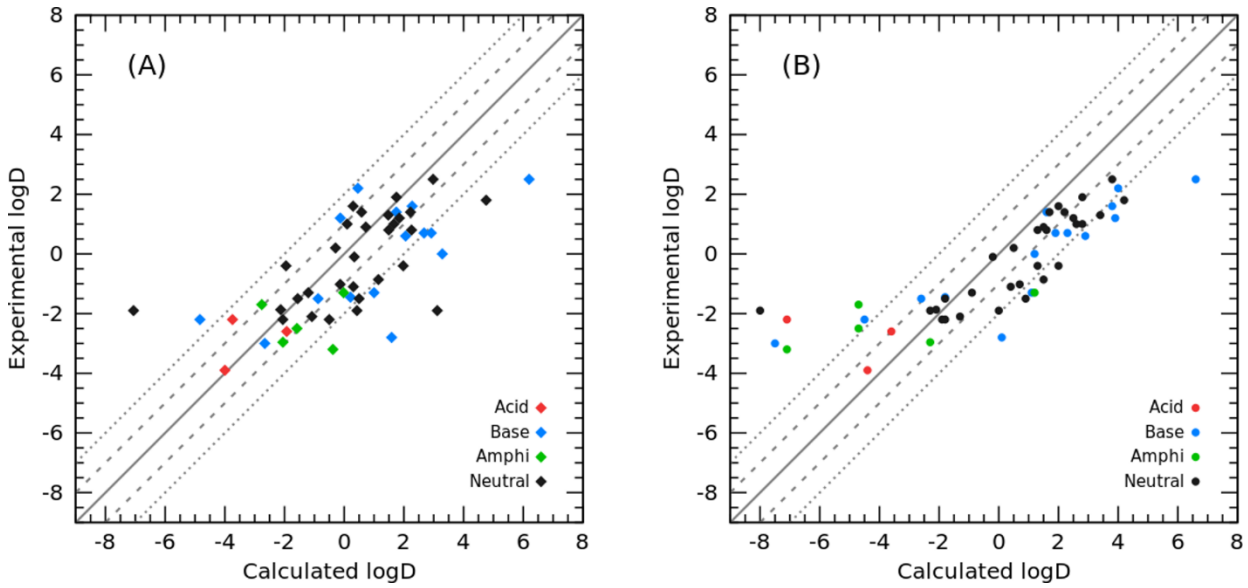

Figure 6 Correlation between experimental and calculated logD. Figure 6A Calculated with PBSA method (this work); Figure 6B Calculated using the COSMO-RS method.

Table 2 Experimental logD, calculated logP and logD values of the PBSA and COSMO-RS methods. The pKa values adopted to correct the ionization effect were from Klamt et al. If the molecule is an amphipathic molecule, the acidic pKa was used to compute the correction factor.

Compound	Expt	pK_a		Calc logP		Calc logD	
	logD	Acid	Base	COSMO-RS	PBSA	COSMO-RS	PBSA
SAMPL5_002	1.40			1.70	0.58	1.70	0.58
SAMPL5_003	1.90			2.80	1.75	2.80	1.75
SAMPL5_004	2.20		6.85	4.10	0.57	4.00	0.46
SAMPL5_005	-0.86			1.50	1.15	1.50	1.15
SAMPL5_006	-1.02			0.70	-0.14	0.70	-0.14
SAMPL5_007	1.40		7.02	1.80	1.90	1.60	1.74
SAMPL5_010	-1.70	4.86	6.03	-2.20	-0.22	-4.70	-2.76
SAMPL5_011	-2.96	4.01	4.55	1.10	1.33	-2.30	-2.06
SAMPL5_013	-1.50			0.90	0.50	0.90	0.50
SAMPL5_015	-2.20	4.35		-4.00	-0.70	-7.10	-3.74

SAMPL5_017	2.50			3.80	2.98	3.80	2.98
SAMPL5_019	1.20		6.55	4.00	-0.08	3.90	-0.13
SAMPL5_020	1.60			2.00	0.30	2.00	0.30
SAMPL5_021	1.20			2.50	1.85	2.50	1.85
SAMPL5_024	1.00			2.60	1.66	2.60	1.66
SAMPL5_026	-2.60	4.73		-0.90	0.74	-3.60	-1.93
SAMPL5_027	-1.87			-2.10	-2.13	-2.10	-2.13
SAMPL5_033	1.80			4.20	4.76	4.20	4.76
SAMPL5_037	-1.50		8.17	-1.70	-0.04	-2.60	-0.88
SAMPL5_042	-1.10			0.40	0.31	0.40	0.31
SAMPL5_044	1.00			2.80	0.10	2.80	0.10
SAMPL5_045	-2.10			-1.30	-1.08	-1.30	-1.08
SAMPL5_046	0.20			0.50	-0.29	0.50	-0.29
SAMPL5_047	-0.40			2.00	-1.95	2.00	-1.95
SAMPL5_048	0.90			1.50	0.72	1.50	0.72
SAMPL5_049	1.30			3.40	1.48	3.40	1.48
SAMPL5_050	-3.20	7.24	3.86	-6.70	0.01	-7.10	-0.38
SAMPL5_055	-1.50			-1.80	-1.56	-1.80	-1.56
SAMPL5_056	-2.50	8.09	-4.19	-4.60	-1.51	-4.70	-1.59
SAMPL5_058	0.80			1.60	1.49	1.60	1.49
SAMPL5_059	-1.30			-0.90	-1.20	-0.90	-1.20
SAMPL5_060	-3.90	4.95		-1.90	-1.55	-4.40	-4.00
SAMPL5_061	-1.45		7.03	-1.70	0.36	-1.80	0.21
SAMPL5_063	-3.00		9.05	-5.80	-1.00	-7.50	-2.66
SAMPL5_065	0.70		8.43	3.40	3.99	2.30	2.92
SAMPL5_067	-1.30		8.85	2.60	2.46	1.10	1.00
SAMPL5_068	1.40			2.20	2.22	2.20	2.22
SAMPL5_069	-1.30	8.91	7.74	1.70	-0.01	1.20	-0.02
SAMPL5_070	1.60		9.32	5.80	4.20	3.80	2.28

SAMPL5_071	-0.10		-0.20	0.34	-0.20	0.34
SAMPL5_072	0.60	8.62	4.10	3.30	2.90	2.06
SAMPL5_074	-1.90		-8.00	-7.06	-8.00	-7.06
SAMPL5_075	-2.80	8.50	1.30	2.72	0.10	1.59
SAMPL5_080	-2.20		-1.90	-2.06	-1.90	-2.06
SAMPL5_081	-2.20	8.28	-3.60	-3.90	-4.50	-4.84
SAMPL5_082	2.50	8.11	7.40	6.98	6.60	6.20
SAMPL5_083	-1.90		-2.30	3.12	-2.30	3.12
SAMPL5_084	0.00	8.18	2.00	4.13	1.20	3.29
SAMPL5_085	-2.20		-1.80	-0.50	-1.80	-0.50
SAMPL5_086	0.70	9.52	4.00	4.80	1.90	2.68
SAMPL5_088	-1.90		0.00	0.43	0.00	0.43
SAMPL5_090	0.80		1.30	2.26	1.30	2.26
SAMPL5_092	-0.40		1.30	1.98	1.30	1.98
R			0.79	0.55	0.85	0.68
MSE			1.05	1.26	0.49	0.71
MUE			1.79	1.84	1.65	1.44
RMSE			2.26	2.34	2.10	1.88

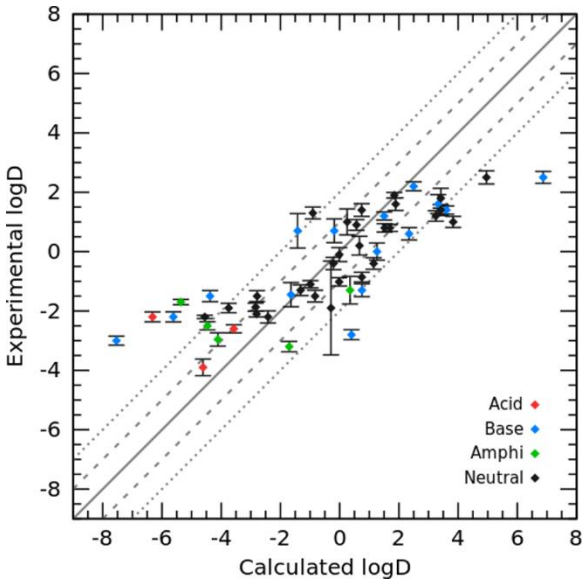

Figure 7 Correlation between experimental logD and TI calculated logD. Uncertainties were standard deviations from three independent TI runs.

