Published in final edited form as:

J Comput Chem. 2023 May 15; 44(13): 1300–1311. doi:10.1002/jcc.27086.

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

Yuchen Sun[1], **Tingjun Hou**[2], **Xibing He**[1], **Viet Hoang Man**[1], **Junmei Wang**[1,*]

[1]Department of Pharmaceutical Sciences and Computational Chemical Genomics Screening Center, School of Pharmacy, University of Pittsburgh, Pittsburgh, PA 15261, USA.

[2]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China

Abstract

The logarithm of n-octanol-water partition coefficient (logP) is frequently used as an indicator of lipophilicity in drug discovery, which has substantial impacts on the absorption, distribution, metabolism, excretion, and toxicity of a drug candidate. Considering that the experimental measurement of the property is costly and time-consuming, it is of great importance to develop reliable prediction models for logP. In this study, we developed a transfer free energy-based logP prediction model-FElogP. FElogP is based on the simple principle that logP is determined by the free energy change of transferring a molecule from water to n-octanol. The underline physical method to calculate transfer free energy is the molecular mechanics Poisson Boltzmann surface area (MM-PBSA), thus this method is named as Free Energy-based logP (FElogP). The superiority of FElogP model was validated by a large set of 707 structurally diverse molecules in the ZINC database for which the measurement was of high quality. Encouragingly, FElogP outperformed several commonly-used QSPR or machine learning-based logP models, as well as some continuum solvation model-based methods. The root mean square error (RMSE) and Pearson correlation (R) between the predicted and measured values are 0.91 log units and 0.71, respectively, while the runner-up, the logP model implemented in OpenBabel had an RMSE of 1.13 log units and R of 0.67. Given the fact that FElogP was not parameterized against experimental logP directly, its excellent performance is likely to be expanded to arbitrary organic molecules covered by the general AMBER force fields.

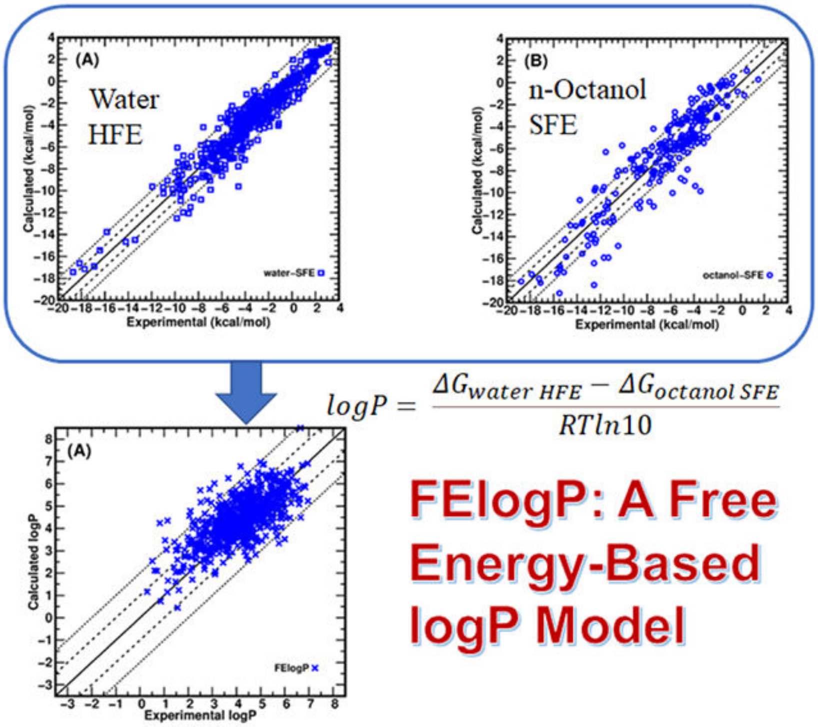
[*]Corresponding author, junmei.wang@pitt.edu.

Supporting information includes the following tables: the list of experimental hydration free energies and the predicted values using FElogP (Table S1), the list of solvation free energies in n-octanol and the predicted values using ab initio SMD and PBSA model developed in this work for 243 molecules (Table S2); the list of experimental logP and the predicted values using FElogP and ab initio SMD approaches for the 156 training set molecules (Table S3); List of experimental logP and the predicted values using seven approaches for the Martel dataset (Table S4). List of FElogP calculations for 21 outliers using multiple conformations with Method4 and Method5 (Table S5); the radius parameters for PBSA calculations (Table S6), and the molecular names and the corresponding experimental values for the three molecular sets, i.e., the 544-molecule set for hydration free energy calculation, the 243-molecule set for solvation free energy in n-octanol calculations, the 164-molecule set for evaluation of FElogP model (Table S7).

Conflict of Interest

There are no conflicts to declare.

Graphical Abstract



Keywords

logP; MM-PBSA; MM-GBSA; Transfer Free Energy; GAFF2; ABCG2

1. Introduction.

Partition coefficient is defined as the ratio of the distribution of a molecule in two immiscible solvents. The logarithm of partition coefficient (logP) between n-octanol and water is a critical property related to the physicochemical and physiological properties of a drug molecule, such as absorption, distribution, metabolism, excretion and toxicity (ADMET) and target protein binding.[1–4] Thus it is frequently used in describing the druggability of a pharmaceutical molecule. The first logP calculation method was developed by Fujita et al. in 1960s[5,6] Early on, the structural information of receptor is not easy to obtain, medicinal chemists frequently used quantitative structure-activity relationship (QSAR) to correlate the binding energy of a ligand with its physicochemical characteristics, and logP was regarded as a key descriptor measuring a drug's hydrophobicity. Nowadays, Lipinski's rule of five,[7,8] published in the 1990s, has been widely used as simple rules to determine the druglikeness of a pharmaceutical molecule. logP smaller than 5 is one of the four rules in Lipinski's rule of 5.

Several methods have been developed to experimentally measure logP,[5,9,10] such as the shake flask method and reversed phase high performance liquid chromatography.[11,12] However, the experimental determination of logP for some molecules is not easy, especially when the tested molecules are unstable in solvents. In addition, some compounds that need to be tested maybe difficult to synthesize or not purchasable. Such scenario is frequently encountered at the early stages of drug discovery when the physicochemical properties for a large number of structures need to be determined[13]. Hence, there is an increasing demand on the development of high-quality computational methods to predict logP accurately and efficiently.

Currently, a variety of free logP calculation tools are available in public domain. Furthermore, popular commercial molecular simulation platforms such as Discovery Studio, Molecular Operating Environment (MOE), and Schrödinger also support logP calculation, with either open-source models (Discovery Studio supports AlogP[14]) or their own unpublished models (MOE uses both unpublished model "Labute"[15] and hybrid SlogP[16]).

