J Comput Aided Mol Des (2016) 30:1019–1033
DOI 10.1007/s10822-016-9928-x

Extended solvent-contact model approach to blind SAMPL5 prediction challenge for the distribution coefficients of drug-like molecules

Kee-Choo Chung[1] · **Hwangseo Park**[1]

Received: 21 June 2016 / Accepted: 20 July 2016 / Published online: 23 July 2016
© Springer International Publishing Switzerland 2016

Abstract The performance of the extended solvent-contact model has been addressed in the SAMPL5 blind prediction challenge for distribution coefficient (LogD) of drug-like molecules with respect to the cyclohexane/water partitioning system. All the atomic parameters defined for 41 atom types in the solvation free energy function were optimized by operating a standard genetic algorithm with respect to water and cyclohexane solvents. In the parameterizations for cyclohexane, the experimental solvation free energy (ΔG_{sol}) data of 15 molecules for 1-octanol were combined with those of 77 molecules for cyclohexane to construct a training set because ΔG_{sol} values of the former were unavailable for cyclohexane in publicly accessible databases. Using this hybrid training set, we established the LogD prediction model with the correlation coefficient (R), average error (AE), and root mean square error (RMSE) of 0.55, 1.53, and 3.03, respectively, for the comparison of experimental and computational results for 53 SAMPL5 molecules. The modest accuracy in LogD prediction could be attributed to the incomplete optimization of atomic solvation parameters for cyclohexane. With respect to 31 SAMPL5 molecules containing the atom types for which experimental reference data for ΔG_{sol} were available for both water and cyclohexane, the accuracy in LogD prediction increased remarkably with the R, AE, and RMSE values of 0.82, 0.89, and 1.60, respectively. This significant enhancement in performance stemmed from the better optimization of atomic solvation parameters by limiting the element of training set to the molecules with experimental ΔG_{sol} data for cyclohexane. Due to the simplicity in model building and to low computational cost for parameterizations, the extended solvent-contact model is anticipated to serve as a valuable computational tool for LogD prediction upon the enrichment of experimental ΔG_{sol} data for organic solvents.

Keywords SAMPL5 · Distribution coefficient · Solvation free energy · Extended solvent-contact model · Genetic algorithm

Introduction

Partition coefficient (P) refers to the ratio of the equilibrium concentration of a substance in organic solvent to that in water, which is distributed in the mixture of the two immiscible solvents. Due to the hydrophobicity in organic solvent layer, logarithm of P (LogP) can represent the lipophilicity of solute molecules. Because 1-octanol serves as a prototype of organic solvents, LogP with respect to 1-octanol/water system is used as the most popular molecular descriptor. LogP values have thus been measured as an important physicochemical property pertinent to drug-likeness [1], toxicity [2], blood brain barrier to reach the central nervous system [3], and ADMET properties [4–7]. Furthermore, LogP values can also be related with the molecular permeability with respect to the cell membrane that has a lipophilic central layer [8–10]. In addition to the role of a yardstick to measure the molecular lipophilicities, the usefulness of LogP has also been

Electronic supplementary material The online version of this article (doi:10.1007/s10822-016-9928-x) contains supplementary material, which is available to authorized users.

✉ Hwangseo Park
hspark@sejong.ac.kr

[1] Department of Bioscience and Biotechnology, Sejong University, 209 Neungdong-ro, Kwangjin-gu, Seoul 143-747, Republic of Korea

appreciated in estimating the dehydration cost for binding of a small molecule to the receptor protein [11]. LogP is thus a representative of the most prevalent molecular descriptors to quantify the pharmacological properties of drug candidates.

Due to the necessity in the discovery of new drugs and materials, a great deal of efforts has been devoted to the development of the reliable computational methods for LogP prediction. It was assumed in a large part of the computational methods that molecular LogP values would be obtained by summing all the contributions from the individual atoms or from the dissected fragments [12, 13]. A number of quantitative structure-property relationship (QSPR) models with reasonable accuracy were also developed for LogP prediction using a variety of molecular descriptors [14–18]. LogP values of small organic molecules were also predicted successfully by comparing the solvation free energies with respect to water and 1-octanol calculated based on the solvent-accessible surface area model [19], the extended solvent-contact approach [20], and 3D density distribution function [21].

Although the usefulness of LogP was well-appreciated in contemporary drug discovery, its weak correlation has been observed with the membrane permeability as well as with the aqueous solubility of some small molecules [22–24]. These discrepancies stem in a large part from the presence of a polar hydroxyl group in 1-octanol, which would be the potential source of error in mimicking the hydrophobic environment [25, 26]. Hence, the logarithm of distribution coefficient (LogD) with respect to the cyclohexane/water partitioning system may serve as a good alternative for the general-purpose molecular descriptor because cyclohexane is an absolutely lipophilic solvent with no polar moiety. Actually, LogD differs from LogP in that the former is measured under consideration of all the possible ionization and tautomerization states of a substance instead of taking into account only a single tautomeric form [27]. LogD proved be the better descriptor than LogP in particular for the molecules capable of establishing the intermolecular or intramolecular hydrogen bonds [28].

Like LogP, LogD has been estimated with reasonable accuracy with various computational methods based on the parametrizations of molecular surface area [29–31], molecular interaction fields [32, 33], and solubility-diffusion theory [34]. Driven by the successful predictions of molecular solvation free energy and LogP [20, 35, 36], we address the usefulness of the solvent-contact model in estimating the molecular LogD values through the participation in SAMPL5 blind prediction challenge for distribution coefficient. To improve the solvation free energy function required for computing the ratio of solute concentrations with respect to cyclohexane and water, we augmented the number of atomic parameters to cope with various chemical environments encountered in 53 SAMPL5 molecules. This modification would have the effect of enhancing the accuracy in predicting the solvation free energy and LogD because the electronic structures and bonding patterns peculiar to drug-like molecules in SAMPL5 dataset can be described appropriately by the extension of atom types. The fundamental assumptions to calculate molecular LogD with the extended solvent-contact model are presented and discussed. We also address the limitations to practical applicability inherent in the extended solvent-contact model, and suggest the reasonable methods for further improvement.

Theory and computational methods

When the solute molecules diffuse passively across the two immiscible solvents, the ratio of equilibrium concentrations of the solute in the two solvents yields the partition coefficient (P). The P value of a solute molecule (S) with respect to cyclohexane/water partitioning system can be defined as follows.

$$\mathrm{P} = \frac{[S]_{chx}}{[S]_{wat}} \tag{1}$$

Because P is expressed in the form of equilibrium coefficient for the diffusion of a solute molecule from water to cyclohexane, its LogP can be related with the difference in solvation free energies ($\Delta\Delta G^0$) with respect to the two solvents. Here, the solvation free energy (ΔG_{sol}) refers to the free energy change for the transfer of a solute molecule from the gas phase to solvent. LogP of a molecule can thus be related with $\Delta\Delta G^0$ as follows when the latter is given in kcal/mol.

$$\mathrm{LogP} = -\frac{\Delta\Delta G^0}{1.364} \tag{2}$$

whereas P is measured from the concentrations of a single neutral form, D is defined with respect to all the forms of a solute molecule available in the two solvents. D should therefore be expressed with the summations running over all the possible ionization and tautomerization states a solute molecule.

$$\mathrm{D} = \frac{\sum_i \gamma_i [T_i]_{chx}}{\sum_k \gamma_k [T_k]_{wat}} \tag{3}$$

Here, γ_i and T_i represent the activity coefficient and the concentration of a single ionization or tautomerization state of the solute molecule in each solvent, respectively.

When a solute molecule with low ionization constant dissolves into the two solvents to form dilute solutions, D can be approximated to P because the D value would be dominated by a single tautomeric form of the solute. We note in this regard that SAMPL5 molecules involve only weakly or hardly ionizable moieties such as carboxylic

acid, amine, and phenol with the ionization constant smaller than 10^{-4}. The LogD value of each SAMPL5 molecule can thus be estimated by the difference between ΔG_{sol} in water and that in cyclohexane.

$$\mathrm{LogD} = \frac{\Delta G_{sol}^{wat} - \Delta G_{sol}^{cyx}}{1.364} \quad (4)$$

Here, ΔG_{sol}^{wat} and ΔG_{sol}^{cyx} refer to the solvation free energies of the solute at 298.15 K in water and in cyclohexane, respectively.