3.3.4 Test of the PBSA method on Additional logP Datasets

Finally, to further validate the developed PBSA models for toluene and cyclohexane, additional test molecules were collected to predict the $logP_{tol/wat}$ and $logP_{cyc/wat}$ values. For 110 organic molecules in toluene, the PBSA method achieved an RMSE of 1.83 log units. In contrast, the QM-based SMD method calculated at the B3LYP/6-31G* level of theory had a prediction error of 2.31 log units. The comparison results were shown in a scatter plot between the experimental logP and calculated logP (Figure 8).

Interestingly, there was a strong agreement between the PBSA method and the SMD method for molecules with large prediction errors, which are: 8-Hydroxyquinoline, 2-Methyl-8-Quinolinol, Bromothymol blue, and Schiff base. Some others with larger errors by the PBSA method are phosphorus-containing molecules, for which the phosphorus-related bond charge correction parameters were not adequately adjusted for the ABCG2 charge model. Still other six

molecules with experimental logP values between 3.0 - 4.0 have systematic errors in the PBSA calculations, but not in SMD calculations. Examination on their structures revealed that most of them are halogen-substituted benzenes except for cyclohexene. This systematic error is probably due to the inability of the implicit solvent model described by the dielectric constant to adequately model the π-π interactions arising from the benzene rings in the toluene and solute molecules. Of course, the systematic error may also come from the inadequate description of the σ-hole effect by the ABCG2 charge model. This systematic error in structure-dependent SFE calculations recurs in the PBSA model and has attracted our attention to deal with those "difficult" molecules in the future.

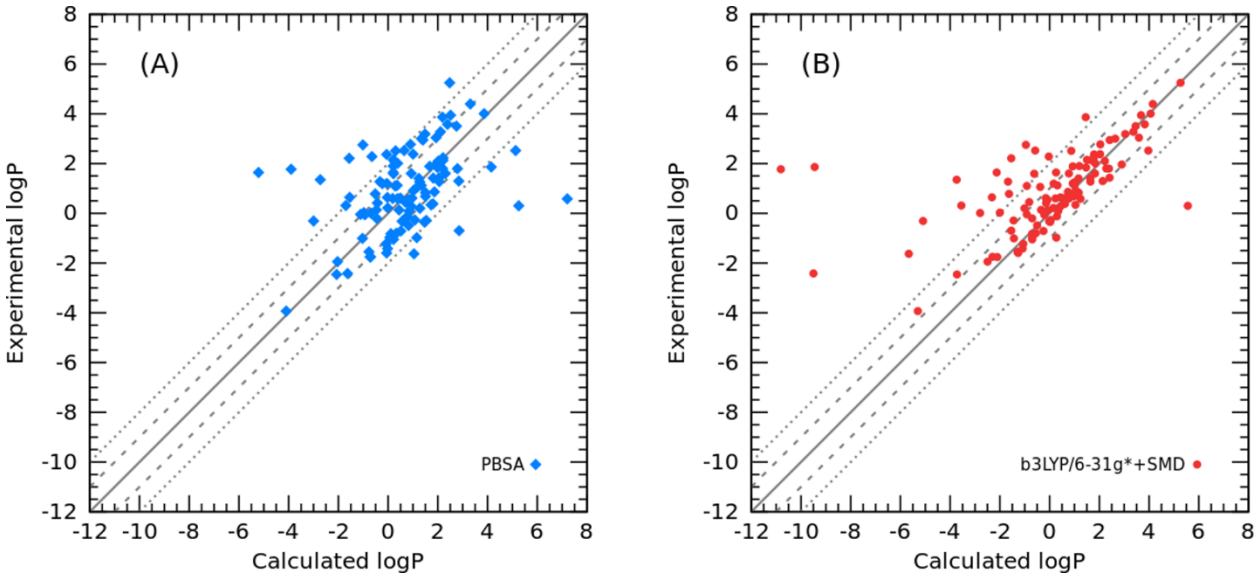

Figure 8 Correlation between experimental and calculated $logP_{tol/wat}$. Figure 8A Calculated $logP_{tol/wat}$ using PBSA method; Figure 8B Calculated $logP_{tol/wat}$ using SMD method.

As to the 87 organic molecules in the additional cyclohexane test set, the PBSA method achieved an RMSE of 1.11 log units, which is slightly larger than that of the SMD method (RMSE=0.99) as shown in Figure 9. Nevertheless, the prediction error is much lower than the RMSE of logD prediction in SAMPL5 challenge.

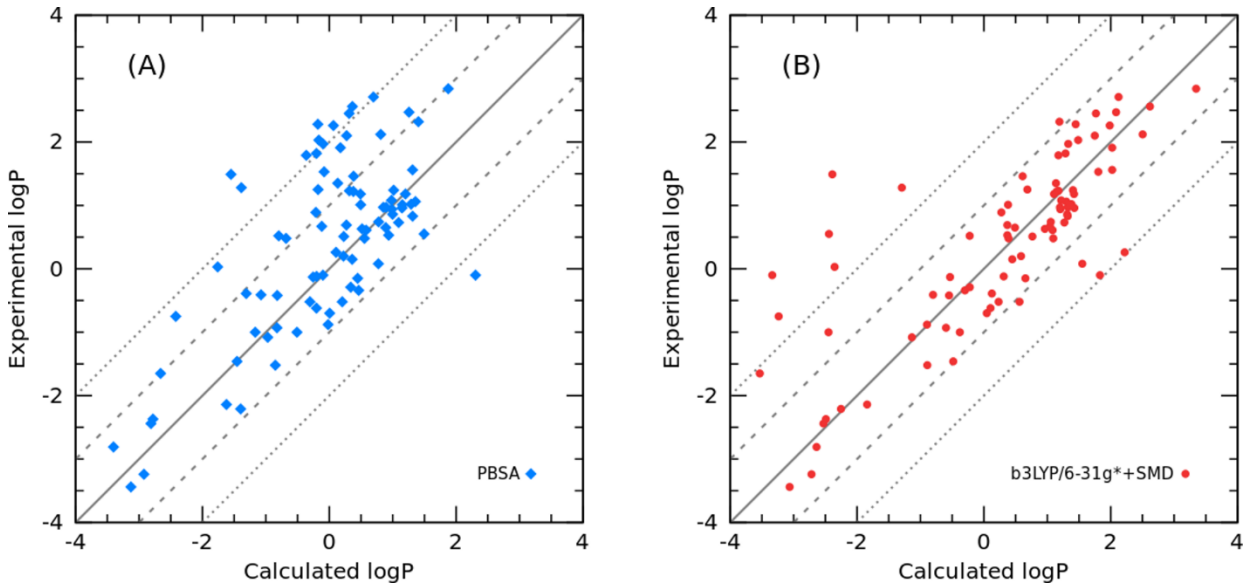

Figure 9 Correlation between the experimental and calculated $logP_{cyc/wat}$. Figure 9A. Calculated $logP_{cyc/wat}$ using the PBSA method; Figure 9B. Calculated $logP_{cyc/wat}$ using the SMD method.