Existing calculation methods for logP can be roughly classified into four families: (1) atom-based methods, (2) fragment-based methods, (3) topology or graph-based method, and (4) structural property-based methods. Atom-based or atom-additive methods for logP prediction, such as AlogP[14], simply sum up the additive contributions of all atoms. Atom-based methods are suitable for small molecules, but they may fail for a very complex structure or a specific molecule for which the logP prediction is greatly affected by its electronic structure. To overcome these shortcomings, enhanced atom-based or hybrid methods such as XlogP[17] and SlogP[16] have been proposed by applying additional corrections (such as the influence from neighboring atoms) to achieve better prediction accuracy for large molecular systems. Fragment-based methods (e.g., Clogp[18] and KlogP[19]) for logP calculations are also based on the additive hypothesis, assuming logP can be calculated by the summation of the hydrophobic contributions of each constitutive fragment in a molecule. The fragment constant, i.e., the hydrophobic contribution of a fragment is determined by experimental logP of a compound without "surprise interactions"[20]. Other interactions such as hydrogen bonding, proximity interactions, hydrophobic shielding effect, bond and branching influence etc. are taken into account as the additional correction factors.[21] This class of models overall have better prediction performance than the atom-based methods for large molecules. The third family is topology or graph-based models, such as MlogP,[22] which apply topological descriptors to construct models. Topological descriptors are usually generated using 2D-structures. Recently, Ulrich et al. applied deep neural networks (DNN) to train molecular graphs for logP prediction.[23] The DNN model achieved an RMSE of 0.47 log units in the test dataset. The last family of methods is structural property-based methods which calculate logP from a more rigorous physical-chemical perspective.[24] This type of methods typically require 3D structures for performing quantum mechanics (QM) or molecular mechanics (MM) calculations. Molecular simulations are sometimes performed to obtain reliable prediction of some properties, such as free energy. For example, Procacci et al. conducted logP prediction using a non-equilibrium alchemical technique called non-equilibrium switching (NES) for the druglike molecules in the SAMPL6 challenge and was able to achieve mean unsigned error of 1.06 and correlation coefficient R of 0.79 for training set molecules which is among the top ranked submissions[25]. Ogata et al. conducted alchemical free energies to calculate logP using 27 different charge models for 58 compounds, and the best correlation coefficient R of 0.92 was achieved.[26]

The logP models from different families have their own advantages and disadvantages. Although simple methods such as AlogP[14] from the first family and ClogP[18] from the second family have been widely used, they are not very accurate especially for large and flexible molecules. The past decades have seen a trend of increase in molecular weight for approved small-molecule drugs[27]. According to the review by Shultz[28], in general, ClogP overestimates logP for molecules that have been approved by FDA after the publication of the famous “Lipinski rule of five”. One explanation is that large and flexible molecules have their polar atoms buried leading to the hydrophobic groups collapsed, an effect not being taken into account in ClogP model development.[29]

On the other hand, for structural property-based logP model, accurate prediction relies on theoretically rigorous interpretation of solvation process. For example, high level QM calculations with implicit solvation models can achieve relatively good accuracy, and those methods are the best ranking physical methods in the SAMPL6 challenge[30]. Nevertheless, such methods require high computational cost, making them less feasible and less attractive in practical applications.

Recently, Martel et al.[31] determined the logP values of 707 molecules from ZINC database[32] using ultra-high performance liquid chromatography (UHPLC) followed by ultraviolet (UV) or mass spectrometry (MS) detection. The molecules were selected out of 4.5 million compounds to guarantee the structural diversity in chemical space. Moreover, the experimental data was determined by one research lab to minimize the error stemming from different experiment protocols. The prediction performance of a set of widely used models for this dataset is much poorer compared to the reported performance for individual models. For example, ACD/GALAS has an RMDE of 1.44 log units, the DNN model by Ulrich et. al. has an RMSE of 1.23 log units.[23] Thus, the performance of logP models from the first three families is strongly influenced by the training set molecules.

In this work, we set forth to developing a high-quality logP model which can overcome the limitations of the models belonging to the first three families, i.e., their performance is training-set-dependent, and the models belong to the fourth families, i.e., the computational efficiency is very poor. To achieve this goal, we will first develop a high-quality Poisson Boltzmann surface area (MM-PBSA) model to accurately predict solvation free energy of an arbitrary molecule in n-octanol solvent. We then predict logP of a molecule using its transfer free energy from water phase to n-octanol phase. This is the second publication of the paper series “development and test of highly accurate endpoint free energy methods”. To the best of our knowledge, this is the first attempt to apply MM-PBSA method to calculate logP.

2. Methods

2.1 logP calculation using transfer free energy

In this study, a transfer free energy-based logP prediction method FElogP was developed and evaluated. From the thermodynamic point of view, logP is proportional to the Gibbs free energy of transferring a molecule from water to octanol according to the equation shown below:[33]

$$-RTln10 * logP = \quad \Delta G_{transfer} \tag{1}$$

where R represents the gas constant (8.314 J·mol^{-1}·K^{-1}), and T is the thermodynamic temperature (K). Consequently, logP can be calculated by the solvation free energies (SFE) of molecule in water and n-octanol phases:

$$logP = \quad \frac{\Delta G_{water\ HFE} - \Delta G_{octanol\ SFE}}{RTln10} \tag{2}$$

There are various computational methods to estimate the SFE of molecules in different solvents, ranging from rigorous alchemical free energy methods to simple quantitative structure-property relationship (QSPR) methods. The molecular mechanics Poisson Boltzmann surface area (MM-PBSA) and molecular mechanics Generalized Born surface area (MM-GBSA) methods are becoming increasingly popular among the existing free energy calculation methods due to their good balance between accuracy and efficiency. In MM-PBSA/GBSA calculations, the free energy of solvation can be decomposed into two different terms[34]. The polar part of SFE ($\Delta G_{PB/GB}$) which corresponds with the polarization free energy in the solvation process, is usually calculated by solving the Poisson-Boltzmann (PB) equation or the Generalized Born (GB) equation. The nonpolar part is associate with the free energy cost of the creating a cavity in solvent to accommodate the solute molecule, and the dispersion and repulsive interactions between the solute and solvent. The nonpolar solvation free energy is typically estimated by scaling the solvent accessible surface area (SASA) of the solute molecule[35] as shown in Equations 3 and 4, where γ is the surface tension parameter related to the solvent, and b is the constant term.

$$\Delta G_{solv} = \quad \Delta G_{PB/GB} + \Delta G_{nonpolar} \tag{3}$$

$$\Delta G_{nonpolar} = \gamma SASA + b \tag{4}$$

In the first article of this series, we have optimized radius parameters for PB calculation to reproduce the polar solvation free energies calculated by thermodynamic integration (TI) with the ABCG2 charge model.[36] The newly developed MM-PBSA model achieved an outstanding performance with RMSE of 1.05 kcal/mol and Pearson correlation coefficient of 0.95 for 544 molecules. The PB radius parameters as well as the γ and b parameters for $\Delta G_{nonpolar}$ calculations were listed in Table S6 and Equation S1. The hydration free energy (HFE) of a molecule $\Delta G_{water\ HFE}$ will be calculated using the above model in our FElogP method. For the solvation free energy of a molecule in n-octanol, we will evaluate the performance of the developed PB radius parameters and make adjustment whenever it is necessary. A new nonpolar model reflecting the n-octanol solvent environment will be developed. To make FElogP an efficient method, the calculations of SFE in water (HFE) and SFE in n-octanol will be performed using the same single conformation.

2.2 Data preparation

All the experiment SFE data in water and in n-octanol were collected from the FreeSolv v0.52 database[37] and Minnesota Solvation Database version 2012[38], respectively. All the structures of the molecules were downloaded from the FreeSolv database. 663 small molecules have the experimental SFE data in water, out of which 243 molecules have the experimental SFE data in octanol in the Minnesota Solvation Database. We excluded some molecules whose HFE by thermodynamic integration (TI) differed more than 1.5 kcal/mol from the experimental values, resulting in 544 molecules in the training set for HFE, and 243 molecules for n-octanol SFE. 156 molecules have both experimental values of HFE and SFE in n-octanol solvent. The molecular names and the corresponding experimental values were listed in Table S7.