To calculate the LogD values using Eq. (4), we constructed the molecular solvation free energy functions based on the extended solvent-contact model as detailed in the previous papers [35, 36].

$$\Delta G_{sol} = \sum_{i}^{atoms} S_i \left(O_i^{\max} - \sum_{j \neq i}^{atoms} V_j e^{-\frac{r_{ij}^2}{2\sigma^2}} \right) \quad (5)$$

Here, Gaussian envelope function with respect to the interatomic distances (r_{ij}'s) between solute atoms is introduced to define the occupied volume to which the approach of solvent molecules is restricted. S_i, O_i^{max}, and V_i parameters represent the atomic solvation energy per unit volume, maximum atomic occupancy, and atomic fragmental volume, respectively. O_i^{max} and V_i values are related with the volume of a solute atom in the isolated state and that in molecules, respectively. The negative and positive signs of S_i parameter indicate the stabilization and destabilization of the solute atom, respectively, as a consequence of the interactions with solvent molecules. These three atomic parameters assigned to each atom type should be determined for the solvation free energy function to be used in LogD calculations. To optimize all the atomic parameters with respect to water and cyclohexane, a standard genetic algorithm was employed with the training set comprising the molecules for which experimental ΔG_{sol} data were available for both solvents. As widely adopted in the literature, the σ value in Eq. (5) was set equal to 3.5 Å during the parameterizations.

With respect to partitioning the two-solvent system, the cyclohexane phase contains only 0.04 % of water at 298.15 K [37] in contrast to 4 % in 1-octanol/water system. Such an exceptionally high purity is not surprising because cyclohexane is the hydrophobic molecule with no polar group. Therefore, the experimental data for training set molecules were adopted without a composition-weighted correction as the reference ΔG_{sol} values to optimize the atomic parameters in the solvation free energy function.

Preparation of training set

As a preliminary step to optimize the atomic parameters in the solvation free energy function, we had to prepare the training set containing a sufficient number of molecules whose experimental ΔG_{sol} values were available for both water and cyclohexane. In contrast to the abundance of LogP values for a variety of organic molecules, the rarity of experimental LogD data for cyclohexane/water partitioning system made it difficult to establish a proper training set. Because the experimental ΔG_{sol} data for cyclohexane were also insufficient to cope with all SAMPL5 molecules, the training set was constructed by combining the molecules with ΔG_{sol} values for cyclohexane and water with those for which ΔG_{sol} data were available for 1-octanol and water. This was inevitable due to the rarity of experimental ΔG_{sol} data even for the other hydrocarbon solvents such as hexane. Actually, such a combination was the only way to optimize the atomic parameters of the atom types missing in the molecules for which the experimental ΔG_{sol} values in cyclohexane were available. 1-Octanol is likely to serve as an effective surrogate for cyclohexane because it categorizes into a hydrophobic solvent. Indeed, 1-octanol has often been used as a simplified model system for lipid [38] because of high lipophilicity with the low dielectric constant of 10.3 due to the presence of a long hydrocarbon chain.

A total of 92 molecules were collected to construct the training set, the structures of which are illustrated in Fig. 1. The ΔG_{sol} values in cyclohexane for 77 elements served as the reference data for parameterization while those for the rest 15 molecules were approximated with the corresponding ΔG_{sol} values in 1-octanol. This minor subset included the molecules containing the atom types for sp^2 carbon, amide nitrogen, and sulfur. The ΔG_{sol} values of most elements in the training set with respect to cyclohexane, water, and 1-octanol were extracted from Minnesota Solvation Database of version 2012 [39], while those of the remaining molecules including sp^2 carbon, planar and amide nitrogens with 3 substituents, and sp^2 sulfur with two substituents were retrieved from literature [40, 41].

Definition of atom types

The contributions of individual atoms to solvation free energy vary even among the same elements due to the diverse chemical environments with which the atoms in molecule are faced. The atom types should therefore be assigned under consideration of the detailed atomic properties including the hybridization state, electronegativity, and the number of substituents. For example, the specific atom types should be defined for the functional groups with characteristic electronic structure such as carbonyl carbon, phenolic oxygen, and the hydrogen atoms attached to varying heteroatoms. A total of 41 atom types were assigned in this study to discriminate the differences in the

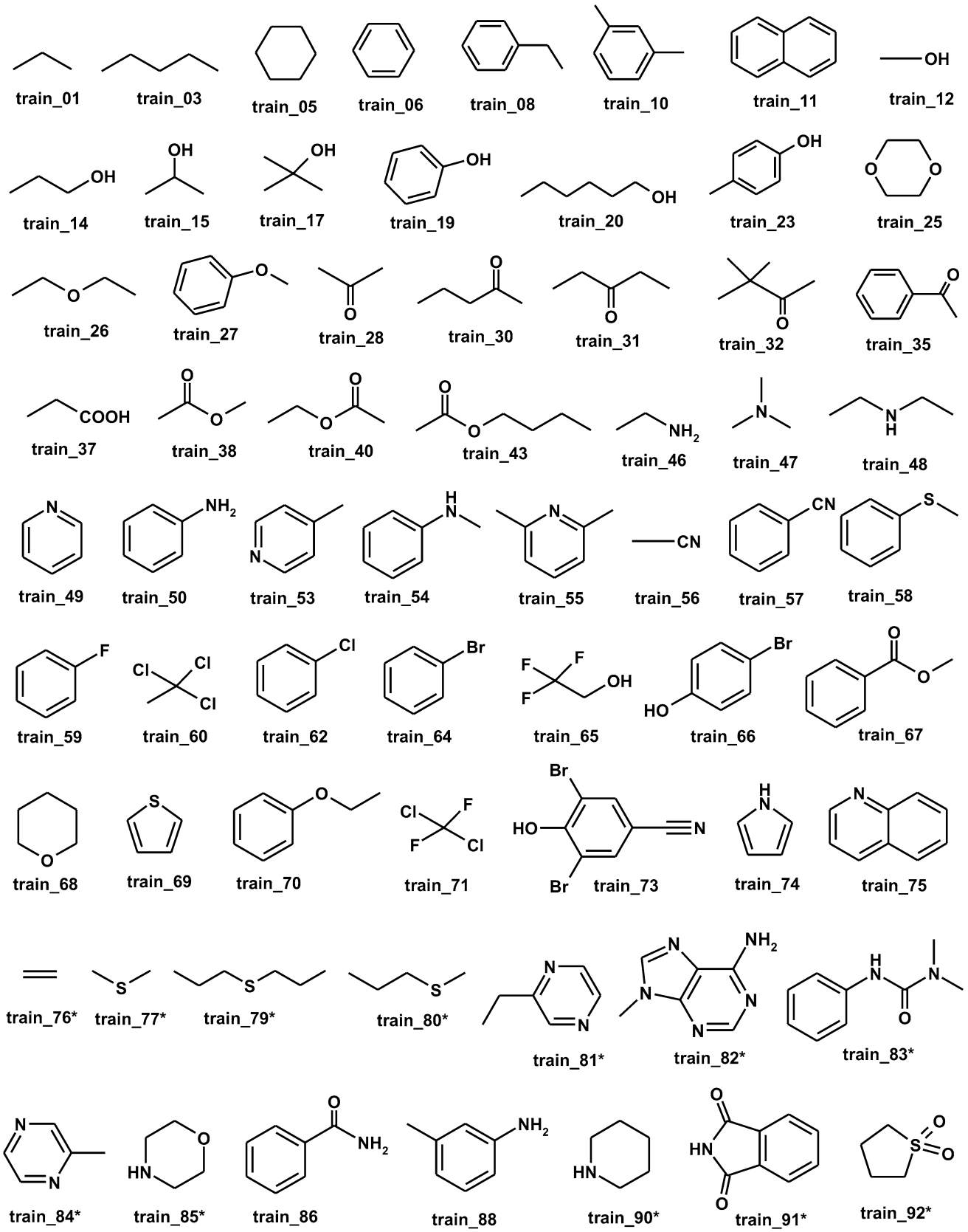


Fig. 1 Chemical structures of the selected molecules in the training set for the optimization of atomic parameters in the solvation free energy function. *Asterisks* indicate the molecules for which the experimental ΔG_{sol} values in 1-octanol were referenced instead of those in cyclohexane

interactions with solvent among the atoms contained in 53 SAMPL5 molecules. The number of atom types reduced to 33 when the fifteen molecules without ΔG_{sol} values for cyclohexane were excluded, which exemplified their necessity in the optimization of solvation free energy function. All atom types were designated in Sybyl MOL2 format for simplicity in discriminating the similar ones.