3.3.5 GA optimized atomic radii and non-polar parameters for PB calculations

After 5 iterations, we stop the process and test the performance of PBSA with the tuned radii and non-polar parameters. The new radii were listed in Appendix xxx and the non-polar parameters were listed in Table 3.

Table 3 Non-polar parameters coupled with GA tuned radii

Solvent	ε	γ	b
Water	80.00	0.0053	1.03
Toluene	2.37	0.0238	3.90
Cyclohexane	2.23	0.0235	4.40

To test these parameters, we use $logP_{tol/wat}$ and $logD_{cyc/wat}$ from SAMPL9 and SAMPL5 again. The results are shown in Table 4. Although the perforamnce of the new radii set is similar

to the previous radii, it should be a more rubust set of parameters since we used more drug like molecules during the optimization and considered more molecular mechanics atom types.

Table 4 Test results of new PB parameters

Solvent System	Number of Ligands	Previous radii	Current radii
		RMSE	RMSE
SAMPL9 Tol/wat	16	1.33 kcal/mol	1.30 kcal/mol
SAMPL5 Cyc/wat	53	1.88 log unit	1.86 log unit

3.4 Conclusion

In this study, we extended the scope of our PBSA method for predicting solvation free energies in toluene and cyclohexane for organic molecules by parameterizing the nonpolar part and successfully applied this model to predict toluene-water partition coefficients in the SAMPL9 challenge. The PBSA method performed the best out of a total of 18 submissions in terms of RMSE. The RMSE error of our submission, 1.52 kcal/mol, was further reduced to 1.33 kcal/mol after using the multi-conformations generated through MD simulations. The distribution coefficient dataset from SAMPL5 challenge was adopted to test the performance of the PBSA SFE model for cyclohexane, and the prediction error of our model, RMSE = 1.88 log units, was better than that of COSMO-RS, which had the lowest prediction error (RMSE = 2.11 log units) among the 63 submissions of the SAMPL5 challenge. The ACES TI was conducted to calculate toluene-water transfer free energy in SAMPL9 dataset and cyclohexane-water logD in SAMPL5 dataset. The RMSE of TI were 2.11 kcal/mol on SAMPL9 dataset and 2.15 log units on SAMPL5 dataset. This further proved the reliability of our PBSA-based approach for partition coefficient prediction. In addition, we discussed the potential sources of errors for some poor predictions. More excitingly, we found the prediction error of our models can be further reduced when using multiple conformations. Among the three conformational ensemble generation methods, MD simulation achieved the best performance. We further evaluated our two PBSA SFE models using two larger molecule sets. Finally, we conducted global optimization of PB parameters. The intermediate version of parameters can even achieve similar accuracy compared with our previous results.

4.0 Future Work Perspectives

To develop a set of robust atomic radii for PB calculation, we intend to use a set of training set with more molecules, involving 1100 solvation free energy data derived from henry's law constant database.

Although adjusting the atomic radii in PB calculations can improve the accuracy of PBSA in predicting SFE and binding free energies. However, another assumption for the practical application of PBSA is that solutes and solvents all have homogeneous dielectric constants. This assumption ignores the fact that solutes, especially biomolecules (proteins, DNA and RNA), usually have highly charged regions, which leads to the inability of the uniform dielectric constant to accurately describe the dielectric properties of solutes. Therefore, using an automated process to differentiate dielectric regions of proteins and using different dielectric constants for PB calculations is expected to further improve the predictive performance of PBSA.

The current widely used molecular mechanics force field still employs atomic partial charges, and although this treatment ensures computational efficiency, the atomic partial charges cannot adequately consider the polarization effect. Therefore, the development of new electrostatic models and the incorporation of explicit polarization effects can depict the electrostatic interactions of molecules more accurately. Polarizable molecular mechanics force fields can also be combined with PBSA to produce more accurate electrostatic potentials.

Appendix A Atomic Radii Used for PBSA Organic Solvent Model

Appendix Table 1 Atomic radii used for PBSA organic solvent models adopted from Sun et al.

Atom type	Old Radius Parameter	Optimized Radius Parameter	Radius Parameters for SASA and WSAS Entropy Calculations	Weight of WSAS
Hydrogen				
h1	1.19	1.19	1.20	0.105257
h2	1.19	1.19	1.20	0.0866113
h3	1.19	1.19	1.20	0.0708034
h4	1.19	1.19	1.20	0.104611
h5	1.19	1.19	1.20	0.0951559
ha	1.19	1.19	1.20	0.114837
hc	1.19	1.19	1.20	0.127134
hn	1.19	1.19	1.20	0.0145069
hn1		1.50	1.20	0.0145069
hn2		1.60	1.20	0.0145069
hn3		1.70	1.20	0.0145069
ho	1.19	1.19	1.20	0.004208
hp	1.19	1.19	1.20	0.0166403
hs	1.19	1.19	1.20	0.0157608
hw	1.19	1.19	1.20	0.0106
hx	1.19	1.19	1.20	0.0574766
HC	1.19	1.19	1.20	0.127134
HA	1.19	1.19	1.20	0.114837
HO	1.19	1.19	1.20	0.004208
HS	1.19	1.19	1.20	0.0157608
HW	1.19	1.19	1.20	0.004208
HP	1.19	1.19	1.20	0.0166403
HZ	1.19	1.19		
H1	1.19	1.19	1.20	0.105257
H2	1.19	1.19	1.20	0.0866113
H3	1.19	1.19	1.20	0.0708034
H4	1.19	1.19	1.20	0.104611
H5	1.19	1.19	1.20	0.0951559
H	1.19	1.19	1.20	0.0145069
Carbon				
c	1.76	1.76	1.70	0.559732
c1	1.76	1.90	1.70	0.826582
c2	1.76	1.76	1.70	0.559732
c3	1.76	1.76	1.70	0.63088
ca	1.76	1.76	1.70	0.559732

cp	1.76	1.76	1.70	0.559732
cq	1.76	1.76	1.70	0.559732
cc	1.76	1.76	1.70	0.559732
cd	1.76	1.76	1.70	0.559732
ce	1.76	1.76	1.70	0.559732
cf	1.76	1.76	1.70	0.559732
cg	1.76	1.76	1.70	0.826582
ch	1.76	1.76	1.70	0.826582
cx	1.76	1.76	1.70	0.63088
cy	1.76	1.76	1.70	0.63088
cz	1.76	1.76	1.70	0.559732
c5	1.76	1.76	1.70	0.63088
c6	1.76	1.76	1.70	0.63088
cu	1.76	1.76	1.70	0.559732
cv	1.76	1.76	1.70	0.559732
CA	1.76	1.76	1.70	0.559732
CB	1.76	1.76	1.70	0.559732
CC	1.76	1.76	1.70	0.559732
CD	1.76	1.76	1.70	0.559732
CK	1.76	1.76	1.70	0.559732
CM	1.76	1.76	1.70	0.559732
CN	1.76	1.76	1.70	0.559732
CQ	1.76	1.76	1.70	0.559732
CR	1.76	1.76	1.70	0.559732
CT	1.76	1.76	1.70	0.63088
CV	1.76	1.76	1.70	0.559732
CW	1.76	1.76	1.70	0.559732
C*	1.76	1.76	1.70	0.559732
CY	1.76	1.76	1.70	0.826582
CZ	1.76	1.76	1.70	0.826582
C	1.76	1.76	1.70	0.826582
C3	1.76	1.76	1.70	0.63088
C4	1.76	1.76	1.70	0.63088
C5	1.76	1.76	1.70	0.559732
C6	1.76	1.76	1.70	0.559732
C8	1.76	1.76	1.70	0.63088
CX	1.76	1.76	1.70	0.63088
2C	1.76	1.76	1.70	0.63088
3C	1.76	1.76	1.70	0.63088
CO	1.76	1.76	1.70	0.559732
CI	1.76	1.76	1.70	0.63088
CP	1.76	1.76	1.70	0.559732
CS	1.76	1.76	1.70	0.559732
Nitrogen				
n	1.73	1.73	1.55	0.635011