All the small molecules in the training set were optimized by Gaussian 16[39] using the Hartree-Fock theory of model and the 6–31G* basis set. The minimized structures were then used to generate the topology files and GAFF2 force field parameter files with the ABCG2 charge model[36] using Antechamber.[40] Each molecule was solvated into a cubic water box using the explicit TIP3P water model[41], and the minimum distance of the solute molecule to the edges of the box was set to 12 Å.

The experimental logP data came from Martel's study as detailed above.[31] To prepare the data for MM-PBSA calculation, we obtained the structure information of the 707 molecules in the simplified molecular-input line-entry system (SMILES) format. The SMILES strings were then converted to the Mol2 format files using Open Babel[42]. Following the aforementioned protocol, we generated the topology files for the 707 molecules using the GAFF2 force field parameters and ABCG2 charge model[36]. This high-quality experimental logP data serves as the test set to evaluate the accuracy of our logP prediction model.

2.3 *Ab-initio* logP calculation

We performed *ab initio* calculations using Gaussian 16[39] for all the training set and test set molecules to calculate SFEs using SMD method.[43] To calculate SFE of a molecule, we first carried out geometric optimization using the B3LYP/6–31G* and/or HF/6–31G* model chemistry with the SCRF and SMD keywords; we then performed a single point calculation at the same level of model chemistry without considering solvent effect. The energy difference was calculated as the solvation free energy of the molecule. logP was calculated in two modes. In the first mode (single point), both HFE and n-octanol SFE were calculated using a single input geometry; in the second mode (geometry optimization), both HFE and n-octanol SFE were calculated normally as described above.

2.4 Minimization and molecular dynamics simulation

The molecular mechanics (MM) minimization and molecular dynamics (MD) simulation were conducted using the pmemd.mpi module in AMBER18[44] package for the molecules in training sets. For each system, energy minimization was conducted before the MD simulation. Both the steepest descent and conjugate gradient methods were applied in this step. The number of the cycles of steepest descent was set to 1000, and then switched to conjugate gradient method and ran another 1000 cycles.

The MD simulation can be divided into three phases: heating, equilibration and production phases. At the heating phase, the system was heated from 0 K to 298.15 K in 100 ps using a time step of 0.001 ps. For the equilibration and production phases, the temperature was set to 298.15 K using Langevin dynamics[45] with the collision frequency of 2.0 ps^{-1}. After 0.1 ns equilibration phase, 5.0 ns MD production was conducted, and the trajectories were saved every 10 ps for the following post-MD analysis. For all the MD phases, periodic boundary condition was applied and the pressure was set to 1.0 bar with 1.0 ps pressure relaxation time. The SHAKE algorithm was applied to constrain the hydrogen atoms.

We also performed implicit MD simulations for some ZINC molecules in the Martel dataset, to study the conformation-dependence of logP calculation using FElogP. The exterior dielectric constant was set to 80 and 9.8629 to mimic aqueous and n-octanol solvent environment. The nonpolar contribution was turned off (*gbsa* = 0). The nonbonded cutoff was set to 999 Å to explicitly calculated all pairs of the nonbonded interactions. 190 snapshots were evenly selected from 1–10 ns for FElogP calculations.

2.5 MM-PBSA calculation.

MM-PBSA calculations were based on 3D structures of solute molecules, which were either multiple conformations extracted from the MD snapshots (for model development), or a single conformation automatically generated by OpenBabel (for standard protocol of FElogP calculation). The trajectories acquired in the MD production phase were used to generate the snapshots for each solute (with water stripped) using the cpptraj[46] module in Amber Tools. All the PB calculations were performed using Delphi 95 software[47,48]. The HFE of the 544 molecules (Table S1) and the SFE in n-octanol of 243 molecules (Table S2) were calculated using the optimized PB radius parameters in the first article of this series (Table S6). For SFE in different solvent environment, different experimentally determined external dielectric constants were applied (80 for water and 9.8629 for n-octanol). The nonpolar contribution was estimated based on SASA calculated by an internal program, and the probe radius was set to 1.4 Å. The detailed procedure to generate the model for the nonpolar SFE in different solvents will be introduced in the next section. The accuracy of the final SFE prediction was assessed by using a set of statistical metrics: root-mean-square-error (RMSE), mean unsigned error (MUE), mean signed error (MSE), prediction index (PI) and Pearson's correlation coefficient R.

2.6 Solvation free energy in n-octanol modeling

In our previous study, the importance of the PB radius parameters in free energy calculation was revealed and a set of radius parameters compatible with the newly developed ABCG2 charge model was developed.[36] We found that ABCG2, although developed using hydration free energies, also had excellent performance in SFE calculations using TI for a variety of solvents. Thus, this set of radius parameters for PB calculations were adopted directly in the n-octanol SFE calculations. We would adjust some radius parameters whenever it was necessary. However, the nonpolar solvation free energy model for n-octanol must be redeveloped due to the different dielectric constants of n-octanol and water. The nonpolar model was trained by conducting the regression between the surface area and the values from the experimental SFE data subtracted by the polar SFE (PB) contribution ($\Delta G_{solv}^{expt} - \Delta G_{solv}^{PB}$).

2.7 logP calculations

Last we performed logP using FElogP methods for two molecular sets: the 156-molecule set which have measured HFE and n-octanol SFE, and the 707-molecule Martel set. Note that after PB radii parametrization and non-polar model construction, the stand protocol of applying FElogP calculation for any solute molecule only utilizes a single conformation which is automatically generated by OpenBabel. For each molecule in the second molecular set, logP was also calculated using several widely-used logP models in the public domain implemented in either freeware or commercial software.

3. Results

3.1 The performance of the optimized PB radii combined with new nonpolar model on octanol SFE prediction

After conducting the PB calculation using the optimized radii from the previous study, the differences between the experiment data and PB results (accounting for the nonpolar part of solvation free energy) were obtained. Then the new nonpolar model was established by fitting the solvent-accessible surface area (SASA) to reproduce $\Delta G_{solv}^{expt} - \Delta G_{solv}^{PB}$. We conducted a 10-fold cross validation for 1000 rounds to minimize the bias of the model. Specifically, the data was randomly divided into 10 groups and each time 9 out of 10 groups of data were used as the training set and the remaining group was used as the test set. The mean RMSE was 1.76 kcal/mol for the total of 1000 rounds. The final model for n-octanol SFE estimation was obtained by fitting all the 243 n-octanol SFE data:

$$SFE_{octanol} = PB - 0.012 * SASA + 2.82 \tag{5}$$

The solvation free energies in octanol calculated by the MM-PBSA model and SMD models were summarized in Table S2. Overall, our MM-PBSA n-octanol solvation model achieved a similar performance (RMSE = 1.83 kcal/mol) as SMD at the HF/6–31G* and B3LYP/6–31G* levels of theory, which had an RMSE of 1.84 and 1.68 kcal/mol, for the two *ab initio* models, respectively.

The scatter plot of the experimental versus calculated solvation free energies in n-octanol for the 243 training set molecules were shown in Figure 1B. The performance of *ab initio* SMD and PBSA models was summarized in Table 1.

It is pointed out that we attempted to optimize the PB radius parameters to further improve the n-octanol SFE model, but the improvement was incremental. For the sake of maximizing the error cancellation in the logP prediction, we decided to use the same PB radius parameters optimized for HFE in n-octanol SFE solvation model.