Optimization of atomic parameters

The molecular structures in the training set and in SAMPL5 dataset were fully optimized with *ab inito* quantum chemical calculations at B3LYP/6-31G** level of theory to prepare the atomic coordinates required to compute the solvation free energies. All the atomic parameters defined for 41 atom types were then determined with respect to cyclohexane and water to estimate the LogD values of SAMPL5 molecules based on the extended solvent-contact model. Because the atomic fragmental volume (V_i) parameters revealed a bad convergence during the simultaneous optimization with O_i^{max} and S_i values, they were optimized in separate using a standard genetic algorithm as described in the previous papers [35, 36]. Due to the convergence problem, V_i values were allowed to vary among all the atoms even with the same atom type. This criterion was necessary because the volumes of individual atoms depended on the overall structure of a solute molecule.

The optimization of V_i parameters started with calculating the volumes (V_{mol}'s) of all the solute molecules. Each molecule was placed in a 3-D box whose length, width, and height corresponded to the maximum distances along the three axes for the coordinate system of molecular van der Waals volume. To construct the van der Waals volume, atomic radii of carbon, nitrogen, oxygen, sulfur, hydrogen, fluorine, chlorine, and bromine atoms are set to 1.53, 1.45, 1.36, 1.70, 1.08, 1.30, 1.65 and 1.80 respectively. Monte Carlo simulations were then carried out to calculate the V_{mol} value by randomly selecting the grid points in the 3-D box embedding the solute molecule. More specifically, the V_{mol} value was obtained by the product of the box volume (V_{box}) and the ratio of the number of trials to select a point in the molecular van der Waals volume (N_{hits}) to the total number of trials (N_{trials}). All the V_{mol} values of the solute molecules were thus calculated with the following equation.

$$V_{mol} = V_{box} \times \frac{N_{hits}}{N_{trials}} \tag{6}$$

Using the calculated V_{mol} values, the V_i parameters were optimized by operating the standard genetic algorithm. This began with the definition of a generation with 100 vectors comprising the V_i parameters for all the atoms in molecules, which was followed by the removal of 50 with a bias toward preserving the most fit with the lowest error. The empty 50 vectors were then filled with the point mutations to alter the value of one of the parameters with probability 0.01, and with the cross breeds with probability 0.6 to select some parameters from one vector to replace the elements of another vector of the top 50. The 50 new vectors created in this way were then evaluated together with the top 50. This cycle was repeated as many times as desired. To evaluate the 100 vectors, we used the error hypersurface (F_v) defined by the sum of the absolute values of the differences between the V_{mol} value and the sum of V_i values of a solute molecule.

$$F_V = \sum_{k}^{molecules} \left| V_{mol}^{k} - \sum_{i}^{atoms} V_i \right| \tag{7}$$

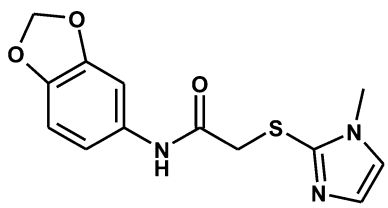
After the parameterizations of V_i, O_i^{max} and S_i values for all the atom types were optimized concurrently using the genetic algorithm to make the solvation free energy function suitable for calculating the ΔG_{sol} values for cyclohexane and water. These second parameterizations began with the construction of a generation consisting of 100 vectors whose elements were O_i^{max} and S_i parameters for all the available atom types. In the second step, 50 of 100 vectors were made empty with a bias toward the best fit with the lowest error. These vacant 50 vectors were filled again with the new elements generated by processing those of the remaining 50 vectors. The new vector elements were obtained by the point mutations with probability 0.01 to alter the O_i^{max} and S_i values as well as by the cross breeds with probability 0.6 to exchange the corresponding elements in the top 50 vectors. The newly generated vectors were then combined with top 50 to be evaluated together. This procedure was repeated until the convergence criterion was met. The evaluation of each vector was carried out using the error hypersurface (F_s) given by summing over the differences between the ΔG_{sol} values of the training set molecules measured from experiment ($\Delta G_{\exp}^{i}$) and those calculated with the solvation free energy function (ΔG_{calc}^{i}). This fitness function can be written in the following form.

$$F_s = \sum_{i}^{molecules} \left| \Delta G_{\exp}^{i} - \Delta G_{calc}^{i} \right| \tag{8}$$

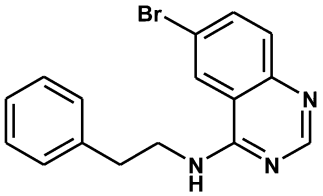
The optimizations tended to converge approximately after 100000 iterations for V_i and 1000 for O_i^{max} and S_i values.

Results and discussion

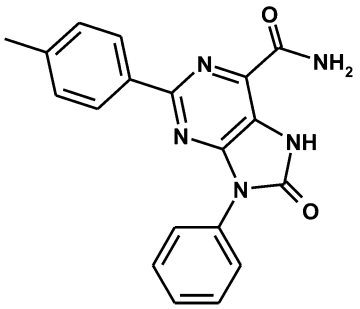
Chemical structures of the selected SAMPL5 molecules are shown in Fig. 2. We note that SAMPL5 dataset has a wide spectrum of shape and size with molecular weights (MWs)