n1	1.73	1.73	1.55	0.567605
n2	1.73	1.73	1.55	0.582155
n3	1.73	1.73	1.55	0.546228
n4	1.73	1.73	1.55	1.56076
n5	1.73	1.73	1.55	0.485127
n6	1.73	1.73	1.55	0.485127
n7	1.73	1.73	1.55	0.485127
n8	1.73	1.73	1.55	0.433329
n9	1.73	1.73	1.55	0.329614
na	1.73	1.73	1.55	0.72638
nb	1.73	1.73	1.55	0.582155
nc	1.73	1.73	1.55	0.582155
nd	1.73	1.73	1.55	0.582155
ne	1.73	1.73	1.55	0.582155
nf	1.73	1.73	1.55	0.582155
nh	1.73	1.73	1.55	0.734254
no	1.73	1.73	1.55	0.546228
ni	1.73	1.73	1.55	0.635011
nj	1.73	1.73	1.55	0.635011
nk	1.73	1.73	1.55	1.38946
nl	1.73	1.73	1.55	1.38946
nm	1.73	1.73	1.55	0.734254
nn	1.73	1.73	1.55	0.734254
np	1.73	1.73	1.55	0.546228
nq	1.73	1.73	1.55	0.546228
ns	1.73	1.73	1.55	0.584969
nt	1.73	1.73	1.55	0.540968
nu	1.73	1.73	1.55	0.676782
nv	1.73	1.73	1.55	0.625821
nx	1.73	1.73	1.55	1.38946
ny	1.73	1.73	1.55	1.24398
nz	1.73	1.73	1.55	1.11956
n+	1.73	1.73	1.55	1.01253
NA	1.73	1.73	1.55	0.72638
NB	1.73	1.73	1.55	0.582155
NC	1.73	1.73	1.55	0.582155
N2	1.73	1.73	1.55	0.72638
N3	1.73	1.73	1.55	0.546228
NT	1.73	1.73	1.55	0.546228
N*	1.73	1.73	1.55	0.72638
NY	1.73	1.73	1.55	0.567605
N	1.73	1.73	1.55	0.635011
Oxygen				
o	1.43	1.70	1.52	0.528811
on		2.00	1.52	0.528811

oi		1.28	1.52	0.528811
oh	1.43	1.70	1.52	0.507605
os	1.43	1.64	1.52	0.413186
ow	1.43	1.64	1.52	0.594825
op	1.43	1.64	1.52	0.413186
oq	1.43	1.64	1.52	0.413186
O2	1.43	1.64	1.52	0.528811
OH	1.43	1.64	1.52	0.507605
OS	1.43	1.64	1.52	0.413186
OW	1.43	1.64	1.52	0.507605
O	1.43	1.64	1.52	0.528811
Sulfur				
s	1.75	2.00	1.80	1.15379
s2	1.75	2.00	1.80	1.15379
s4	1.75	2.00	1.80	1.15379
s6	1.75	2.80	1.80	0.847601
sh	1.75	2.00	1.80	1.15379
ss	1.75	2.00	1.80	1.15379
sx	1.75	2.00	1.80	1.15379
sy	1.75	2.00	1.80	0.847601
sp	1.75	2.00	1.80	1.15379
sq	1.75	2.00	1.80	1.15379
SH	1.75	2.00	1.80	1.15379
S	1.75	2.00	1.80	1.15379
Phosphate				
p2	1.75	2.00	1.80	1.20046
p3	1.75	2.00	1.80	1.20046
p4	1.75	2.00	1.80	1.20046
p5	1.75	2.60	1.80	1.20046
pb	1.75	2.00	1.80	1.20046
pc	1.75	2.00	1.80	1.20046
pd	1.75	2.00	1.80	1.20046
pe	1.75	2.00	1.80	1.20046
pf	1.75	2.00	1.80	1.20046
px	1.75	2.00	1.80	1.20046
py	1.75	2.00	1.80	1.20046
p	1.75	2.00		
P	1.75	2.00	1.80	1.20046
Halide				
f	1.40	1.90	1.47	0.393452
F	1.40	1.90	1.47	0.393452
cl	1.54	2.10	1.75	1.05024
Cl	1.54	2.10	1.75	1.05024
CL	1.54	2.10		
br	1.99	2.15	1.85	1.46244

Br	1.99	2.15	1.85	1.46244
BR	1.99	2.15		
i	2.00	2.20	1.90	2.00408
I	2.00	2.20	1.90	2.00408
Boron				
B	1.50	1.50		
Metal				
Mn	2.00	2.00		
Mg	2.00	2.00		
Fe	2.00	2.00		
Lone pair				
lp	0.00	0.00		
LP	0.00	0.00		
Z5	1.76	1.76	1.70	

Appendix B Tuned atomic radii from GA

Appendix Table 2 Atomic radii for PB from GA search

Atom type	Old Radius Parameter	Optimized Radius Parameter	Radius Parameters for SASA and WSAS Entropy Calculations	Weight of WSAS
Hydrogen				
h1	1.19	1.00	1.20	0.105257
h2	1.19	1.00	1.20	0.0866113
h3	1.19	1.00	1.20	0.0708034
h4	1.19	1.00	1.20	0.104611
h5	1.19	1.00	1.20	0.0951559
ha	1.19	1.00	1.20	0.114837
hc	1.19	1.00	1.20	0.127134
hn	1.19	1.13	1.20	0.0145069
hn1		1.50	1.20	0.0145069
hn2		1.60	1.20	0.0145069
hn3		1.70	1.20	0.0145069
ho	1.19	1.41	1.20	0.004208
hp	1.19	1.00	1.20	0.0166403
hs	1.19	1.00	1.20	0.0157608
hw	1.19	1.00	1.20	0.0106
hx	1.19	1.00	1.20	0.0574766
HC	1.19	1.00	1.20	0.127134
HA	1.19	1.00	1.20	0.114837
HO	1.19	1.00	1.20	0.004208
HS	1.19	1.00	1.20	0.0157608
HW	1.19	1.00	1.20	0.004208
HP	1.19	1.00	1.20	0.0166403
HZ	1.19	1.00		
H1	1.19	1.00	1.20	0.105257
H2	1.19	1.00	1.20	0.0866113
H3	1.19	1.00	1.20	0.0708034
H4	1.19	1.00	1.20	0.104611
H5	1.19	1.00	1.20	0.0951559
H	1.19	1.00	1.20	0.0145069
Carbon				
c	1.76	1.90	1.70	0.559732
c1	1.76	1.90	1.70	0.826582
c2	1.76	2.15	1.70	0.559732
c3	1.76	1.90	1.70	0.63088
ca	1.76	1.90	1.70	0.559732