3.2 logP prediction for the 156-molecule training set

As shown in Eq. 1 and Eq. 2, logP can be estimated using the transfer free energy from water to n-octanol phases. The hydration free energy and the solvation free energy were calculated using the newly developed PB models in conjunction with ABCG2 charge model. We first test the rigorousness of the theory of calculating logP from the transfer free energy. As shown in Table S3, for the 156 molecules which have measured logP, HFE and n-octanol SFE, the calculated logP using the measured HFE and SFE agreed very well with the measured logP for all molecules except for p-bromophenol, which has a prediction error of 0.89 (the measured and calculated values are 2.59 and 3.48 respectively). The MUE and RMSE for all 156 molecules are only 0.07 and 0.14 log units, which demonstrate the theory rigorousness of Eq. 1 and Eq. 2. We also evaluated the SMD model at the B3LYP/6–31G* level of model chemistry in logP prediction. The performance metrics of MUE, RMSE and R are 0.47, 0.69, 0.86 respectively. The performance of FElogP is marginally worse, whose MUE, RMSE and R are 0.78, 0.92, 0.77 respectively. The comparison between the experimental versus predicted logP values were illustrated in Figure 2. It is pointed out that the two outliers in Figure 2B are chi030 and choni01, two iodine-containing molecules.

3.3 logP prediction for the 707-molecule druglike test set

To validate our newly proposed logP model, the logP prediction accuracy was further tested on the 707 druglike molecules reported by the Martel's study[31]. All these molecules collected from the ZINC database have high structural diversity, and do not contain any permanently charged molecules. Moreover, the molecular weights for those molecules are distributed between 160 and 600 Da. As shown in Table 2, the RMSE of the predictions using FElogP is 0.91, which is the lowest among several commonly-used logP models including the logP models implemented in Open Babel[42] (RMSE = 1.13), Schrodinger's Qikprop (RMSE = 1.25), Tripos's Sybyl (RMSE = 1.55, ClogP model) and SimulationPlus's ADMET Predictor (RMSE = 2.03, MlogP model). The scatter plots of the experimental versus predicted logP using five different methods were shown in Figure 3. The detailed calculation results by using different methods were listed in Table S4.

4. Discussion

4.1 Advantage of applying the principle of transfer free energy in logP prediction

The accurate prediction of logP is an important topic in molecular modeling and computer aided drug design (CADD) due to its significant role in drug delivery. A full spectrum of calculation methods has been adopted in logP prediction, ranging from the fast QSPR methods to time-consuming alchemical free energy methods. Although many logP models have been developed and deployed in public domain, scientists are still seeking more ideal models which are accurate, efficient, and more importantly, robust for arbitrary molecules no mater they are similar to the training set molecules or not. FElogP, a transfer free energy-based physical model entails those features. First, FElogP was developed without using the measured logP data, thus, its performance is more irrelevant with the training set molecules for logP model construction. The number of parameters optimized for solvation free energy calculations using MM-PBSA is limited (13 for PB raii and 4 for two nonpolar solvation free energy models), so the two solvation models are quite robust. Second, the current computational protocol using only one single conformation, so the computational efficiency is satisfactory, albeit is slower than those simple atom or fragment-based methods.

The performance of FElogP model was critically evaluated using a set of 707 druglike molecules measured by Martel et al.[31] The logP values of those molecules were experimentally determined by UHPLC followed by UV or MS detection. According to Martel, previous logP experiment datasets suffer from a poor coverage of chemical space, inadequate and heterogenous experimental condition for obtaining these logP data[31], thus led to unsatisfying predictive power of the logP prediction tools. For the ideal test set, FElogP outperformed several popular logP models implemented in freeware and commercial software packages. Thus, implicit solvent model-based approaches have a great advantage over those QSPR and even machine learning-based approaches in term of model robustness.

4.2 Comparison of FElogP with other transfer free energy-based methods

We developed a nonpolar model for SFE in n-octanol calculations, while for the polar solvation free energy, the PB radius parameter set was the same as for HFE calculations. The RMSE of SFE calculations was 1.83 kcal/mol for 243 molecules (Table 1), worse than RMSE of HFE calculations, which was 1.05 kcal/mol for 544 molecules. We were curious if the performance drop-off indicated our n-octanol SFE model was very poor. We conducted SFE calculations for all the molecules in the training set using *ab initio* methods. As shown in Table 1, the performance of our PBSA model is between that of two model chemistry, HF/6–31G* + SMD and B3LYP/6–31G* + SMD. We expected the n-octanol SFE by using MM-PBSA can be improved with more experimental values being determined for more structurally diverse molecules. We further explored how well B3LYP/6–31G* + SMD perform in logP prediction applying the same principle of transfer free energy. As shown in Table S3, the *ab initio* method achieved a slightly better performance than FElogP (0.69 versus 0.92 log units).

We also calculated logP for the Martel molecular set using B3LYP/6–31G* + SMD model chemistry. Two computational protocols were applied, one used the same single conformations generated by OpenBabel to conduct single point calculations; and the other performed geometry optimizations using polarizable continuum model (PCM) with the aqueous and n-octanol solvents (Figure 4). The logP calculation results for the Martel dataset are quite astonishing. The RMSE between experimental and calculated logP are 2.16 and 2.67 log units for the first and second protocols, respectively. Compared to our FElogP method, logP predicted with SMD not only have larger prediction deviation, but also have worse correlation (Table 2). Note that this performance is dramatically different from that for the 156-molecular set, for which the RMSE are 0.92 and 0.69 log units for FElogP and the *ab initio* model, respectively.

Ab initio calculation is theoretically more rigorous than MM-PBSA method and has long been regard as an accurate calculation method. One possible explanation of our finding is that the universal parameters for calculating the nonpolar term of any solvents in SMD maybe overfitted. In our FElogP model, we used the same PB radius parameter set for the polar solvation free energy calculations in both aqueous and n-octanol solvents, and the same radius parameter set for solvent accessible surface area calculation. This strategy can maximize the error cancellation in solvation free energy calculations. We hypothesized that applying SASA to estimate the nonpolar part of solvation free energy, although quite simple, is more suitable for logP prediction using the principle of transfer free energy.

4.3 Conformation-dependence of logP prediction with FElogP

To make FElogP method efficient, we performed logP calculation of a molecule using a single 3D structure generated by OpenBabel. It is very important to investigate if FElogP predicted values are conformation-dependent. As shown in Table S4, there are 21 out of 707 molecules have been identified as outliers. Note that we recognize a molecule is an outlier if the prediction error is equal to or larger than 2 log units. The molecular structures of those outliers were illustrated in Figure 5. Unfortunately, there are little common structural features of those outliers.

We then investigated if we could improve the prediction using multiple conformations. First of all, we conducted two implicit MD simulations on the 21 outlier molecules using a GB model of Hawkins et al.[49] to sample multiple conformations, mimicking aqueous and n-octanol solvent environment. The resulted MD snapshots were labeled as the aqueous and n-octanol conformational sets, correspondingly. We then designed five different methods to calculate logP, which are (1) HFE and SFE respectively used the aqueous and n-octanol conformational sets; (2) both HFE and SFE used the aqueous conformational set; (3) both HFE and SFE used the n-octanol conformational set; (4) FElogP calculations were carried out for every conformation in the aqueous conformation set and then take an average; and (5) FElogP calculations were performed for every conformation in the n-octanol conformation set and then take an average. For the last two methods, the mean values and the standard deviations were calculated.