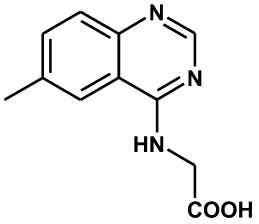
SAMPL5_005



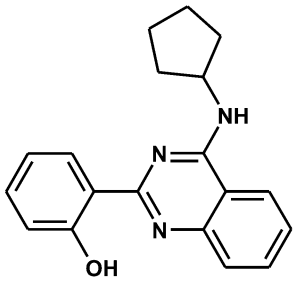
SAMPL5_007



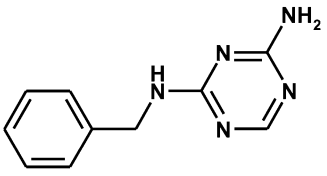
SAMPL5_013



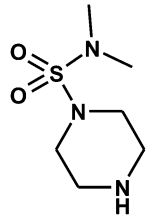
SAMPL5_015



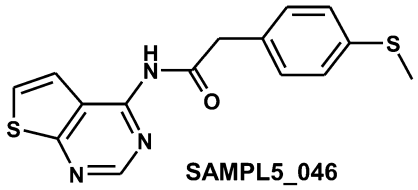
SAMPL5_017



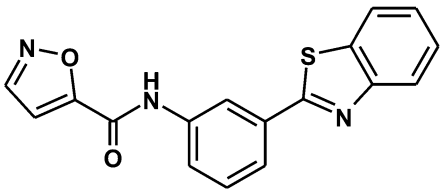
SAMPL5_027



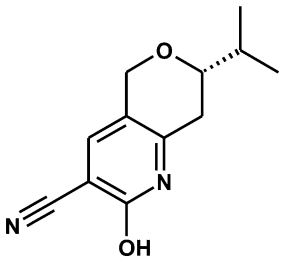
SAMPL5_037



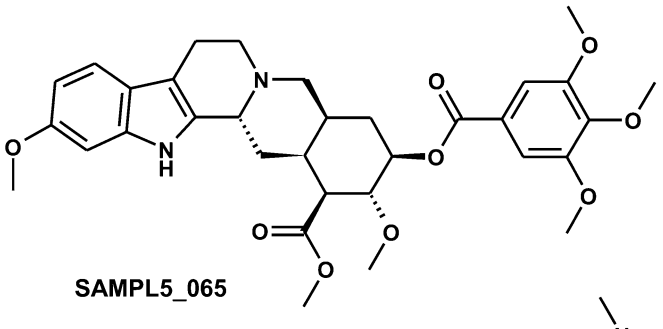
SAMPL5_046



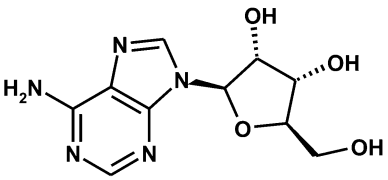
SAMPL5_048



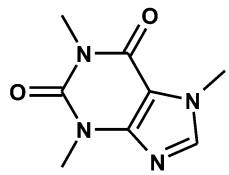
SAMPL5_056



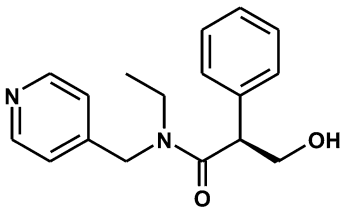
SAMPL5_065



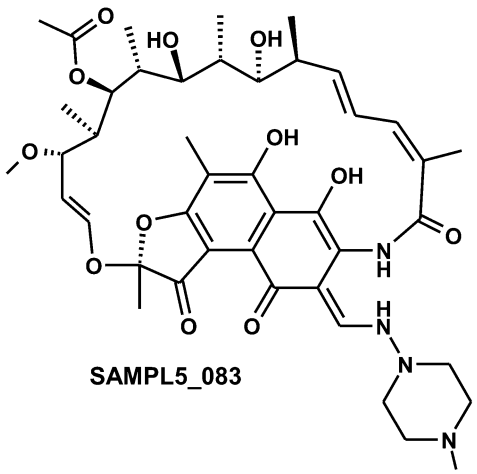
SAMPL5_074



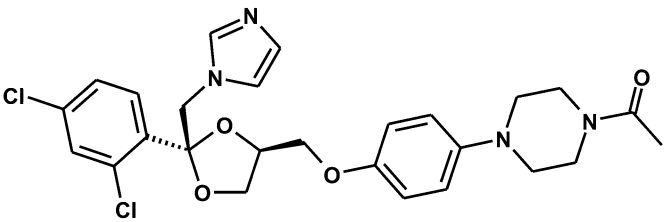
SAMPL5_080



SAMPL5_088



SAMPL5_083



SAMPL5_092

◀**Fig. 2** Chemical structures of the selected molecules included in SAMPL5 dataset. The structures of all SAMPL5 molecules are presented in Supplementary Materials

ranging from 170 to 810 amu in comparison to that of SAMPL4 which included only the small molecules with MW lower than 280 amu. This indicates that more rigorous computational methods would be required in SAMPL5 prediction challenge than those adopted for SAMPL4 molecules to get similar achievements in performance. Nonetheless, the augmentation of new atom types should be minimal in the extended solvent-contact model lest the optimization leads to overtraining due to the excessive atomic parameters.

It should also be noted that several SAMPL5 molecules can exist in different tautomeric forms. Although the accuracy in LogD prediction would be enhanced by considering the structural multiplicity, we used only the major tautomeric form of each SAMPL5 molecule for computational simplicity. For example, the enol form was adopted when the ring system involving the –OH moiety satisfied the aromaticity conditions as in SAMPL5_50, SAMPL5_56, and SAMPL5_83, whereas the keto form was selected in the other cases.

LogD values of SAMPL5 molecules are expected to be similar to LogP ones because they include only weakly or hardly ionizable groups such as carboxylic acid, amine, and phenol moieties, which belong to a weak acid/base with the ionization constant smaller than 10^{-4}. All SAMPL5 molecules were therefore assumed to be neutral in this study to make it straightforward to determine their solvation free energies. Furthermore, the experimental LogD values of all SAMPL5 molecules were measured at the concentrations lower than 0.1 mM. It is difficult to form the solute dimer in such dilute solutions, which would have the effect of further reducing the difference between LogD and LogP values of SAMPL5 molecules. Taken together, LogD values of SAMPL5 molecules may be estimated with the solvation free energy functions optimized for water and cyclohexane.

To calculate the ΔG_{sol} values of each SAMPL5 molecule in water and cyclohexane, all atomic parameters in the solvation free energy function were optimized with respect to both solvents using the experimental data for 92 training set molecules. Table 1 lists the optimized O_i^{max} and S_i values for 41 atom types introduced to describe all the atoms in SAMPL5 molecules under a variety of chemical circumstances. V_i parameters for all the atoms in SAMPL5 molecules are presented in Supplementary Materials. They have to be presented in separate from O_i^{max} and S_i parameters because they can vary among the atoms with the same atom type. Despite the complexity in parametrizations, the O_i^{max} and S_i values tend to vary in a manner consistent with general atomic properties. For example, the O_i^{max} parameters of the second-period atoms (C, N, O, and F) range from 270 to 400 irrespective of the solvent as compared to those of hydrogen atoms smaller than 250. In comparison, most O_i^{max} values of sulfur and bromine atoms exceed 400. This trend can be understood in terms of the conceptual similarity of the O_i^{max} parameter to the atomic volume.

The S_i parameters appear to vary significantly with the atom types even among the same elements whereas the O_i^{max} values for the atoms in the same period are relatively similar. For instance, the S_i value of carbonyl carbon (C.CO_2) with respect to water is even more negative than those of alkyl and aromatic carbons. This may be related with the accumulation of positive charge on the carbonyl carbon, which stems from the electron withdrawal by the adjacent carbonyl oxygen. As can be expected from the large differences in physicochemical properties between the two solvents, the S_i parameters for water are quite different from those for cyclohexane. We note in this regard that all the S_i values corresponding to carbon atoms converge to the negative values in cyclohexane as a consequence of reflecting the attractive van der Waals interactions between solute carbon atoms and the solvent. On the other hand, the S_i values of carbon atoms become positive or less negative in water due to the weakening of hydrophobic interactions with solvent.

Consistent with the major contributions of nitrogen and oxygen atoms to molecular solubility in polar solvents, their S_i values for most atom types are optimized to be highly negative in water. Actually, the interactions of polar solute atoms with water molecules are expected to be attractive because they can be stabilized in aqueous solution not only by the long-range electrostatic interactions with bulk solvent but also by the local hydrogen bonds with solvent molecules. However, most S_i values of nitrogen and oxygen atoms become much less negative or positive with the change of solvent from water to cyclohexane. This can be understood in the context of the weakening of solute-solvent interactions due to the lack of polarity in solvent molecules. In case of hydrogen atoms, the S_i parameters tend to be more negative as the adjacent atom changes from carbon to polar atoms, which can also be attributed to the facilitation of the electrostatic interactions with solvent.

Using the V_i, O_i^{max}, and S_i parameters of all 41 atom types optimized with 92 training set molecules, the solvation free energies of SAMPL5 molecules were calculated for water and cyclohexane to produce the ultimate LogD values. The correlation diagram of experimental versus calculated ΔG_{sol} values of 92 molecules in the training set and that of experimental versus calculated LogD values of 53 SAMPL5 molecules are shown in Figs. 3 and 4, respectively. All ΔG_{sol} values for the molecules in the

Table 1 Optimized O_i^{max} and S_i parameters for 41 atom types in SAMPL5 molecules