cp	1.76	1.90	1.70	0.559732
cq	1.76	1.90	1.70	0.559732
cc	1.76	1.90	1.70	0.559732
cd	1.76	1.90	1.70	0.559732
ce	1.76	2.15	1.70	0.559732
cf	1.76	1.90	1.70	0.559732
cg	1.76	1.90	1.70	0.826582
ch	1.76	1.90	1.70	0.826582
cx	1.76	1.07	1.70	0.63088
cy	1.76	1.90	1.70	0.63088
cz	1.76	1.90	1.70	0.559732
c5	1.76	1.90	1.70	0.63088
c6	1.76	1.90	1.70	0.63088
cu	1.76	1.90	1.70	0.559732
cv	1.76	1.90	1.70	0.559732
CA	1.76	1.90	1.70	0.559732
CB	1.76	1.90	1.70	0.559732
CC	1.76	1.90	1.70	0.559732
CD	1.76	1.90	1.70	0.559732
CK	1.76	1.90	1.70	0.559732
CM	1.76	1.90	1.70	0.559732
CN	1.76	1.90	1.70	0.559732
CQ	1.76	1.90	1.70	0.559732
CR	1.76	1.90	1.70	0.559732
CT	1.76	1.90	1.70	0.63088
CV	1.76	1.90	1.70	0.559732
CW	1.76	1.90	1.70	0.559732
C*	1.76	1.90	1.70	0.559732
CY	1.76	1.90	1.70	0.826582
CZ	1.76	1.90	1.70	0.826582
C	1.76	1.90	1.70	0.826582
C3	1.76	1.90	1.70	0.63088
C4	1.76	1.90	1.70	0.63088
C5	1.76	1.90	1.70	0.559732
C6	1.76	1.90	1.70	0.559732
C8	1.76	1.90	1.70	0.63088
CX	1.76	1.90	1.70	0.63088
2C	1.76	1.90	1.70	0.63088
3C	1.76	1.90	1.70	0.63088
CO	1.76	1.90	1.70	0.559732
CI	1.76	1.90	1.70	0.63088
CP	1.76	1.90	1.70	0.559732
CS	1.76	1.90	1.70	0.559732
Nitrogen				
n	1.73	1.15	1.55	0.635011

n1	1.73	1.74	1.55	0.567605
n2	1.73	1.74	1.55	0.582155
n3	1.73	1.50	1.55	0.546228
n4	1.73	1.74	1.55	1.56076
n5	1.73	1.74	1.55	0.485127
n6	1.73	1.59	1.55	0.485127
n7	1.73	1.59	1.55	0.485127
n8	1.73	1.74	1.55	0.433329
n9	1.73	1.74	1.55	0.329614
na	1.73	1.74	1.55	0.72638
nb	1.73	1.70	1.55	0.582155
nc	1.73	1.74	1.55	0.582155
nd	1.73	1.74	1.55	0.582155
ne	1.73	1.74	1.55	0.582155
nf	1.73	1.74	1.55	0.582155
nh	1.73	1.74	1.55	0.734254
no	1.73	2.85	1.55	0.546228
ni	1.73	1.74	1.55	0.635011
nj	1.73	1.74	1.55	0.635011
nk	1.73	1.74	1.55	1.38946
nl	1.73	1.74	1.55	1.38946
nm	1.73	1.74	1.55	0.734254
nn	1.73	1.74	1.55	0.734254
np	1.73	1.74	1.55	0.546228
nq	1.73	1.74	1.55	0.546228
ns	1.73	1.94	1.55	0.584969
nt	1.73	1.64	1.55	0.540968
nu	1.73	1.74	1.55	0.676782
nv	1.73	1.74	1.55	0.625821
nx	1.73	1.74	1.55	1.38946
ny	1.73	1.74	1.55	1.24398
nz	1.73	1.74	1.55	1.11956
n+	1.73	1.74	1.55	1.01253
NA	1.73	1.74	1.55	0.72638
NB	1.73	1.74	1.55	0.582155
NC	1.73	1.74	1.55	0.582155
N2	1.73	1.74	1.55	0.72638
N3	1.73	1.74	1.55	0.546228
NT	1.73	1.74	1.55	0.546228
N*	1.73	1.74	1.55	0.72638
NY	1.73	1.74	1.55	0.567605
N	1.73	1.74	1.55	0.635011
Oxygen				
o	1.43	1.62	1.52	0.528811
on		2.00	1.52	0.528811

oi		1.28	1.52	0.528811
oh	1.43	1.60	1.52	0.507605
os	1.43	1.81	1.52	0.413186
ow	1.43	1.62	1.52	0.594825
op	1.43	1.62	1.52	0.413186
oq	1.43	1.62	1.52	0.413186
O2	1.43	1.62	1.52	0.528811
OH	1.43	1.62	1.52	0.507605
OS	1.43	1.62	1.52	0.413186
OW	1.43	1.62	1.52	0.507605
O	1.43	1.62	1.52	0.528811
Sulfur				
s	1.75	2.40	1.80	1.15379
s2	1.75	2.40	1.80	1.15379
s4	1.75	2.82	1.80	1.15379
s6	1.75	1.58	1.80	0.847601
sh	1.75	2.40	1.80	1.15379
ss	1.75	2.25	1.80	1.15379
sx	1.75	2.40	1.80	1.15379
sy	1.75	2.40	1.80	0.847601
sp	1.75	2.40	1.80	1.15379
sq	1.75	2.40	1.80	1.15379
SH	1.75	2.40	1.80	1.15379
S	1.75	2.40	1.80	1.15379
Phosphate				
p2	1.75	1.72	1.80	1.20046
p3	1.75	1.72	1.80	1.20046
p4	1.75	1.72	1.80	1.20046
p5	1.75	1.72	1.80	1.20046
pb	1.75	1.72	1.80	1.20046
pc	1.75	1.72	1.80	1.20046
pd	1.75	1.72	1.80	1.20046
pe	1.75	1.72	1.80	1.20046
pf	1.75	1.72	1.80	1.20046
px	1.75	1.72	1.80	1.20046
py	1.75	1.72	1.80	1.20046
p	1.75	1.72		
P	1.75	1.72	1.80	1.20046
Halide				
f	1.40	2.91	1.47	0.393452
F	1.40	2.91	1.47	0.393452
cl	1.54	2.15	1.75	1.05024
Cl	1.54	2.15	1.75	1.05024
CL	1.54	2.15		
br	1.99	2.18	1.85	1.46244

Br	1.99	2.18	1.85	1.46244
BR	1.99	2.18		
i	2.00	1.92	1.90	2.00408
I	2.00	1.92	1.90	2.00408
Boron				
B	1.50	1.50		
Metal				
Mn	2.00	2.00		
Mg	2.00	2.00		
Fe	2.00	2.00		
Lone pair				
lp	0.00	0.00		
LP	0.00	0.00		
Z5	1.76	1.76	1.70	

Appendix C Calculated toluene-water logP using TI

Appendix Table 3 The experimental logP from SAMPL9 dataset and the calculated toluene-water logP using TI. Standard Deviation was calculated from three independent TI runs.

Solute ID	$logP_{expt}$	$logP_{calc}$	Standard Deviation
1	-5.11	-2.56	0.15
2	-3.26	-5.14	0.36
3	-7.49	-8.25	0.17
4	-7.44	-8.34	0.28
5	-4.91	-6.57	0.16
6	1.67	3.72	0.17
7	-5.94	-7.98	0.23
8	-3.79	-6.92	1.08
9	-6.87	-8.46	0.14
10	-3.36	-3.57	0.19
11	-1.99	-2.63	0.31
12	2.16	4.55	0.25
13	-0.49	-0.44	0.24
14	-1.92	-5.49	0.41
15	1.01	2.76	0.11
16	-5.13	-8.96	0.21
MSE		-0.71	
MUE		1.81	
RMSE		2.11	

Appendix D Calculated cyclohexane-water logP and logD using TI

Appendix Table 4 The experimental logD from SAMPL5 dataset and the calculated cyclohexane-water logP and logD using TI. Standard Deviation was calculated from three independent TI runs.