It turned out that the logP prediction errors using multiple conformations were decreased slightly compared to the standard protocol for which only one conformation was applied to calculate logP. As shown in Table 3, the RMSE was dropped reduced from 2.57 for the standard protocol, to 2.42 for Method 1, 2.16 for Method 2, and 2.15 for Method 3. As shown in Table S5, the RMSE were 2.16 for Method 4 and 2.15 for Method 5, respectively. The average standard deviation of FElogP calculations for 190 snapshots are 0.26 and 0.27 for Method 4 and Method 5, respectively. The low standard deviation values of Methods 4 and 5 suggested that our FElogP method is basically conformation-independent. When multiple conformations are used in FElogP calculations, the prediction performance can be improved using either of the last four methods. It is reasonable for Method 1 being inferior to the other methods, as the error cancellation became less effective when two different conformational sets were applied in solvation free energy calculations.

5. Conclusion:

Due to its important role in drug discovery and development, accurate prediction of logP of an arbitrary pharmaceutical molecule is highly important. An ideal logP model is accurate, efficient and robust for any druglike molecules. A transfer free energy-based approach has a potential to satisfy the above requirement. We developed the FElogP model in the spirit of applying the transfer free energy principal to predict logP. Our method avoided using the measured logP values to construct models, thus achieved high mode robustness. Indeed, the RMSE values for the 156-molecule training set and the 707-molecule test set are very similar, which are 0.92 and 0.91 log units, respectively. Note that our FElogP calculation on a molecule only utilize one 3D conformation which is automatically generated by free software OpenBabel. We further explored the influence of considering multiple solute conformations to the logP prediction results. It turned out that calculate logP from multiple conformations could reduce the prediction to some degree. Meanwhile, a solvation model for n-octanol using MM-PBSA was developed and its performance was comparable with the SMD model using B3LYP/6–31G* level of theory. We believe our FElogP will have great applications in logP prediction, especially for molecules which are distinct from those molecules in the training sets for constructing logP models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

This work was supported by National Institutes of Health (Grant No: R01GM147673 to JW) and the National Science Foundation (Grant No: 1955260 to JW). The authors also thank the computing resources provided by the Center for Research Computing (CRC) at University of Pittsburgh.

Data And Software Availability

All solvation free energy and protein ligand binding data, and software packages (Delphi, AMBER) come from public domain

References

1. Hughes JD; Blagg J; Price DA; Bailey S; DeCrescenzo GA; Devraj RV; Ellsworth E; Fobian YM; Gibbs ME; Gilles RW; Greene N; Huang E; Krieger-Burke T; Loesel J; Wager T; Whiteley L; Zhang Y Bioorganic & Medicinal Chemistry Letters 2008, 18(17), 4872–4875. [PubMed: 18691886]
2. Waring MJ Bioorganic & Medicinal Chemistry Letters 2009, 19(10), 2844–2851. [PubMed: 19361989]
3. Gleeson MP Journal of Medicinal Chemistry 2008, 51(4), 817–834. [PubMed: 18232648]
4. Johnson TW; Dress KR; Edwards M Bioorganic & Medicinal Chemistry Letters 2009, 19(19), 5560–5564. [PubMed: 19720530]
5. Fujita T; Iwasa J; Hansch C Journal of the American Chemical Society 1964, 86(23), 5175–5180.
6. Iwasa J; Fujita T; Hansch C Journal of Medicinal Chemistry 1965, 8(2), 150–153. [PubMed: 14332653]
7. Lipinski CA; Lombardo F; Dominy BW; Feeney PJ Adv Drug Deliv Rev 2001, 46(1–3), 3–26. [PubMed: 11259830]
8. Lipinski CA Drug Discov Today Technol 2004, 1(4), 337–341. [PubMed: 24981612]
9. Sangster J Journal of Physical and Chemical Reference Data 1989, 18(3), 1111–1229.
10. Leo A; Hansch C; Elkins D Chemical Reviews 1971, 71(6), 525–616.
11. Arup Ghose VV Combinatorial Library Design and Evaluation: Principles, Software, Tools, and Applications in Drug Discovery; CRC Press, 2001.
12. Waterbeemd H. v. d. Chemometric Methods in Molecular Design; John Wiley & Sons, 2008.
13. Liao Q; Yao J; Yuan S Molecular Diversity 2006, 10(3), 301–309. [PubMed: 17031534]

14. Ghose AK; Pritchett A; Crippen GM Journal of Computational Chemistry 1988, 9(1), 80–90.

15. Labute P; Chemical Computing Group, Inc.: 1010 Sherbrooke Street West, Suite 910, Montreal, Quebec, H3A 2R7, Canada 1998.

16. Wildman SA; Crippen GM Journal of Chemical Information and Computer Sciences 1999, 39(5), 868–873.

17. Tetko IV; Tanchuk VY Journal of Chemical Information and Computer Sciences 2002, 42(5), 1136–1145. [PubMed: 12377001]

18. Leo AJ; Hoekman D Perspectives in Drug Discovery and Design 2000, 18(1), 19–38.

19. Klopman G; Li J-Y; Wang S; Dimayuga M Journal of Chemical Information and Computer Sciences 1994, 34(4), 752–781.

20. Leo A; Jow PY; Silipo C; Hansch C J Med Chem 1975, 18(9), 865–868. [PubMed: 1159707]

21. Chou JT; Jurs PC Journal of Chemical Information and Computer Sciences 1979, 19(3), 172–178.

22. Moriguchi I; Hirono S; Nakagome I; Hirano H Chemical & Pharmaceutical Bulletin 1994, 42(4), 976–978.

23. Ulrich N; Goss KU; Ebert A Commun Chem 2021, 4(1).

24. Tetko IV; Poda GI In Molecular Drug Properties, 2007, p 381–406.

25. Procacci P; Guarnieri G Journal of Computer-Aided Molecular Design 2020, 34(4), 371–384. [PubMed: 31624982]

26. Ogata K; Hatakeyama M; Nakamura S Molecules 2018, 23(2).

27. Bryant MJ; Black SN; Blade H; Docherty R; Maloney AGP; Taylor SC Journal of Pharmaceutical Sciences 2019, 108(5), 1655–1662. [PubMed: 30615878]

28. Shultz MD Journal of Medicinal Chemistry 2019, 62(4), 1701–1714. [PubMed: 30212196]

29. Pozzan A In Quantum Mechanics in Drug Discovery; Heifetz A, Ed.; Springer US: New York, NY, 2020, p 285–305.

30. Işık M; Bergazin TD; Fox T; Rizzi A; Chodera JD; Mobley DL Journal of Computer-Aided Molecular Design 2020, 34(4), 335–370. [PubMed: 32107702]

31. Martel S; Gillerat F; Carosati E; Maiarelli D; Tetko IV; Mannhold R; Carrupt PA Eur J Pharm Sci 2013, 48(1–2), 21–29. [PubMed: 23131797]

32. Irwin JJ; Sterling T; Mysinger MM; Bolstad ES; Coleman RG J Chem Inf Model 2012, 52(7), 1757–1768. [PubMed: 22587354]

33. Bannan CC; Calabró G; Kyu DY; Mobley DL Journal of Chemical Theory and Computation 2016, 12(8), 4015–4024. [PubMed: 27434695]

34. Wang J; Hou T; Xu X Current Computer - Aided Drug Design 2006, 2(3), 287–306.

35. Izairi R; Kamberaj H J Chem Inf Model 2017, 57(10), 2539–2553. [PubMed: 28880080]

36. He X; Man VH; Yang W; Lee TS; Wang J J Chem Phys 2020, 153(11), 114502. [PubMed: 32962378]

37. Mobley DL; Guthrie JP Journal of Computer-Aided Molecular Design 2014, 28(7), 711–720. [PubMed: 24928188]

38. Marenich AVK, C. P.; Thompson JD; Hawkins GD; Chambers CC; Giesen DJ; Winget P; Cramer CJ; Truhlar DG: University of Minnesota, Minneapolis, 2012.