Atom type	Description	$O_{i,max}$ (Å^3)		S_i (kcal/mol Å^3)	
		c-Hexane	Water	c-Hexane	Water
C.3_1	sp^3 carbon with 1 substituent	320.6	330.0	−2.857	0.492
C.3_2	sp^3 carbon with 2 substituents	318.3	342.1	−3.492	0.651
C.3_3	sp^3 carbon with 3 substituents	328.6	341.3	−4.127	0.397
C.3_4	sp^3 carbon with 4 substituents	303.2	304.0	−2.937	−0.079
C.2_1*	sp^2 carbon with 1 substituent	317.5	308.7	−0.667	−0.397
C.2_2*	sp^2 carbon with 2 substituents	307.1	334.1	−3.254	0.048
C.2_3*	sp^2 carbon with 3 substituents	315.1	346.0	−2.460	−1.651
C.1_2	*sp* carbon with 2 substituents	323.8	281.1	−4.444	0.651
C.ar_2	Aromatic carbon with 2 substituents	323.8	323.2	−2.778	−0.667
C.ar_3	Aromatic carbon with 3 substituents	334.1	304.0	−3.889	−2.000
C.CO_2	Carbonyl carbon with 2 substituents	306.3	325.4	−7.063	−8.413
N.1_1	*sp* nitrogen with 1 substituent	336.5	326.2	−0.079	−10.238
N.3_2	sp^3 nitrogen with 2 substituents	354.3	303.2	−8.095	−11.191
N.3_3	sp^3 nitrogen with 3 substituents	318.3	301.1	−4.048	−13.730
N.ar	Aromatic nitrogen	271.4	344.2	−5.079	−9.905
N.pl_1	Planar nitrogen with 1 substituent	398.4	315.9	3.143	−15.000
N.pl_2	Planar nitrogen with 2 substituents	315.9	340.5	−1.667	−13.095
N.pl_3*	Planar nitrogen with 3 substituents	396.8	320.6	−13.683	−4.349
N.am_1	Amide nitrogen with 1 substituent	339.7	300.0	2.079	−8.810
N.am_2*	Amide nitrogen with 2 substituents	360.3	321.4	−3.571	−0.476
N.am_3*	Amide nitrogen with 3 substituents	327.0	342.9	−12.429	−14.048
O.3_1	sp^3 oxygen with 1 substituent	270.0	315.9	0.556	−14.603
O.3_2	sp^3 oxygen with 2 substituents	350.8	307.9	−1.730	−11.825
O.pl_1	Planar oxygen with 1 substituent	285.9	312.7	−1.190	−15.873
O.pl_2	Planar oxygen with 2 substituents	387.3	297.8	−1.508	−4.841
O.es_1	sp^3 oxygen in carboxylic acids	382.5	300.8	1.032	−6.349
O.es_2	sp^3 oxygen in esters	393.7	336.5	−0.333	2.619
O.2	sp^2 oxygen	330.2	331.7	0.333	−6.429
S.12*	Sulfur with 12 valence electrons	441.7	431.0	−9.048	−14.921
S.3_2*	sp^3 sulfur with 2 substituents	439.7	400.8	−3.714	−5.159
S.pl	Planar sulfur	379.4	397.5	−2.302	−1.111
F	Fluorine	325.4	301.0	1.349	1.746
Cl	Chlorine	361.3	402.9	−2.937	0.317
Br	Bromine	484.1	465.7	−0.238	0.476
H.C	Hydrogen bonded to carbon	158.6	157.9	0.714	−0.159
H.N3	Hydrogen bonded to sp^3 nitrogen	164.3	238.7	0.079	−11.191
H.Np	Hydrogen bonded to planar nitrogen	221.1	228.9	−6.571	−0.540
H.Na	Hydrogen bonded to amidic nitrogen	246.8	206.5	−9.429	−11.619
H.O3	Hydrogen bonded to planar oxygen	243.7	236.5	−3.095	−6.905
H.Op	Hydrogen bonded to sp^3 oxygen	234.1	236.2	−3.810	−1.698
H.Oa	Hydrogen bonded to carboxylic acid group	183.3	168.3	0.079	−6.508

Asterisk indicates the atom type missing in the molecules with ΔG_{sol} values for cyclohexane

training set and SAMP5 dataset are provided in Supplementary Material. Although a good correlation is observed between experimental and computational solvation free energies of training set molecules with the correlation constant (R) larger than 0.96, the prediction accuracy decreases significantly in the estimation of LogD data for the SAMPL5 molecules with the associated R value of 0.55. The average error (AE) and root mean square error

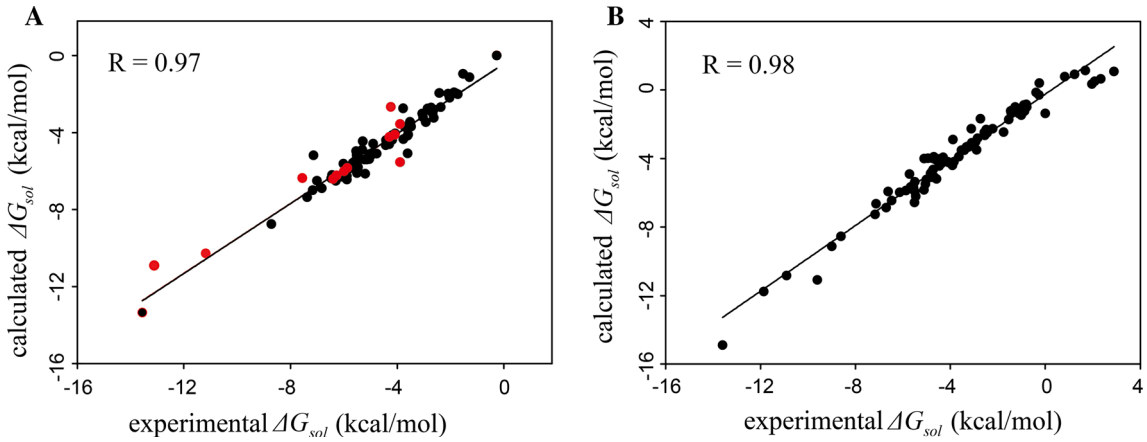


Fig. 3 Correlation diagrams for the experimental versus calculated solvation free energies of 92 molecules in the training set with respect to **A** cyclohexane and **B** water. Indicated in *red circles* are the training set molecules for which the experimental ΔG_{sol} values in 1-octanol were referenced

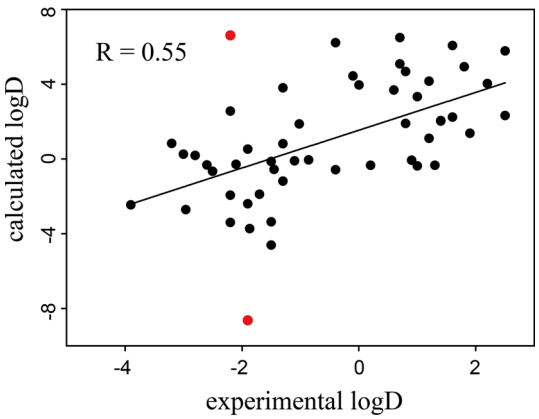


Fig. 4 Correlation diagram between the experimental and calculated LogD values of 53 SAMPL5 molecules whose atomic parameters are optimized with 92 training set molecules. The *upper* and the *lower red circles* indicate SAMPL5_80 and SAMPL5_74, respectively, which reveal a large deviation between the experimental and calculated LogD values

(RMSE) amount to 1.53 and 3.03, respectively, which rank 33th and 17th among 62 participants in SAMPL5 prediction challenge for the unitless LogD values.

With respect to the modest accuracy in LogD prediction, we note that the training set could not be constituted completely with the molecules for which the experimental ΔG_{sol} data were available for both water and cyclohexane because they lacked some atom types present in SAMPL5 molecules. For example, the atom types for sp^2 carbon (C.2_1, C.2_2, and C.2_3) were missing in the molecules with the experimental ΔG_{sol} values for cyclohexane. Therefore, the roles of the elements of training set possessing the sp^2 carbon had to be played by the molecules with the experimental ΔG_{sol} values for 1-octanol. The same was true of the training set molecules containing the atom types of N.pl_3, N.am_2, N.am_3, S.12, and S.3_2. These vicarious selections of the training set molecules can lead to the incomplete optimization of the solvation free energy function for cyclohexane, which would culminate in the impairment of accuracy in LogD predictions. Indeed, the largest differences between the experimental and calculated LogD values are observed in SAMPL5_074 and SAMPL5_080 as indicated in Fig. 4, both of which contain at least two atom types missing in the molecules with the experimental ΔG_{sol} data for cyclohexane.

We now turn to the second prediction challenge for LogD values with the subset of SAMPL5 molecules for which all the atomic parameters can be fully optimized with the reference ΔG_{sol} data for cyclohexane. By comparing the new results with those for all SAMPL5 molecules, it would be possible to address the influence of replacing the ΔG_{sol} values for cyclohexane with those for 1-octanol on the accuracy in LogD prediction. This comparative analysis started with the reoptimization of solvation free energy functions using only the training set molecules for which experimental ΔG_{sol} data were available for cyclohexane. Accordingly, we excluded some SAMPL5 molecules containing the atom types missing in the new training set. As a consequence, 77 and 31 molecules remained in the training set and SAMPL5 test set, respectively, along with the decrease in the number of atom types from 41 to 33.

