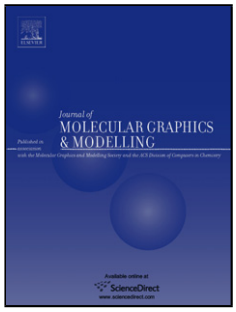


Title: Computational Prediction of Octanol-Water Partition Coefficient Based on the Extended Solvent-Contact Model

Author: Taeho Kim Hwangseo Park

PII: S1093-3263(15)30010-3
DOI: http://dx.doi.org/doi:10.1016/j.jmgm.2015.06.004
Reference: JMG 6558

To appear in: *Journal of Molecular Graphics and Modelling*

Received date: 10-3-2015
Revised date: 9-6-2015
Accepted date: 10-6-2015

Please cite this article as: Taeho Kim, Hwangseo Park, Computational Prediction of Octanol-Water Partition Coefficient Based on the Extended Solvent-Contact Model, Journal of Molecular Graphics and Modelling http://dx.doi.org/10.1016/j.jmgm.2015.06.004

Computational Prediction of Octanol-Water Partition Coefficient Based on the Extended Solvent-Contact Model

Taeho Kim and Hwangseo Park*

Department of Bioscience and Biotechnology, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 143-747, Korea

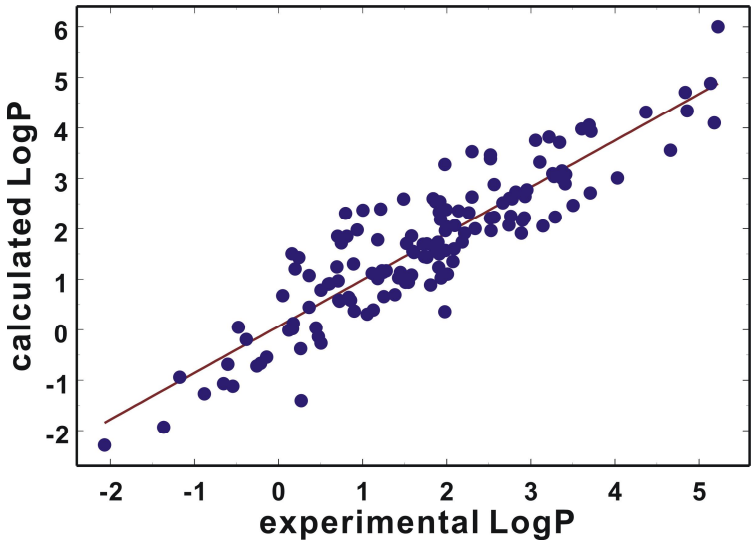
* Author to whom correspondences should be addressed.

Phone: +82-2-3408-3766

FAX: +82-2-3408-4334

E-Mail: hspark@sejong.ac.kr

Graphical Abstract



We have developed and evaluated a method for predicting the molecular LogP values based on the extended solvent-contact model.

Highlights

- We proposed a method for predicting the molecular LogP values based on the extended solvent-contact model.
- Molecular solvation free energy functions for water and 1-octanol were optimized with 139 organic molecules.
- The LogP values were calculated using the two solvation free energy functions.
- The predicted LogP results compared reasonably well with the experimental ones with the squared correlation coefficient and root mean square error of 0.824 and 0.697, respectively.

Abstract

The logarithm of 1-octanol/water partition coefficient (LogP) is one of the most important molecular design parameters in drug discovery. Assuming that LogP can be calculated from the difference between the solvation free energy of a molecule in water and that in 1-octanol, we propose a method for predicting the molecular LogP values based on the extended solvent-contact model. To obtain the molecular solvation free energy data for the two solvents, a proper potential energy function was defined for each solvent with respect to atomic distributions and three kinds of atomic parameters. Total 205 atomic parameters were optimized with the standard genetic algorithm using the training set consisting of 139 organic molecules with varying shapes and functional groups. The LogP values estimated with the two optimized solvation free energy functions compared reasonably well with the experimental results with the associated squared correlation coefficient and root mean square error of 0.824 and 0.697, respectively. Besides the prediction accuracy, the present method has the merit in practical applications because molecular LogP values can be computed straightforwardly from the simple potential energy functions without the need to calculate various molecular descriptors. The methods for enhancing the accuracy of the present prediction model are also discussed.