39. Frisch MJT, G. W.; Schlegel HB; Scuseria GE; Robb MA; Cheeseman JR; Scalmani G; Barone V; Petersson GA; Nakatsuji H; Li X; Caricato M; Marenich AV; Bloino J; Janesko BG; Gomperts R; Mennucci B; Hratchian HP; Ortiz JV; Izmaylov AF; Sonnenberg JL; Williams-Young D; Ding F; Lipparini F; Egidi F; Goings J; Peng B; Petrone A; Henderson T; Ranasinghe D; Zakrzewski VG; 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 JA Jr.; Peralta JE; Ogliaro F; Bearpark MJ; Heyd JJ; Brothers EN; Kudin KN; Staroverov VN; Keith TA; Kobayashi R; Normand J; Raghavachari K; Rendell AP; Burant JC; Iyengar SS; Tomasi J; Cossi M; Millam JM; Klene M; Adamo C; Cammi R; Ochterski JW; Martin RL; Morokuma K; Farkas O; Foresman JB; Fox DJ. Gaussian, Inc: Wallingford, CT, 2016.

40. Wang J; Wang W; Kollman PA; Case DA J Mol Graph Model 2006, 25(2), 247–260. [PubMed: 16458552]

41. Jorgensen WL; Chandrasekhar J; Madura JD; Impey RW; Klein ML The Journal of Chemical Physics 1983, 79(2), 926–935.

42. O'Boyle NM; Banck M; James CA; Morley C; Vandermeersch T; Hutchison GR Journal of Cheminformatics 2011, 3(1), 33. [PubMed: 21982300]

43. Marenich AV; Cramer CJ; Truhlar DG J Phys Chem B 2009, 113(18), 6378–6396. [PubMed: 19366259]

44. Case DA, B.-S. IY, Brozell SR, Cerutti DS, Cheatham TE III, Cruzeiro VWD, Darden TA, Duke RE, Ghoreishi D, Gilson MK, Gohlke H, Goetz AW, Greene D, R Harris N Homeyer, Huang Y, Izadi S, Kovalenko A, Kurtzman T, Lee TS, LeGrand S, Li P, Lin C, Liu J, Luchko T, Luo R, Mermelstein DJ, Merz KM, Miao Y, Monard G, Nguyen C, Nguyen H, Omelyan I, Onufriev A, Pan F, Qi R, Roe DR, Roitberg A, Sagui C, Schott-Verdugo S, Shen J, Simmerling CL, Smith J, SalomonFerrer R, Swails J, Walker RC, Wang J, Wei H, Wolf RM, Wu X, Xiao L, York DM and Kollman PA. University of California, San Francisco: San Francisco, 2018.

45. Larini L; Mannella R; Leporini D J Chem Phys 2007, 126(10), 104101. [PubMed: 17362055]

46. Roe DR; Cheatham TE Journal of Chemical Theory and Computation 2013, 9(7), 3084–3095. [PubMed: 26583988]

47. Rocchia W; Alexov E; Honig B The Journal of Physical Chemistry B 2001, 105(28), 6507–6514.

48. Li L; Li C; Sarkar S; Zhang J; Witham S; Zhang Z; Wang L; Smith N; Petukh M; Alexov E BMC Biophys 2012, 5, 9. [PubMed: 22583952]

49. Hawkins GD; Cramer CJ; Truhlar DG J Phys Chem-Us 1996, 100(51), 19824–19839.

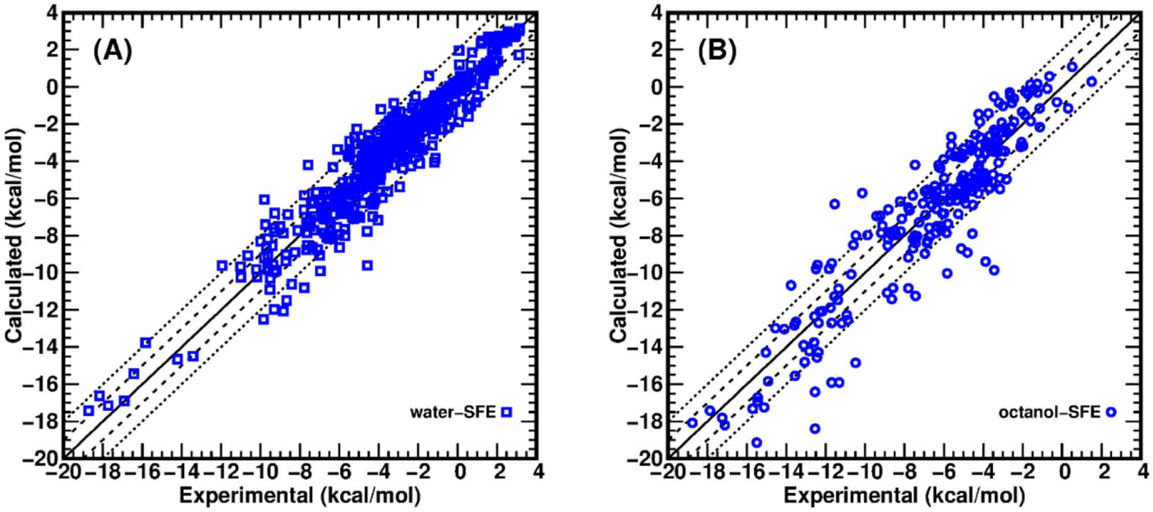


Figure 1.

Comparison of the experimental and calculated solvation free energies for the training set molecules. A: hydration free energies of 544 training set molecules. B: solvation free energies of 243 training set molecules in n-octanol solvent. Eye-guided lines are shown for ideal matching of calculation vs. experiment (solid line), with error of ±1 kcal/mol (dashed line), and with error of ±2 kcal/mol (dotted line), respectively.

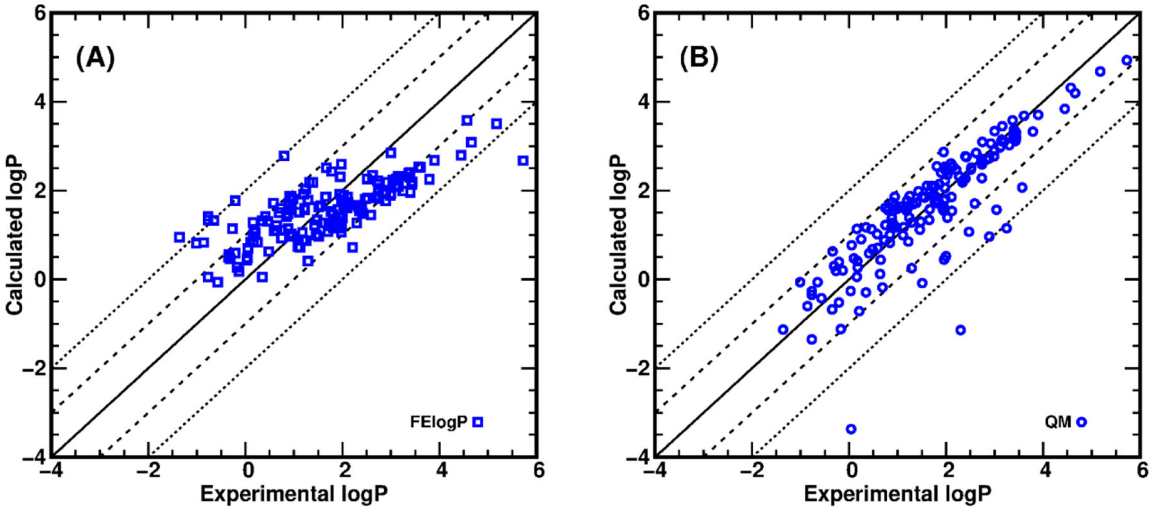


Figure 2.