Solute ID	$logD_{expt}$	$logP_{calc}$	$logD_{calc}$	Standard Deviation
SAMPL5_002	1.40	0.74	0.74	0.21
SAMPL5_003	1.90	1.84	1.84	0.09
SAMPL5_004	2.20	2.60	2.50	0.15
SAMPL5_005	-0.86	0.75	0.75	0.16
SAMPL5_006	-1.02	-0.02	-0.02	0.14
SAMPL5_007	1.40	3.75	3.60	0.14
SAMPL5_010	-1.70	-2.82	-5.36	0.09
SAMPL5_011	-2.96	-0.72	-4.11	0.23
SAMPL5_013	-1.50	-0.83	-0.83	0.20
SAMPL5_015	-2.20	-3.27	-6.32	0.17
SAMPL5_017	2.50	4.96	4.96	0.22
SAMPL5_019	1.20	1.56	1.50	0.14
SAMPL5_020	1.60	1.89	1.89	0.21
SAMPL5_021	1.20	3.25	3.25	0.18
SAMPL5_024	1.00	3.83	3.83	0.19
SAMPL5_026	-2.60	-0.90	-3.57	0.14
SAMPL5_027	-1.87	-2.84	-2.84	0.14
SAMPL5_033	1.80	3.42	3.42	0.33
SAMPL5_037	-1.50	-3.54	-4.38	0.19
SAMPL5_042	-1.10	-0.99	-0.99	0.12
SAMPL5_044	1.00	0.25	0.25	0.44
SAMPL5_045	-2.10	-2.81	-2.81	0.10
SAMPL5_046	0.20	0.67	0.67	0.32
SAMPL5_047	-0.40	1.15	1.15	0.19
SAMPL5_048	0.90	0.57	0.57	0.18
SAMPL5_049	1.30	-0.91	-0.91	0.20
SAMPL5_050	-3.20	-1.32	-1.70	0.17
SAMPL5_055	-1.50	-2.79	-2.79	0.18
SAMPL5_056	-2.50	-4.39	-4.47	0.14
SAMPL5_058	0.80	1.70	1.70	0.12
SAMPL5_059	-1.30	-1.32	-1.32	0.18
SAMPL5_060	-3.90	-2.16	-4.61	0.28

SAMPL5_061	-1.45	-1.49	-1.64	0.41
SAMPL5_063	-3.00	-5.88	-7.54	0.15
SAMPL5_065	0.70	0.89	-0.18	0.40
SAMPL5_067	-1.30	2.22	0.75	0.22
SAMPL5_068	1.40	3.41	3.41	0.18
SAMPL5_069	-1.30	0.36	0.35	0.46
SAMPL5_070	1.60	5.25	3.33	0.32
SAMPL5_071	-0.10	-0.01	-0.01	0.24
SAMPL5_072	0.60	3.59	2.34	0.21
SAMPL5_074	-1.90	-9.53	-9.53	0.08
SAMPL5_075	-2.80	1.53	0.40	0.17
SAMPL5_080	-2.20	-4.55	-4.55	0.05
SAMPL5_081	-2.20	-4.68	-5.62	0.17
SAMPL5_082	2.50	7.66	6.87	0.20
SAMPL5_083	-1.90	-0.29	-0.29	1.58
SAMPL5_084	0.00	2.11	1.26	0.29
SAMPL5_085	-2.20	-2.42	-2.42	0.20
SAMPL5_086	0.70	0.70	-1.42	0.58
SAMPL5_088	-1.90	-3.75	-3.75	0.16
SAMPL5_090	0.80	1.52	1.52	0.14
SAMPL5_092	-0.40	-0.21	-0.21	0.21
MSE			-0.10	
MUE			1.62	
RMSE			2.15	

Bibliography

1.Hansch, C.; Fujita, T., p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *Journal of the American Chemical Society* **1964,** *86* (8), 1616-1626.
2.Cramer, R. D.; Patterson, D. E.; Bunce, J. D., Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* **1988,** *110* (18), 5959-5967.
3.Meng, E. C.; Shoichet, B. K.; Kuntz, I. D., Automated docking with grid-based energy evaluation. *Journal of Computational Chemistry* **1992,** *13* (4), 505-524.
4.Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P., A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society* **1984,** *106* (3), 765-784.
5.Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A., An all atom force field for simulations of proteins and nucleic acids. *Journal of Computational Chemistry* **1986,** *7* (2), 230-252.
6.Tanaka, S.; Scheraga, H. A., Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. *Macromolecules* **1976,** *9* (6), 945-950.
7.Muegge, I.; Martin, Y. C., A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach. *Journal of Medicinal Chemistry* **1999,** *42* (5), 791-804.
8.Nina, M.; Beglov, D.; Roux, B., Atomic Radii for Continuum Electrostatics Calculations Based on Molecular Dynamics Free Energy Simulations. *The Journal of Physical Chemistry B* **1997,** *101* (26), 5239-5248.
9.Cortis, C. M.; Friesner, R. A., Numerical solution of the Poisson–Boltzmann equation using tetrahedral finite-element meshes. *Journal of Computational Chemistry* **1997,** *18* (13), 1591-1608.
10.Holst, M.; Baker, N.; Wang, F., Adaptive multilevel finite element solution of the Poisson–Boltzmann equation I. Algorithms and examples. *Journal of Computational Chemistry* **2000,** *21* (15), 1319-1342.
11.Shestakov, A. I.; Milovich, J. L.; Noy, A., Solution of the Nonlinear Poisson–Boltzmann Equation Using Pseudo-transient Continuation and the Finite Element Method. *Journal of Colloid and Interface Science* **2002,** *247* (1), 62-79.
12.Xie, D.; Zhou, S., A new minimization protocol for solving nonlinear Poisson–Boltzmann mortar finite element equation. *BIT Numerical Mathematics* **2007,** *47* (4), 853-871.
13.Miertuš, S.; Scrocco, E.; Tomasi, J., Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chemical Physics* **1981,** *55* (1), 117-129.
14.Hoshi, H.; Sakurai, M.; Inoue, Y.; Chûjô, R., Medium effects on the molecular electronic structure. I. The formulation of a theory for the estimation of a molecular electronic structure surrounded by an anisotropic medium. *The Journal of chemical physics* **1987,** *87* (2), 1107-1115.
15.Zauhar, R. J.; Morgan, R. S., The rigorous computation of the molecular electric potential. *Journal of Computational Chemistry* **1988,** *9* (2), 171-187.
16.Rashin, A. A., Hydration phenomena, classical electrostatics, and the boundary element method. *Journal of physical chemistry* **1990,** *94* (5), 1725-1733.