Keywords: Partition coefficient; Solvation free energy; Solvent-contact model; Genetic algorithm

1. Introduction

1-octanol/water partition coefficient (P) is defined as the ratio of the concentration of a neutral molecule in 1-octanol to that in water in a two-phase system at equilibrium. Since the first use of 1-octanol/water partitioning system by Collander [1], it has served as a fundamental physicochemical property related with cell permeability, metabolism, bioavailability, and toxicity of molecules. For example, the molecular lipophilicity can be quantified by the logarithm (to base 10) of P, which is one of the most important molecular design parameters [2]. Besides the role of the indicator for lipophilicity, molecular LogP values have also been useful for the estimation of the desolvation cost for binding of a ligand to the receptor protein [3]. LogP is indeed the most widely used molecular descriptor in contemporary drug discovery.

Actually the LogP value of a compound can be determined in a straightforward way by the shake-flask method or reverse phase high performance liquid chromatography. This has made it possible that the experimental LogP values for a large number of simple organic molecules are available in public chemical databases such as PubChem and ChEMBL. However, it has become difficult to cope with all the molecules in the experimental determination of LogP because of the rapid increase in the number of compounds due for example to the advent of combinatorial chemistry [4]. It is therefore necessary to develop the fast and accurate theoretical methods for estimating the LogP values of organic molecules.

Accordingly, a variety of theoretical/computational methods for estimating the molecular LogP have been proposed and explored since the pioneering work of Hansch and coworkers [5]. The most popular methods for estimating LogP include the CLogP and ALogP methods in which the LogP value of a molecule is computed by the summation of all contributions made by the dissected fragments and the individual atoms, respectively [6,7]. To supplement the deficiencies of the fragment and atom-based methods, a whole-molecule

approach using the topological indices was proposed as implemented in the MLogP method [8]. Various quantitative structure-property relationship (QSPR) models with high accuracy were also suggested with the classical [9-11] and some novel molecular descriptors such as the semi-empirical electrotopological index [12], intramolecular interactions between functional groups [13], and SMILES-based optimal descriptors [14]. Very recently, Daina et al. developed a rigorous method for predicting LogP by combining the generalized Born and solvent accessible surface area models [15], which was referred to as iLogP method. Besides the high accuracy in estimating the molecular LogP values, iLogP method was also found to be adequate for coping with a large number of molecules in reasonably short time frame despite the requirement for computing various molecular descriptors. This indicated the usefulness of iLogP method in practical applications of drug discovery.

The present study is undertaken with the aim to establish a potential function from which molecular LogP values can be calculated in a straightforward way based on the difference in solvation free energies of a molecule with respect to water and 1-octanol solvents. For this purpose, P is assumed to be the equilibrium constant for the diffusion reaction of a molecule from water to 1-octanol solvent. Because the difference between the free energy of a molecule in water and that in 1-octanol should be calculated in this model prior to the estimation of LogP, we define the molecular solvation free energy functions for the two solvents based on the extended solvent-contact model. The present method is expected to be useful for coping with a large compound library in the early stage of drug discovery because molecular LogP values can be obtained directly using 3D atomic coordinates and the optimized atomic parameters without any additional calculation.

2. Theory and Computational Methods

2.1. Relation between partition coefficient and solvation free energy

Because P is defined as the ratio of molecular concentration in 1-octanol to that in water at equilibrium, it can be expressed as the equilibrium coefficient for the diffusion reaction of a molecule from water to 1-octanol solvent, which is illustrated in Fig. 1. Within this framework, LogP of a molecule can be expressed in the following form.

$$\mathrm{LogP} = -\frac{\Delta \boldsymbol{G}^{0}}{2.303\boldsymbol{RT}} \tag{1}$$

Because ΔG^0 denotes the difference between the free energy of the solute molecule in 1-octanol and that in water, it can be approximated as the difference in solvation free energies of the solute with respect to the two solvents. LogP of the solute molecule at room temperature (298.15 K) is then expressed as follows in the unit of kcal/mol for ΔG.

$$\mathrm{LogP} = \frac{\Delta \boldsymbol{G}_{\boldsymbol{s}}^{\boldsymbol{wat}} - \Delta \boldsymbol{G}_{\boldsymbol{s}}^{\boldsymbol{oct}}}{1.364} \tag{2}$$

Here ΔG_s^{wat} and ΔG_s^{oct} denote solvation free energies of the solute molecule in water and 1-octanol, respectively.