Comparison of the experimental and calculated logP for the 156 molecules which have measured log P and solvation free energies in water and in n-octonal. A: logP was calculated using FElogP; B: logP was calculated using the SMD at the B3LYP/6–31G* level of model chemistry. Eye-guided lines are shown for ideal matching of calculation vs. experiment (solid line), with error of ±1 kcal/mol (dashed line), and with error of ±2 kcal/mol (dotted line), respectively.

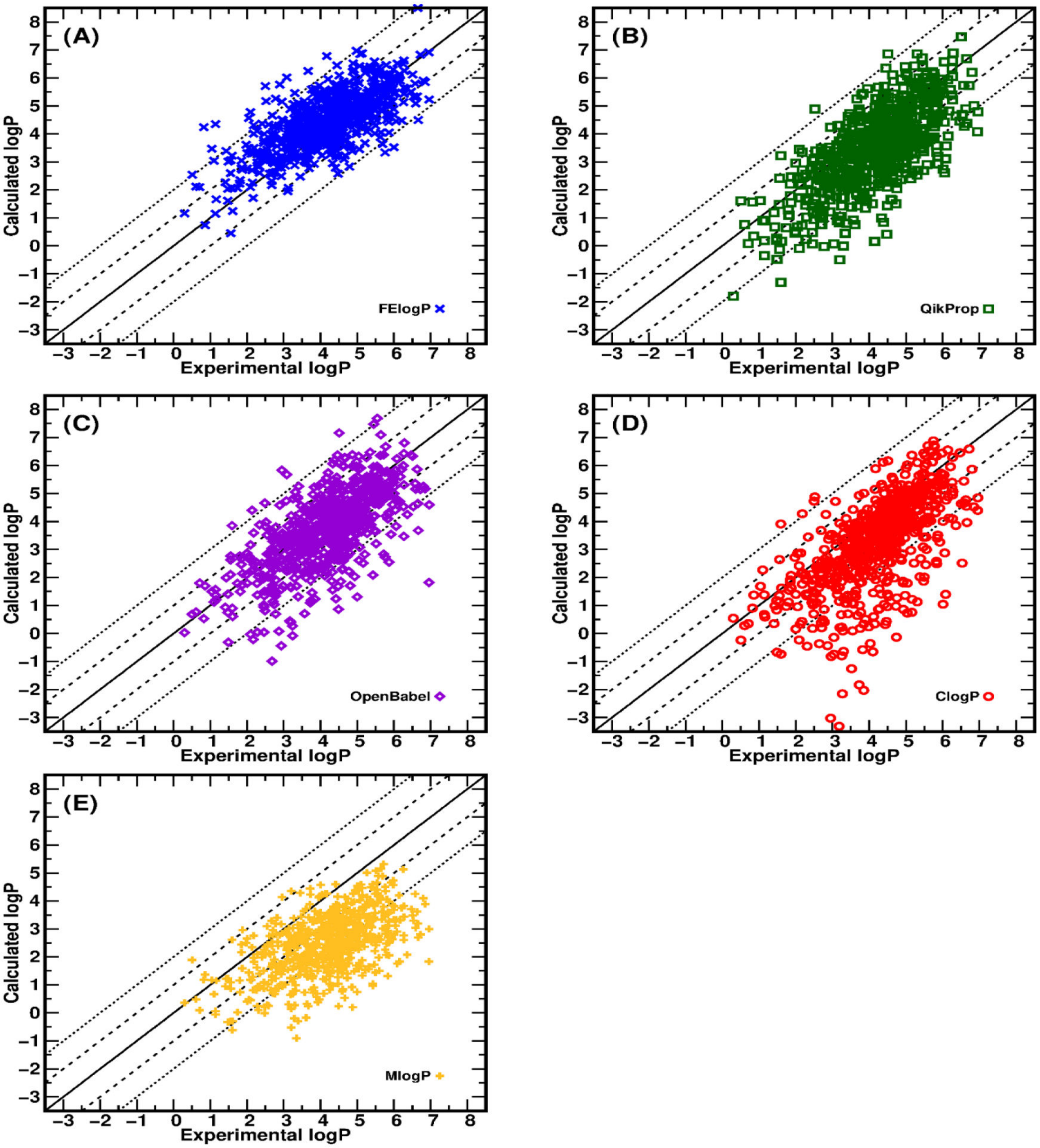


Figure 3.

Comparison of the experimental and calculated logP for the 707 ZINC molecules in the Martel dataset. logP was calculated using five different models, which are FElogP (A), QikProp module in Schrodinger Maestro (B), OpenBabel (C), ClogP in Sybyl (D), and MlogP in ADMET Predictor (E). Eye-guided lines are shown for ideal matching of calculation vs. experiment (solid line), with error of ±1 kcal/mol (dashed line), and with error of ±2 kcal/mol (dotted line), respectively.

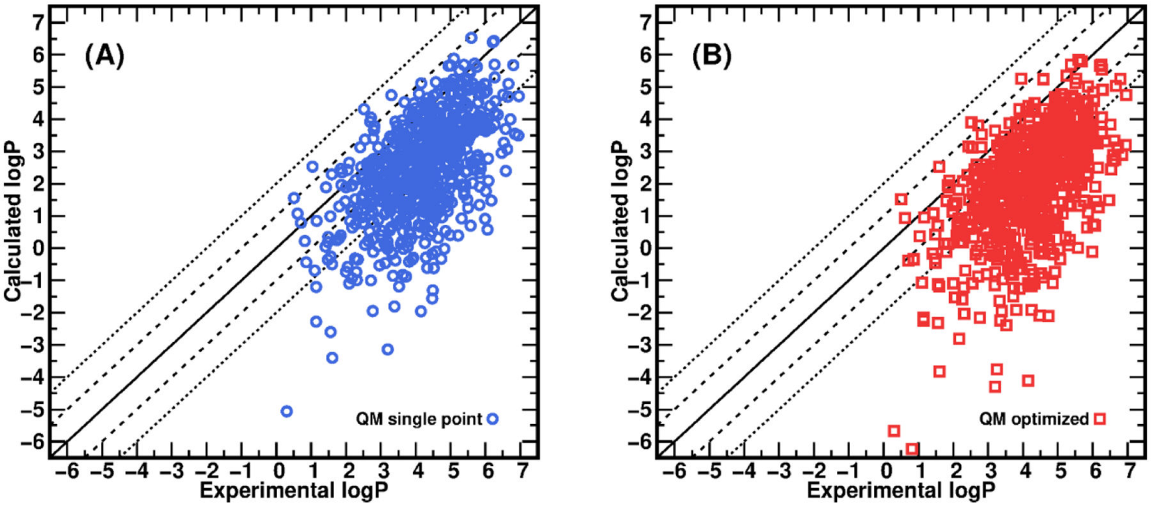


Figure 4.

Comparison of the experimental and calculated logP for the 707 ZINC molecules in the Martel dataset. logP was calculated using two QM methods. A: SMD calculation with the input geometries (single point). B: SMD calculations using optimized geometries.