Table 2 lists the newly optimized atomic parameters using the modified training set with the same procedure as described in the previous section. The S_i parameters for water and the O_i^{max} values for both solvents remain strongly correlated with those in the parameterizations with the full training set (Table 1). The R values associated with the comparisons of the new and previous S_i (water), O_i^{max} (cyclohexane), and O_i^{max} (water) parameters amount to

Table 2 Optimized O_i^{max} and S_i parameters for 33 atom types in SAMPL5 molecules for which the experimental ΔG_{sol} values are available for cyclohexane

Atom type	Description	$O_{i,max}$ (Å^3)		S_i (kcal/mol Å^3)	
		Water	1-Octanol	Water	1-Octanol
C.3_1	sp^3 carbon with 1 substituent	366.5	350.0	−1.587	1.778
C.3_2	sp^3 carbon with 2 substituents	382.4	302.9	−1.905	1.746
C.3_3	sp^3 carbon with 3 substituents (1)	303.8	293.8	−0.714	0.984
C.3_4	sp^3 carbon with 4 substituents	319.7	339.7	−2.302	0.190
C.1_2	sp carbon with 2 substituents	325.4	330.2	−2.143	1.048
C.ar_2	Aromatic carbon with 2 substituents	315.6	348.9	−2.619	−0.206
C.ar_3	Aromatic carbon with 3 substituents	341.3	334.9	−2.190	−0.968
C.CO_2	Carbonyl carbon with 1 substituents	336.5	350.0	−1.476	−4.921
N.1_1	sp nitrogen with 1 substituent	342.1	302.4	−1.667	−12.937
N.3_2	sp^3 nitrogen with 2 substituents	315.6	344.4	−2.460	−6.905
N.3_3	sp^3 nitrogen with 3 substituents (1)	309.5	314.4	−0.873	−14.191
N.ar	Aromatic nitrogen	324.8	375.2	−3.413	−5.000
N.pl_1	Planar nitrogen with 1 substituent	314.3	309.5	−12.302	−15.000
N.pl_2	Planar nitrogen with 2 substituents (2)	366.7	346.0	−5.079	−9.921
N.am_1	Amide nitrogen with 1 substituent (1)	385.7	322.2	−3.413	−7.143
O.3_1	sp^3 oxygen with 1 substituent	279.5	337.6	0.111	−6.984
O.3_2	sp^3 oxygen with 2 substituents	331.7	342.1	−1.286	−9.286
O.pl_1	Planar oxygen with 1 substituent	341.4	301.1	−5.000	−15.079
O.pl_2	Planar oxygen with 2 substituents (2)	331.7	378.9	−2.048	−3.492
O.es_1	sp^3 oxygen in carboxylic acids (2)	295.7	315.1	−1.778	−5.587
O.es_2	sp^3 oxygen in esters	307.9	313.5	−1.222	1.349
O.2	sp^2 oxygen	396.8	349.0	−2.492	−6.667
S.pl	Planar sulfur	354.8	393.3	−3.730	−5.397
F	Fluorine	328.9	306.7	0.810	1.571
Cl	Chlorine	340.6	383.8	−2.619	0.651
Br	Bromine	520.0	445.7	−2.000	−1.556
H.C	Hydrogen bonded to carbon	185.6	183.3	−0.476	−1.333
H.N3	Hydrogen bonded to sp^3 nitrogen	220.6	193.3	0.571	−16.349
H.Np	Hydrogen bonded to planar nitrogen	260.0	186.3	3.524	−1.937
H.Na	Hydrogen bonded to amidic nitrogen	243.3	198.6	−5.524	−16.571
H.O3	Hydrogen bonded to sp^3 oxygen	231.0	231.7	−3.730	−15.524
H.Op	Hydrogen bonded to planar oxygen	223.8	216.7	2.810	−4.095
H.Oa	Hydrogen bonded to carboxylic acid group	240.0	222.1	1.524	−7.667

Numbers in parenthesis indicate the number of occurrences in the training set

0.86, 0.79, and 0.88, respectively. Some general tendencies are therefore also found in the newly optimized O_i^{max} and S_i parameters of varying atom types. For example, the S_i values of most carbon atoms are negative and positive in cyclohexane and in water, respectively. The electronegative nitrogen and oxygen atoms have highly negative S_i values with respect to water, which is consistent with their major contributions to the stabilization of a parent molecule in aqueous solution. However, the newly optimized S_i values with respect to cyclohexane appear to become quite different from those obtained with the full dataset with the associated R value of 0.13. This indicates that the calculated ΔG_{sol} values of SAMPL5 molecules for cyclohexane can vary significantly due to the removal of the training set molecules lacking the reference ΔG_{sol} data for cyclohexane, which would in turn have the effect of changing the results of LogD prediction in a large part.

The correlation diagrams of experimental versus calculated ΔG_{sol} values of 77 training set molecules with 33 atoms types are shown in Fig. 5. Both R values for cyclohexane and water remain similar to those obtained with the original training set comprising 92 molecules and 41 atom types (Fig. 3). As shown in Fig. 6, however, the accuracy in LogD prediction appears to be improved remarkably with the increase of R value from 0.55 to 0.82. We note that the R value becomes close to the top-ranked

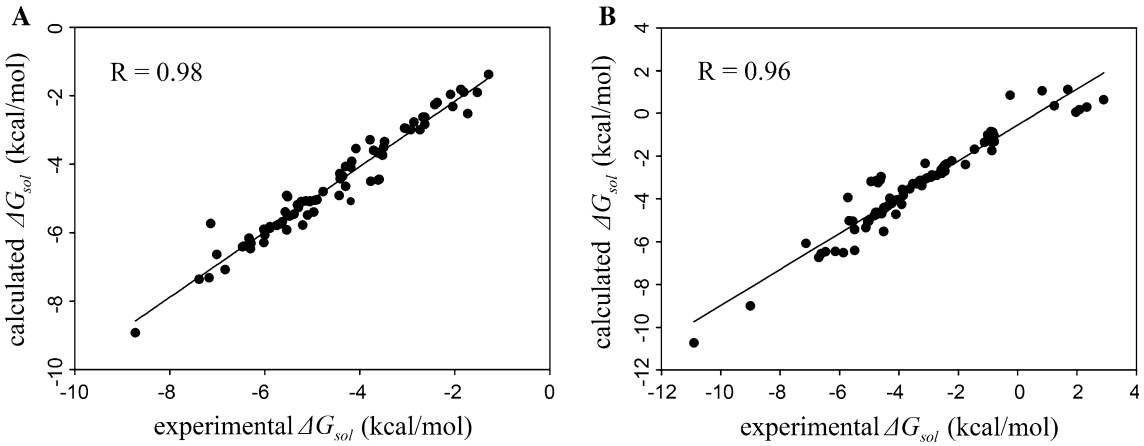


Fig. 5 Correlation diagrams for the experimental versus calculated solvation free energies with respect to **A** cyclohexane and **B** water for 77 training set molecules for which experimental ΔG_{sol} values are available for both solvents

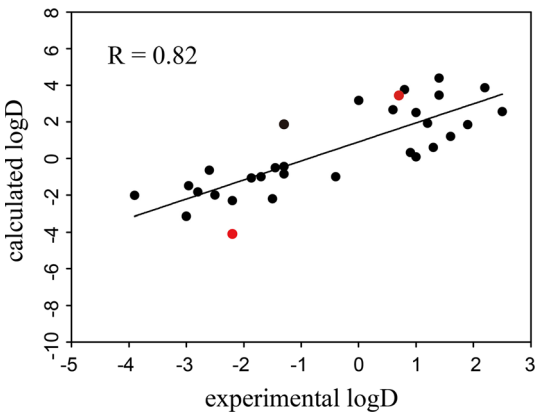


Fig. 6 Correlation diagrams between the experimental and calculated LogD values of 31 SAMPL5 molecules for which all atomic parameters can be optimized using 77 training set molecules with experimental ΔG_{sol} data for cyclohexane. The *upper* and the *lower red circles* indicate SAMPL5_065 and SAMPL5_081, respectively, which reveal a large deviation between the experimental and calculated LogD values

one (0.84) in SAMPL5 blind prediction challenge for LogD. This accuracy enhancement is apparently attributed to the exclusion of the training set molecules without the reference ΔG_{sol} values for cyclohexane. In particular, the S_i parameters for cyclohexane seem to be optimized better than before by limiting the elements of training set to the molecules for which the experimental ΔG_{sol} data are available, which leads to the better prediction of ΔG_{sol} values for cyclohexane and culminates in the accuracy enhancement in LogD predictions. This result exemplifies the importance of constructing a proper training set for the extended solvent-contact model to be useful for predicting the physicochemical properties of drug-like molecules.

Listed in Table 3 are the LogD values of SAMPL5 molecules calculated with and without the ΔG_{sol} data for 1-octanol in the training set in comparison with the corresponding experimental results. Consistent with the increase in R value, both AE and RMSE decrease from 1.53 and 3.03 to 0.89 and 1.60, respectively, due to the modification of the training set. It is remarkable to note that the RMSE value becomes lower than that of the best scored one in the SAMPL5 blind prediction challenge for LogD. Although it makes little sense to compare our new computational results with those obtained with the full SAMPL5 dataset, it can at least be argued that the extended solvent-contact model would be one of the most efficient methods for LogD prediction upon the availability of sufficient experimental ΔG_{sol} data for cyclohexane.