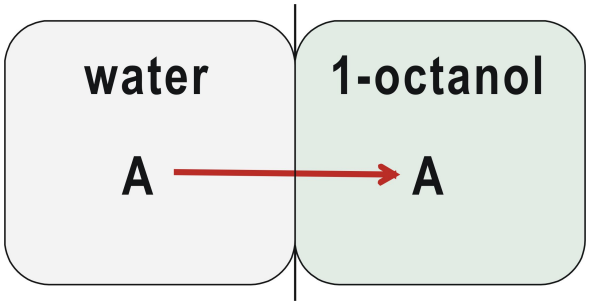


Fig. 1. Diffusion reaction of molecule A from water to 1-octanol solvent as the model system to calculate LogP.

To make it possible to calculate LogP using Eq. (2), we define the molecular solvation free energy functions in the two solvents based on the solvent-contact model [16,17]. As detailed in our previous papers on the extended solvent-contact model [18,19], the solvation free energy of a molecule can be obtained with the interatomic distances (r_{ij}'s) between solute atoms and the three atomic parameters.

$$\Delta G_s^{wat} = \sum_{i}^{atoms} S_i^{wat} \left(O_{i,\max}^{wat} - \sum_{j \neq i}^{atoms} V_j e^{-\frac{r_{ij}^2}{2\sigma^2}} \right) \tag{3}$$

$$\Delta G_s^{oct,pure} = \sum_{i}^{atoms} S_i^{oct} \left(O_{i,\max}^{oct} - \sum_{j \neq i}^{atoms} V_j e^{-\frac{r_{ij}^2}{2\sigma^2}} \right) \tag{4}$$

Here, we assume that the solute molecule can be stabilized in solution through the coordination between the intermolecular solvent-solute interactions and the intramolecular interactions between solute atoms. The gaussian-type envelope function is employed in the potential solvation free energy functions to reflect the effects of all surrounding atoms in the solute on the solvation of each atom in the distance-dependent manner. The key atomic parameters introduced in the solvation free energy function are the atomic solvation (S_i) energy per unit volume, the maximum atomic occupancy ($O_{i,max}$), and the atomic fragmental volume (V_i). The negative and positive signs of S_i parameter indicate the stabilization and destabilization of the solute atom *i*, respectively, due to the combined effects of intermolecular interactions with solvent molecules and intramolecular interactions with the rest of solute atoms. The optimizations of these three atomic parameters for all possible atom types are required to calculate the solvation free energies of the solute with respect to water and 1-octanol, both of which are needed to estimate the LogP values. In the calculation of solvation free energies, we used the same V_i parameters for water and 1-octanol under the assumption that the change of the solvent would have little effect on the molecular volume of

the solute molecule. On the other hand, S_i and $O_{i,max}$ values were optimized separately in the two solvents using the corresponding experimental data for solvation free energies of the molecules in the data set.

Actually the 1-octanol phase is a mixture containing 96% of 1-octanol and 4% of water in the shake flask experiment. Therefore, we used the composition-weighted free energy function given by Eq. (5) to calculate the reference solvation free energy data for 1-octanol required for the estimation of LogP values.

$$\Delta G_s^{oct} = 0.96\Delta G_s^{oct,pure} + 0.04\Delta G_s^{wat} \quad (5)$$

2.2. Preparation of training and test set

To obtain the complete forms of solvation free energy functions, we needed a reference data set with which all the atomic parameters could be optimized. Therefore, we constructed a chemical library containing 178 organic molecules for which experimental solvation free energy values were available for both solvents [20]. These 178 molecules were categorized into 39 subsets according to the structural similarity. More specifically, the structurally similar molecules with Tanimoto coefficient larger than 0.7 were collected into the same cluster. Structural similarities among the molecules were measured using the fingerprints of each molecule generated with the Daylight software as an ASCII string of 1's and 0's. For a subset containing *n* molecules, a single representative one was randomly selected with the probability 1/*n* as an element of the test set. The resultant training and test sets comprising 139 and 39 molecules, respectively, were utilized to build up and validate the LogP prediction method within the framework of solvent-contact model. The structures of the molecules selected as the elements of training and test set are shown in part in Fig. 2 and 3, respectively.