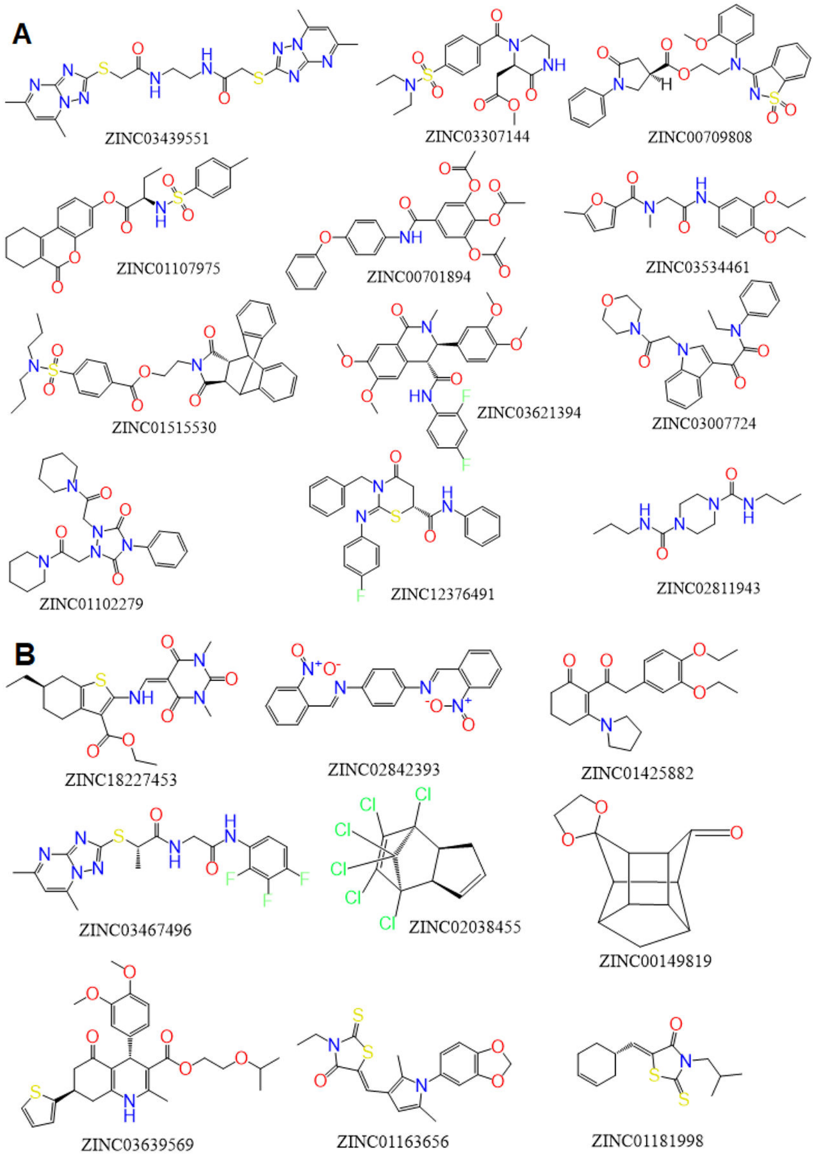


Figure 5.

2D-Structures of outliers who had the predicted logP different from the measured one at least 2 log units. A: molecules 1–14; B: molecules 15–21.

Table 1.

The comparison of solvation free energy calculations in n-octanol for 243 molecules

Method	MSE	MUE	RMSE	PI	R
PBSA (new model in this work)	−0.10	1.37	1.83	0.92	0.92
HF/6–31G* + SMD	−0.71	1.29	1.84	0.93	0.92
B3LYP/6–31G* + SMD	0.53	1.21	1.68	0.93	0.92

Table 2.

The comparison of logP prediction using seven different approaches/models for Martel dataset (N = 707).

Method	MSE	MUE	RMSE	PI	R
FElogP (this work)	0.28	0.70	0.91	0.70	0.71
Maestro (QikProp) *	−0.70	0.99	1.25	0.71	0.71
OpenBabel	−0.43	0.86	1.13	0.69	0.67
Sybyl (ClogP)	−0.93	1.12	1.55	0.70	0.65
ADMET Predictor (MlogP)	−1.72	1.77	2.03	0.55	0.55
B3LYP/6–31G + SMD (single point)	−1.68	1.77	2.16	0.57	0.56
B3LYP/6–31G + SMD (geometry optimization)	−2.24	2.27	2.67	0.54	0.53

*For QikProp, four molecules did not yield results in logP prediction.

Table 3.

The influence of conformational sampling on FElogP calculations. Implicit MD simulations were sampled to model both aqueous and n-octanol solvent environments. The compared methods including “Single” (single conformations generated by OpenBabel), “Method1” (HFE and SFE were calculated using conformational sets sampled in aqueous solution and n-octanol, respectively), “Method2” (both HFE and SFE were calculated using the conformational set sampled in aqueous solution), and “Method3” (both HFE and SFE were calculated using the conformational set sampled in n-octanol solution).

Molecule ID	Expt.	FElogP		Method1		Method2		Method3	
		Calc.	Diff.	Calc.	Diff.	Calc.	Diff	Calc.	Diff.
ZINC00149819	0.51	2.54	2.03	2.51	2.00	2.54	2.03	2.52	2.01
ZINC00701894	2.89	5.77	2.88	5.45	2.56	5.39	2.50	5.38	2.49
ZINC00709808	2.50	5.71	3.21	4.63	2.13	4.55	2.05	4.56	2.06
ZINC01102279	2.78	5.26	2.48	4.84	2.06	5.05	2.27	5.10	2.32
ZINC01107975	3.08	6.02	2.94	5.40	2.32	5.52	2.44	5.54	2.46
ZINC01163656	2.52	5.17	2.65	5.29	2.77	5.29	2.77	5.29	2.77
ZINC01181998	6.66	4.50	−2.16	4.64	−2.02	4.53	−2.13	4.51	−2.15
ZINC01425882	3.30	5.46	2.16	4.30	1.00	4.38	1.08	4.40	1.10
ZINC01515530	4.17	6.78	2.61	6.46	2.29	5.89	1.72	6.43	2.26
ZINC02038455	5.69	3.64	−2.05	3.68	−2.01	3.68	−2.01	3.67	−2.02
ZINC02811943	1.04	3.47	2.43	3.47	2.43	3.46	2.42	3.45	2.41
ZINC02842393	2.95	5.15	2.20	5.11	2.16	5.17	2.22	5.22	2.27
ZINC03007724	2.05	4.57	2.52	4.55	2.50	4.34	2.29	4.35	2.30
ZINC03307144	1.14	4.35	3.21	4.30	3.16	3.94	2.80	3.84	2.70
ZINC03439551	0.82	4.24	3.42	6.50	5.68	4.04	3.22	3.69	2.87
ZINC03467496	2.18	4.31	2.13	3.21	1.03	3.48	1.30	3.17	0.99
ZINC03534461	2.09	4.78	2.69	4.91	2.82	4.65	2.56	4.62	2.53
ZINC03621394	3.66	6.25	2.59	5.78	2.12	5.79	2.13	5.81	2.15
ZINC03639569	4.45	6.45	2.00	5.13	0.68	5.44	0.99	5.37	0.92
ZINC12376491	6.00	3.52	−2.48	4.33	−1.67	4.31	−1.69	4.26	−1.74
ZINC18227453	5.76	3.34	−2.42	4.94	−0.82	4.93	−0.83	4.93	−0.83
RMSE	-	2.57		2.42		2.16		2.15	