With respect to the improvement of the accuracy in LogD prediction, the S_i parameters of the planar nitrogens bonded to aromatic rings appear to change most significantly in the optimizations with the new training set. For example, the S_i values of N.pl_1 and N.pl_2 for cyclohexane decrease from 3.143 and −1.667 (Table 1) to −12.302 and −5.079 (Table 2), respectively, due to the exclusion of the molecules lacking the reference ΔG_{sol} data for cyclohexane in the training set. The highly negative S_i values of planar nitrogens can be understood in the context that their hydrophobic interactions with cyclohexane molecules would be facilitated along with the delocalization of lone-pair electrons into the adjacent aromatic ring, which has the effect of decreasing the polarity on the nitrogens. In this regard, the S_i values seem to be more negative in cyclohexane than in 1-octanol because the former is more hydrophobic than the latter. Because N.pl_1 and N.pl_2 are the most abundant heteroatoms in the SAMPL5 dataset, a significant enhancement in LogD prediction is anticipated by the better optimization of their S_i values with respect to cyclohexane. Indeed, the deviations between the experimental and calculated LogD values of

Table 3 Comparison of experimental and calculated LogD values of SAMPL5 molecules

Compound ID	$LogD_{exp}$	$LogD_{calc}$	
		Training with cyclohexane and 1-octanol data	Training with cyclohexane data only
SAMPL5_002	1.40	2.02	NA
SAMPL5_003	1.90	1.36	1.86
SAMPL5_004	2.20	4.02	3.87
SAMPL5_005	−0.86	−0.06	NA
SAMPL5_006	−1.02	1.87	NA
SAMPL5_007	1.40	2.04	3.46
SAMPL5_010	−1.70	−1.89	−1.00
SAMPL5_011	−2.96	−2.72	−1.49
SAMPL5_013	−1.50	−0.15	NA
SAMPL5_015	−2.20	−3.40	−2.29
SAMPL5_017	2.50	2.32	2.56
SAMPL5_019	1.20	1.09	1.93
SAMPL5_020	1.60	2.22	1.20
SAMPL5_021	1.20	4.15	NA
SAMPL5_024	1.00	3.33	2.52
SAMPL5_026	−2.60	−0.32	−0.64
SAMPL5_027	−1.87	−3.74	−1.06
SAMPL5_033	1.80	4.92	NA
SAMPL5_037	−1.50	−4.61	NA
SAMPL5_042	−1.10	−0.11	NA
SAMPL5_044	1.00	−0.37	0.09
SAMPL5_045	−2.10	−0.29	NA
SAMPL5_046	0.20	−0.35	NA
SAMPL5_047	−0.40	−0.58	−0.99
SAMPL5_048	0.90	−0.07	0.33
SAMPL5_049	1.30	−0.34	0.60
SAMPL5_050	−3.20	0.82	NA
SAMPL5_055	−1.50	−3.37	−2.20
SAMPL5_056	−2.50	−0.67	−2.00
SAMPL5_058	0.80	4.67	NA
SAMPL5_059	−1.30	−1.20	−0.44
SAMPL5_060	−3.90	−2.46	−2.01
SAMPL5_061	−1.45	−0.57	−0.50
SAMPL5_063	−3.00	0.25	−3.14
SAMPL5_065	0.70	6.48	3.44
SAMPL5_067	−1.30	3.80	1.86
SAMPL5_068	1.40	2.02	4.40
SAMPL5_069	−1.30	0.81	−0.85
SAMPL5_070	1.60	6.05	NA
SAMPL5_071	−0.10	4.43	NA
SAMPL5_072	0.60	3.680	2.66
SAMPL5_074	−1.90	−8.64	NA
SAMPL5_075	−2.80	0.17	−1.83
SAMPL5_080	−2.20	6.61	NA
SAMPL5_081	−2.20	−1.95	−4.11
SAMPL5_082	2.50	5.77	NA

Table 3 continued

Compound ID	$LogD_{exp}$	$LogD_{calc}$	
		Training with cyclohexane and 1-octanol data	Training with cyclohexane data only
SAMPL5_083	−1.90	−2.40	NA
SAMPL5_084	0.00	3.95	3.17
SAMPL5_085	−2.20	2.55	NA
SAMPL5_086	0.70	5.08	NA
SAMPL5_088	−1.90	0.52	NA
SAMPL5_090	0.80	1.89	3.77
SAMPL5_092	−0.40	6.21	NA
AE		1.53	0.89
RMSE		3.03	1.60

SAMPL5_015, SAMPL5_027, and SAMPL5_065 including the planar nitrogens appear to decrease significantly from 1.20, 1.87, and 5.78 to 0.09, 0.81 and 2.74 (Table 3), respectively, along with the modification of the training set.

Despite the considerable accuracy enhancement in LogD prediction by modifying the training set, a large discrepancy between experimental and computational results is still observable for some molecule such as SAMPL5_065 and SAMPL5_081 as indicated in Fig. 6. We note in this regard that several atoms types (C.3_3, N.3_3, N.pl_2, N.am_1, O.pl_2, and O.es_1) appear only once or twice in the training set due to the rarity of experimental ΔG_{sol} data for cyclohexane. Therefore, it seems to be difficult for the atomic parameters to be fully optimized in such a way to reflect various chemical circumstances around the atoms in molecules during the parameterizations. The low occurrences of the six atom types in the training set are likely to serve as one of the major error sources in LogD prediction because they are present in a number of SAMPL5 molecules.

It is thus found to be a drawback of the extended solvent-contact model to require a sufficient amount of experimental data for the optimization of solvation free energy function. However, this requirement seems not to be severe because LogD values of 31 SAMPL5 molecules were estimated with reasonable accuracy using only 77 training set molecules. The characteristic feature that discriminates the extended solvent-contact model from the other computational methods lies in that one can calculate the molecular LogD values straightforwardly with the solvation free energy functions and the atomic coordinates of solute molecules. This is in contrast with quantitative structure-property relationship (QSPR), quantum mechanical, and statistical simulation methods that require a high computational cost for calculating the molecular descriptors, the electronic structures, and the trajectories in configurational space, respectively. Because of the simplicity in model building and little computational burden for parameterizations, the extended solvent contact model is expected to serve as one of the most efficient computational methods for LogD prediction upon the enrichment of experimental ΔG_{sol} data for organic solvents.

With respect to the accuracy enhancement in LogD prediction, it also noteworthy that the solvation free energy function in Eq. (5) lacks the entropic term. Although the determination of molecular solvation entropy had been considered very difficult for a long time, it proved recently to be estimated with accuracy by means of combining the free energy perturbation method and the scaled particle theory to calculate the electrostatic and hydrophobic contributions of solvent-solute interactions, respectively [42]. Because both enthalpic and entropic contributions to the solvation free energy are experimentally measurable, the potential parameters in the two terms can be optimized independently using the corresponding reference data. This dual parameterization would warrant the better prediction of solvation free energies than the single parameterization because more diverse experimental data can be referenced. Our future studies will focus on the improvement of LogD prediction accuracy through the modification of solvation free energy function by implementing the solvation entropy term.

Conclusions

We addressed the applicability of the extended solvent-contact model to the calculation of molecular LogD values through the participation in SAMPL5 blind prediction challenge. After defining the atomic parameters for 41 atom types to describe a total of 53 SAMPL5 molecules, the solvation free energy function was optimized with respect to water and cyclohexane using 92 training set molecules to obtain the ΔG_{sol} values required to calculate

LogD. Due to the deficiency of experimental data for cyclohexane, the reference ΔG_{sol} values of 15 training set molecules were replaced with those for 1-octanol. The LogD values of SAMPL5 molecules were predicted with modest accuracy with the R, AE, and RMSE values of 0.55, 1.53, and 3.03, respectively, for the comparison of experimental and computational results. The incomplete optimization of the atomic S_i parameters with respect to cyclohexane proved to be the major source of error in LogD prediction. The R, AE, and RMSE values could be improved remarkably to 0.82, 0.89, and 1.60, respectively, when the predictions were made for 31 SAMPL5 molecules containing the atom types for which the experimental reference ΔG_{sol} data were available for cyclohexane. This considerable enhancement in performance stemmed from the better parameterization of S_i values by limiting the element of training set to the molecules with experimental ΔG_{sol} data for cyclohexane. Most significant improvements in LogD prediction were observed for the SAMPL5 molecules including the planar nitrogens whose attractive van der Waals interactions with cyclohexane could be described appropriately only with the S_i values optimized by using the modified training set. Judging from the simplicity in model building and from the low computational cost for parametrizations, the extended solvent-contact model is anticipated to serve as a valuable computational tool for LogD prediction upon the enrichment of experimental ΔG_{sol} data for cyclohexane.