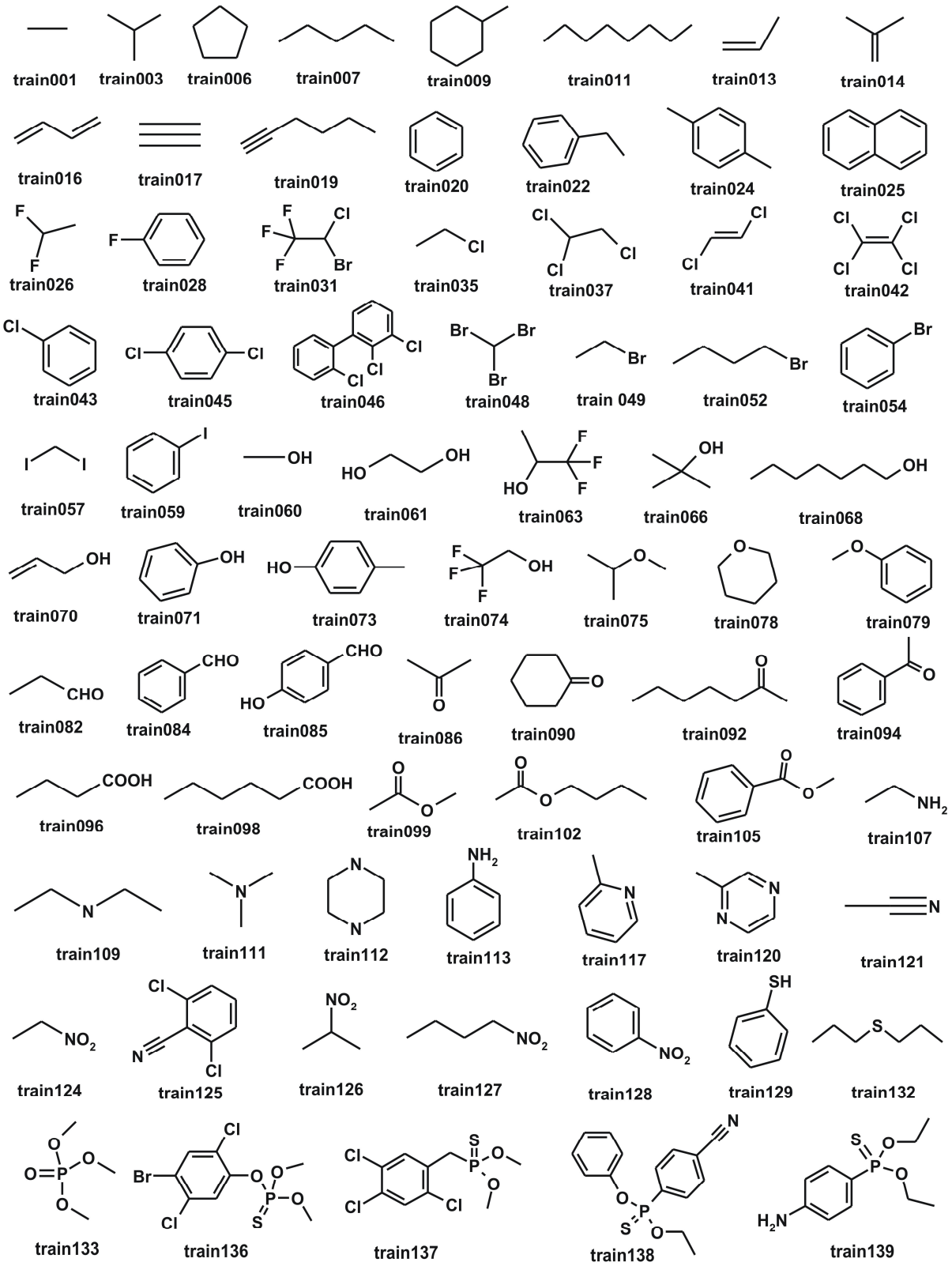


Fig. 2. Chemical structures of the selected molecules in the training set used for the optimization of atomic parameters.