17.Yoon, B. J.; Lenhoff, A. M., A boundary element method for molecular electrostatics with electrolyte effects. *Journal of Computational Chemistry* **1990,** *11* (9), 1080-1086.
18.Zhou, H.-X., Boundary element solution of macromolecular electrostatics: interaction energy between two proteins. *Biophysical journal* **1993,** *65* (2), 955-963.
19.Bharadwaj, R.; Windemuth, A.; Sridharan, S.; Honig, B.; Nicholls, A., The fast multipole boundary element method for molecular electrostatics: An optimal approach for large systems. *Journal of Computational Chemistry* **1995,** *16* (7), 898-913.
20.Purisima, E. O.; Nilar, S. H., A simple yet accurate boundary element method for continuum dielectric calculations. *Journal of Computational Chemistry* **1995,** *16* (6), 681-689.
21.Liang, J.; Subramaniam, S., Computation of molecular electrostatics with boundary element methods. *Biophysical journal* **1997,** *73* (4), 1830-1841.
22.Vorobjev, Y. N.; Scheraga, H. A., A fast adaptive multigrid boundary element method for macromolecular electrostatic computations in a solvent. *Journal of Computational Chemistry* **1997,** *18* (4), 569-583.
23.Klapper, I.; Hagstrom, R.; Fine, R.; Sharp, K.; Honig, B., Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: Effects of ionic strength and amino-acid modification. *Proteins: Structure, Function, and Bioinformatics* **1986,** *1* (1), 47-59.
24.Sharp, K. A.; Honig, B., Electrostatic interactions in macromolecules: theory and applications. *Annual review of biophysics and biophysical chemistry* **1990,** *19* (1), 301-332.
25.Madura, J. D.; Davist, M. E.; Gilson, M. K.; Wades, R. C.; Luty, B. A.; McCammon, J. A., Biological Applications of Electrostatic Calculations and Brownian Dynamics Simulations. In *Reviews in Computational Chemistry*, 1994; pp 229-267.
26.Gilson, M. K.; Sharp, K. A.; Honig, B. H., Calculating the electrostatic potential of molecules in solution: Method and error assessment. *Journal of Computational Chemistry* **1988,** *9* (4), 327-335.
27.Gilson, M. K.; Honig, B., Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. *Proteins: Structure, Function, and Bioinformatics* **1988,** *4* (1), 7-18.
28.Nicholls, A.; Honig, B., A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. *Journal of Computational Chemistry* **1991,** *12* (4), 435-445.
29.Rocchia, W.; Alexov, E.; Honig, B., Extending the Applicability of the Nonlinear Poisson−Boltzmann Equation: Multiple Dielectric Constants and Multivalent Ions. *The Journal of Physical Chemistry B* **2001,** *105* (28), 6507-6514.
30.Burden, R. L.; Faires, J. D., *Numerical analysis*. Brooks Cole: 1997.
31.Press, W. H., *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press: 2007.
32.Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B., 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* **2002,** *23* (1), 128-137.
33.Wang, J.; Hou, T.; Xu, X., Recent Advances in Free Energy Calculations with a Combination of Molecular Mechanics and Continuum Models. *Current Computer-Aided Drug Design* **2006,** *2* (3), 287-306.
34.Taylor, R. D.; Jewsbury, P. J.; Essex, J. W., A review of protein-small molecule docking methods. *Journal of Computer-Aided Molecular Design* **2002,** *16* (3), 151-166.

35.Wu, K.-W.; Chen, P.-C.; Wang, J.; Sun, Y.-C., Computation of relative binding free energy for an inhibitor and its analogs binding with Erk kinase using thermodynamic integration MD simulation. *Journal of Computer-Aided Molecular Design* **2012,** *26* (10), 1159-1169.
36.Lawrenz, M.; Baron, R.; Wang, Y.; McCammon, J. A., Independent-Trajectory Thermodynamic Integration: A Practical Guide to Protein-Drug Binding Free Energy Calculations Using Distributed Computing. In *Computational Drug Discovery and Design*, Baron, R., Ed. Springer New York: New York, NY, 2012; pp 469-486.
37.Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R., Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *Journal of the American Chemical Society* **2015,** *137* (7), 2695-2703.
38.Wang, L.; Deng, Y.; Knight, J. L.; Wu, Y.; Kim, B.; Sherman, W.; Shelley, J. C.; Lin, T.; Abel, R., Modeling Local Structural Rearrangements Using FEP/REST: Application to Relative Binding Affinity Predictions of CDK2 Inhibitors. *Journal of Chemical Theory and Computation* **2013,** *9* (2), 1282-1293.
39.Hou, T.; Wang, J.; Li, Y.; Wang, W., Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* **2011,** *51* (1), 69-82.
40.Sun, Y.; He, X.; Hou, T.; Cai, L.; Man, V. H.; Wang, J., Development and test of highly accurate endpoint free energy methods. 1: Evaluation of ABCG2 charge model on solvation free energy prediction and optimization of atom radii suitable for more accurate solvation free energy prediction by the PBSA method. *Journal of Computational Chemistry* **2023,** *44* (14), 1334-1346.
41.Gautam, V.; Chong, W. L.; Wisitponchai, T.; Nimmanpipug, P.; Zain, S. M.; Rahman, N. A.; Tayapiwatana, C.; Lee, V. S., GPU-enabled molecular dynamics simulations of ankyrin kinase complex. *AIP Conference Proceedings* **2014,** *1621* (1), 112-115.
42.Gohlke, H.; Case, D. A., Converging free energy estimates: MM-PB(GB)SA studies on the protein–protein complex Ras–Raf. *Journal of Computational Chemistry* **2004,** *25* (2), 238-250.
43.Greene, D. A.; Botello-Smith, W. M.; Follmer, A.; Xiao, L.; Lambros, E.; Luo, R., Modeling Membrane Protein–Ligand Binding Interactions: The Human Purinergic Platelet Receptor. *The Journal of Physical Chemistry B* **2016,** *120* (48), 12293-12304.
44.Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A., Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices. *Journal of the American Chemical Society* **1998,** *120* (37), 9401-9409.
45.Islam, B.; Stadlbauer, P.; Neidle, S.; Haider, S.; Sponer, J., Can We Execute Reliable MM-PBSA Free Energy Computations of Relative Stabilities of Different Guanine Quadruplex Folds? *The Journal of Physical Chemistry B* **2016,** *120* (11), 2899-2912.
46.Izadyar, M.; Khavani, M.; Housaindokht, M. R., A combined molecular dynamic and quantum mechanic study of the solvent and guest molecule effect on the stability and length of heterocyclic peptide nanotubes. *Physical Chemistry Chemical Physics* **2015,** *17* (17), 11382-11391.
47.Henriksen, N. M.; Hayatshahi, H. S.; Davis, D. R.; Cheatham, T. E., III, Structural and Energetic Analysis of 2-Aminobenzimidazole Inhibitors in Complex with the Hepatitis C Virus IRES RNA Using Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* **2014,** *54* (6), 1758-1772.

48.Shin, W.; Yang, Z. J., Computational Strategies for Entropy Modeling in Chemical Processes. *Chemistry – An Asian Journal* **2023,** *18* (9), e202300117.
49.Wang, C.; Greene, D. A.; Xiao, L.; Qi, R.; Luo, R., Recent Developments and Applications of the MMPBSA Method. *Frontiers in Molecular Biosciences* **2018,** *4*.
50.Wang, J.; Hou, T., Develop and Test a Solvent Accessible Surface Area-Based Model in Conformational Entropy Calculations. *Journal of Chemical Information and Modeling* **2012,** *52* (5), 1199-1212.
51.Duan, L.; Liu, X.; Zhang, J. Z. H., Interaction Entropy: A New Paradigm for Highly Efficient and Reliable Computation of Protein–Ligand Binding Free Energy. *Journal of the American Chemical Society* **2016,** *138* (17), 5722-5728.
52.Duan, L.; Feng, G.; Wang, X.; Wang, L.; Zhang, Q., Effect of electrostatic polarization and bridging water on CDK2–ligand binding affinities calculated using a highly efficient interaction entropy method. *Physical Chemistry Chemical Physics* **2017,** *19* (15), 10140-10152.
53.Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J. Z. H.; Hou, T., End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chemical Reviews* **2019,** *119* (16), 9478-9508.
54.Wang, E.; Weng, G.; Sun, H.; Du, H.; Zhu, F.; Chen, F.; Wang, Z.; Hou, T., Assessing the performance of the MM/PBSA and MM/GBSA methods. 10. Impacts of enhanced sampling and variable dielectric model on protein–protein Interactions. *Physical Chemistry Chemical Physics* **2019,** *21* (35), 18958-18969.
55.Sun, H.; Li, Y.; Shen, M.; Tian, S.; Xu, L.; Pan, P.; Guan, Y.; Hou, T., Assessing the performance of MM/PBSA and MM/GBSA methods. 5. Improved docking performance using high solute dielectric constant MM/GBSA and MM/PBSA rescoring. *Physical Chemistry Chemical Physics* **2014,** *16* (40), 22035-22045.
56.Xu, L.; Sun, H.; Li, Y.; Wang, J.; Hou, T., Assessing the Performance of MM/PBSA and MM/GBSA Methods. 3. The Impact of Force Fields and Ligand Charge Models. *The Journal of Physical Chemistry B* **2013,** *117* (28), 8408-8421.
57.Su, P.-C.; Tsai, C.-C.; Mehboob, S.; Hevener, K. E.; Johnson, M. E., 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 (FabI). *Journal of Computational Chemistry* **2015,** *36* (25), 1859-1873.
58.Sun, Y.; Hou, T.; He, X.; Man, V. H.; Wang, J., Development and test of highly accurate endpoint free energy methods. 2: Prediction of logarithm of n-octanol–water partition coefficient (logP) for druglike molecules using MM-PBSA method. *Journal of Computational Chemistry* **2023,** *44* (13), 1300-1311.
59.He, X.; Man, V. H.; Yang, W.; Lee, T.-S.; Wang, J., A fast and high-quality charge model for the next generation general AMBER force field. *The Journal of Chemical Physics* **2020,** *153* (11).
60.Zamora, W. J.; Viayna, A.; Pinheiro, S.; Curutchet, C.; Bisbal, L.; Ruiz, R.; Ràfols, C.; Luque, F. J., Prediction of toluene/water partition coefficients in the SAMPL9 blind challenge: assessment of machine learning and IEF-PCM/MST continuum solvation models. *Physical Chemistry Chemical Physics* **2023,** *25* (27), 17952-17965.
61.Bannan, C. C.; Burley, K. H.; Chiu, M.; Shirts, M. R.; Gilson, M. K.; Mobley, D. L., Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. *Journal of Computer-Aided Molecular Design* **2016,** *30* (11), 927-944.