Acknowledgments This research was supported by Creative Materials Discovery Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT and Future Planning (2015M3D1A1069705), and by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (NRF-2016R1D1A1B01014187).

References

1. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 23:3–25
2. Hermens JLM, de Gruijin JHM, Brooke DN (2013) The octanol–water partition coefficient: strengths and limitations. Environ Toxicol Chem 32:732–733
3. van de Waterbeemd H, Camenisch G, Folkers G, Chretien JR, Raevsky OA (1998) Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors. J Drug Target 6:151–165
4. van de Waterbeemd H, Gifford E (2003) ADMET in silico modelling: towards prediction paradise? Nat Rev Drug Discov 2:192–204
5. Dearden JC (2007) In silico prediction of ADMET properties: how far have we come? Expert Opin Drug Metab Toxicol 3:635–639
6. Kenny JR (2013) Predictive DMPK: in silico ADME predictions in drug discovery. Mol Pharm 10:1151–1152
7. Smith DA (2013) Evolution of ADME science: where else can modeling and simulation contribute? Mol Pharm 10:1162–1170
8. Roda A, Minutello A, Angellotti MA, Fini A (1990) Bile acid structure-activity relationship: evaluation of bile acid lipophilicity using 1-octanol/water partition coefficient and reverse phase HPLC. J Lipid Res 31:1433–1443
9. Leung SSF, Sindhikara D, Jacobson MP (2016) Simple predictive models of passive membrane permeability incorporating size-dependent membrane-water partition. J Chem Inf Model 56:924–929
10. Jing P, Rodgers PJ, Amemiya S (2009) High lipophilicity of perfluoroalkyl carboxylate and sulfonate: implications for their membrane permeability. J Am Chem Soc 131:2290–2296
11. Schneider N, Lange G, Hindle S, Klein R, Rarey MA (2013) A consistent description of hydrogen bond and dehydration energies in protein-ligand complexes: methods behind the HYDE scoring function. J Comput Aided Mol Des 27:15–29
12. Ghose AK, Crippen GM (1987) Atomic physicochemical parameters for three-dimensional-structuredirected quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. J Chem Inf Comput Sci 27:21–35
13. Rekker RF, Kort HMD (1979) Hydrophobic fragmental constant—extension to a 1000 data point set. Eur J Med Chem 14:479–488
14. Yaffe D, Cohen Y, Espinosa G, Arenas A, Giralt F (2002) Fuzzy ARTMAP and back-propagation neural networks based quantitative structure-property relationships (QSPRs) for octanol-water partition coefficient of organic compounds. J Chem Inf Comput Sci 42:162–183
15. Hou TJ, Xu XJ (2003) ADME evaluation in drug discovery. 2. Prediction of partition coefficient by atom-additive approach based on atom-weighted solvent accessible surface areas. J Chem Inf Comput Sci 43:1058–1067
16. Wegner JK, Zell A (2003) Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. J Chem Inf Comput Sci 43:1077–1084
17. Souza ES, Zaramello L, Kuhnen CA, Junkes BS, Yunes RA, Heinzen VEF (2011) Estimating the octanol/water partition coefficient for aliphatic organic compounds using semi-empirical electrotopological index. Int J Mol Sci 12:7250–7264
18. Toropov AA, Toropova AP, Raska I, Benfenati E (2010) QSPR modeling of octanol/water partition coefficient of antineoplastic agents by balance of correlations. Eur J Med Chem 45:1639–1647
19. Daina A, Michielin O, Zoete V (2014) iLOGP: a simple, robust, and efficient description of n-octanol/water partition coefficient for drug design using the GB/SA approach. J Chem Inf Model 54:3284–3301
20. Kim T, Park H (2015) Computational prediction of octanol–water partition coefficient based on the extended solvent-contact model. J Mol Graph Model 60:108–117
21. Huang W, Blinov N, Kovalenko A (2015) Octanol−water partition coefficient from 3D-RISM-KH molecular theory of solvation with partial molar volume correction. J Phys Chem B 119:5588–5597
22. Banks WA, Kastin A (1985) Peptides and the blood-brain barrier: lipophilicity as a predictor of permeability. Brain Res Bull 15:287–292
23. Kellogg GE, Burnett JC, Abraham DJ (2001) Very empirical treatment of solvation and entropy: a force field derived from LogPo/w. J Comput-Aided Mol Des 15:381–393
24. Delaney JS (2005) Predicting aqueous solubility from structure. Drug Discov Today 10:289–295
25. Mayer PT, Anderson BD (2002) Transport across 1, 9-decadiene precisely mimics the chemical selectivity of the barrier domain in egg lecithin bilayers. J Pharm Sci 91:640–646

26. Toulmin A, Wood JM, Kenny PW (2008) Toward prediction of alkane/water partition coefficients. J Med Chem 51:3720–3730
27. Young RJ, Green DVS, Luscombe CN, Hill AP (2011) Getting physical in drug discovery II: the impact of chromatographic hydrophobicity measurements and aromaticity. Drug Discov Today 16:822–830
28. Abraham MH, Chadha HS, Whiting GS, Mitchell RC (1994) Hydrogen bonding. 32. An analysis of water-octanol and water-alkane partitioning and the & Δlog P parameter of Seiler. J Pharm Sci 83:1085–1100
29. Saunders RA, Platts JA (2004) Scaled polar surface area descriptors: development and application to three sets of partition coefficients. New J Chem 28:166–172
30. Zerara M, Brickmann J, Kretschmer R, Exner TE (2008) Parameterization of an empirical model for the prediction of n-octanol, alkane and cyclohexane/water as well as brain/blood partition coefficients. J Comput-Aided Mol Des 23:105–111
31. Kenny PW, Montanari CA, Prokopczyk IM (2013) $ClogP_{alk}$: a method for predicting alkane/water partition coefficient. J Comput Aided Mol Des 27:389–402
32. Lamarche O, Platts JA, Hersey A (2004) Theoretical prediction of partition coefficients via molecular electrostatic and electronic properties. J Chem Inf Comput Sci 44:848–855
33. Caron G, Ermondi G (2005) Calculating virtual log P in the alkane/water system $logP_{alk}^{N}$ and its derived parameters Δlog $P_{oct\_alk}^{N}$ and log D_{alk}^{pH}. J Med Chem 48:3269–3279
34. Leung SSF, Sindhikara D, Jacobson MP (2016) Simple predictive models of passive membrane permeability incorporating size-dependent membrane-water partition. J Chem Inf Model 56:924–929
35. Park H (2014) Extended solvent-contact model approach to SAMPL4 blind prediction challenge for hydration free energies. J Comput-Aided Mol Des 28:175–186
36. Chung KC, Park H (2015) Accuracy enhancement in the estimation of molecular hydration free energies by implementing the intramolecular hydrogen bond effects. J Cheminform 7:57
37. Mączyński A, Wiśniewska-Gocłowska B, Góral M (2004) Recommended liquid–liquid equilibrium data. Part 1. Binary alkane–water systems. J Phys Chem Ref Data 33:549–577
38. Escher BI, Schwarzenbach RP, Westall JC (2000) Evaluation of liposome−water partitioning of organic acids and bases. 1. Development of a sorption model. Environ Sci Technol 34:3954–3961
39. Marenich AV, Kelly CP, Thompson JD, Hawkins GD, Chambers CC, Giesen DJ, Winget P, Cramer CJ, Truhlar DG (2012) Minnesota solvation database—version 2012. University of Minnesota, Minneapolis
40. Wang J, Wang W, Huo S, Lee M, Kollman PA (2001) Solvation model based on weighted solvent accessible surface area. J Phys Chem B 105:5055–5067
41. Lee S, Cho KH, Lee CJ, Kim GE, Na CH, In Y, No KT (2011) Calculation of the solvation free energy of neutral and ionic molecules in diverse solvents. J Chem Inf Model 51:105–114
42. Choi H, Kang H, Park H (2015) Computational prediction of molecular hydration entropy with hybrid scaled particle theory and free-energy perturbation method. J Chem Theory Comput 11:4933–4942