62.Bergazin, T. D.; Tielker, N.; Zhang, Y.; Mao, J.; Gunner, M. R.; Francisco, K.; Ballatore, C.; Kast, S. M.; Mobley, D. L., Evaluation of log P, pKa, and log D predictions from the SAMPL7 blind challenge. *Journal of Computer-Aided Molecular Design* **2021,** *35* (7), 771-802.
63.Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *The Journal of Physical Chemistry B* **2009,** *113* (18), 6378-6396.
64.Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G., *Minnesota Solvation Database – version 2012*. University of Minnesota: Minneapolis, **2012**.
65.Mobley, D. L.; Guthrie, J. P., FreeSolv: a database of experimental and calculated hydration free energies, with input files. *Journal of Computer-Aided Molecular Design* **2014,** *28* (7), 711-720.
66.Maestro *Schrödinger*, LLC: New York, NY, **2017**.
67.O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011,** *3* (1), 33.
68.Leo, A.; Hansch, C.; Elkins, D., Partition coefficients and their uses. *Chemical Reviews* **1971,** *71* (6), 525-616.
69.Shalaeva, M.; Caron, G.; Abramov, Y. A.; O'Connell, T. N.; Plummer, M. S.; Yalamanchi, G.; Farley, K. A.; Goetz, G. H.; Philippe, L.; Shapiro, M. J., Integrating Intramolecular Hydrogen Bonding (IMHB) Considerations in Drug Discovery Using ΔlogP As a Tool. *Journal of Medicinal Chemistry* **2013,** *56* (12), 4870-4879.
70.Byrne, F. P.; Hodds, W. M.; Shimizu, S.; Farmer, T. J.; Hunt, A. J., A comparison of the solvation power of the green solvent 2,2,5,5-tetramethyloxolane versus toluene via partition coefficients. *Journal of Cleaner Production* **2019,** *240*, 118175.
71.Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A., Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling* **2006,** *25* (2), 247-260.
72.Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J Comput Chem* **2004,** *25* (9), 1157-1174.
73.Case, D. A.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; Cheatham, I., T.E. ; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Ghoreishi, D.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Greene, D.; Harris, R.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Kurtzman, T.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D. J.; Merz, K. M.; Miao, Y.; Monard, G.; Nguyen, C.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C. L.; Smith, J.; Salomon-Ferrer, R.; Swails, J.; Walker, R. C.; Wang, J.; Wei, H.; Wolf, R. M.; Wu, X.; Xiao, L.; D.M., Y.; Kollman, P. A., AMBER 2018. *University of California, San Francisco* **2018**.
74.Li, L.; Li, C.; Sarkar, S.; Zhang, J.; Witham, S.; Zhang, Z.; Wang, L.; Smith, N.; Petukh, M.; Alexov, E., DelPhi: a comprehensive suite for DelPhi software and associated resources. *BMC Biophysics* **2012,** *5* (1), 9.
75.Bondi, A., van der Waals Volumes and Radii. *The Journal of Physical Chemistry* **1964,** *68* (3), 441-451.
76.Lee, T.-S.; Tsai, H.-C.; Ganguly, A.; York, D. M., ACES: Optimized Alchemically Enhanced Sampling. *Journal of Chemical Theory and Computation* **2023,** *19* (2), 472-487.

77.Kaus, J. W.; Pierce, L. T.; Walker, R. C.; McCammon, J. A., Improving the Efficiency of Free Energy Calculations in the Amber Molecular Dynamics Package. *Journal of Chemical Theory and Computation* **2013,** *9* (9), 4131-4139.
78.Lee, T.-S.; Hu, Y.; Sherborne, B.; Guo, Z.; York, D. M., Toward Fast and Accurate Binding Affinity Prediction with pmemdGTI: An Efficient Implementation of GPU-Accelerated Thermodynamic Integration. *Journal of Chemical Theory and Computation* **2017,** *13* (7), 3077-3084.
79.Lee, T.-S.; Cerutti, D. S.; Mermelstein, D.; Lin, C.; LeGrand, S.; Giese, T. J.; Roitberg, A.; Case, D. A.; Walker, R. C.; York, D. M., GPU-Accelerated Molecular Dynamics and Free Energy Methods in Amber18: Performance Enhancements and New Features. *Journal of Chemical Information and Modeling* **2018,** *58* (10), 2043-2050.
80.Steinbrecher, T.; Mobley, D. L.; Case, D. A., Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *The Journal of Chemical Physics* **2007,** *127* (21), 214108.
81.Steinbrecher, T.; Joung, I.; Case, D. A., Soft-core potentials in thermodynamic integration: Comparing one- and two-step transformations. *Journal of Computational Chemistry* **2011,** *32* (15), 3253-3263.
82.Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, **2016**.
83.Swanson, J. M. J.; Adcock, S. A.; McCammon, J. A., Optimized Radii for Poisson−Boltzmann Calculations with the AMBER Force Field. *Journal of Chemical Theory and Computation* **2005,** *1* (3), 484-493.
84.Friedrich, N. O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J., Benchmarking Commercial Conformer Ensemble Generators. *J Chem Inf Model* **2017,** *57* (11), 2719-2728.
85.Sun, Y.; Hou, T.; He, X.; Man, V. H.; Wang, J., Development and test of highly accurate endpoint free energy methods. 2: Prediction of logarithm of n-octanol-water partition coefficient (logP) for druglike molecules using MM-PBSA method. *J Comput Chem* **2023,** *44* (13), 1300-1311.
86.Klamt, A.; Eckert, F.; Reinisch, J.; Wichmann, K., Prediction of cyclohexane-water distribution coefficients with COSMO-RS on the SAMPL5 data set. *Journal of Computer-Aided Molecular Design* **2016,** *30* (11), 959-967.
87.Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. E., First Principles Calculations of Aqueous pKa Values for Organic and Inorganic Acids Using COSMO−RS Reveal an Inconsistency in the Slope of the pKa Scale. *The Journal of Physical Chemistry A* **2003,** *107* (44), 9380-9386.

88.Klicic, J. J.; Friesner, R. A.; Liu, S. Y.; Guida, W. C., Accurate prediction of acidity constants in aqueous solution via density functional theory and self-consistent reaction field methods. *Journal of Physical Chemistry A* **2002,** *106* (7), 1327-1335.