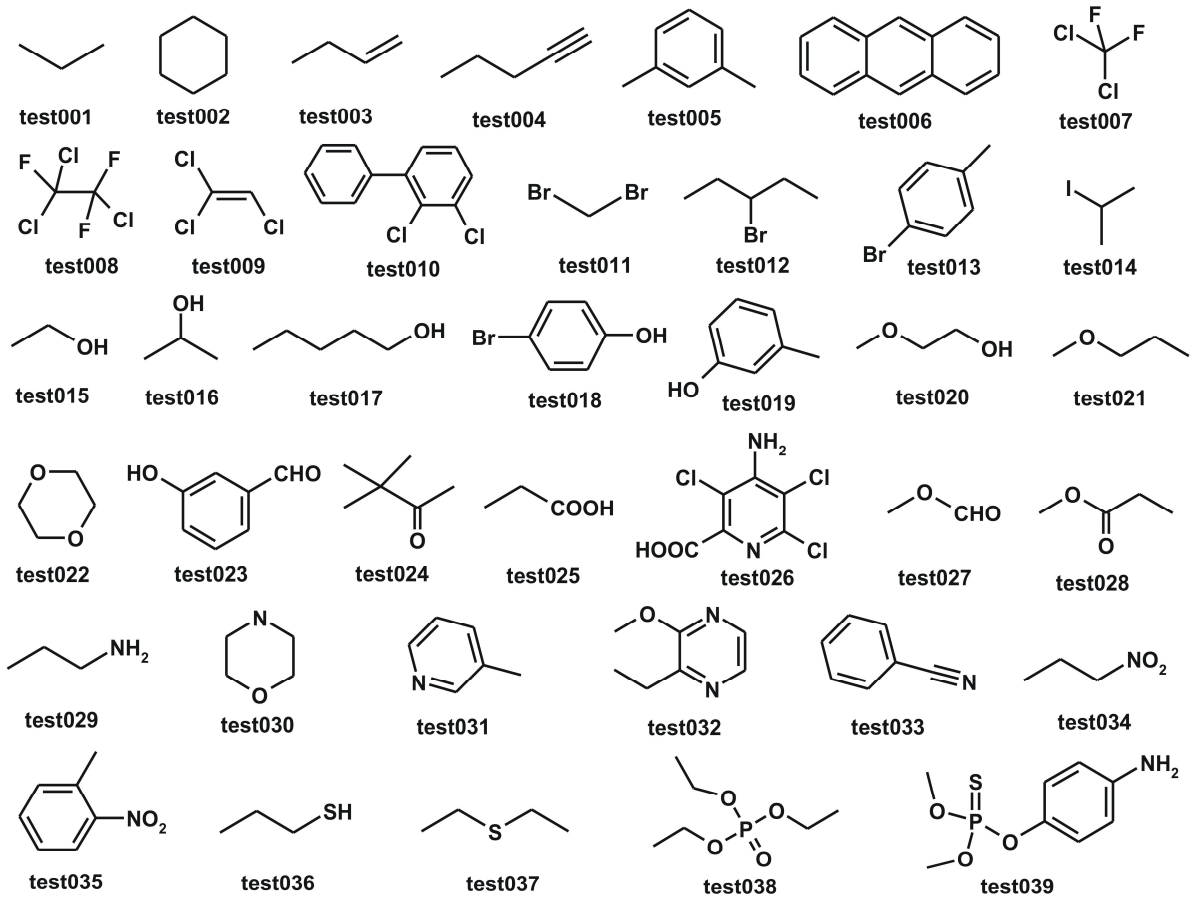


Fig. 3. Chemical structures of the selected molecules in the test set.

3D atomic coordinates of all the molecules in training and test set were obtained using the CORINA program [21] with which a stable molecular conformation could be generated from the structural parameters derived with the X-ray crystal structures of various small molecules. The generated molecular structures were further refined with quantum chemical geometry optimizations at B3LYP/6-31G* level of theory to obtain the final structures from which molecular solvation free energies were calculated. For simplicity, only single molecular conformation was considered in this study although the use of multiple conformations for a molecule would be required to obtain the more reliable solvation free energy value.

2.3. Definition of atom types

Because the individual atoms in a solute molecule have different electronic structures and neighboring atoms, they can contribute to the solvation free energy to different extents.

Therefore, the atom types in molecules should be specified in such a way to reflect the characteristics of each solute atom. Besides the basic atomic properties such as the electronegativity and hybridization state, the number of substituents was also considered in the atom type definition to discriminate the differences in solvent accessibilities among the atoms in a molecule. To describe all the atoms included in 178 molecules in the data set, we extended the space of atom types further to reflect the peculiar electronic structures of some functional groups. For example, carbonyl carbons were distinguished from the normal sp^2 carbons to reflect the significant positive atomic charge. Similarly, specific atom types were assigned to the oxygen atoms of phenolic and carboxylic acid groups because their electronic structures should be discriminated from those of normal sp^3 oxygens. The necessity for these subdivisions becomes apparent by noting the general notion that phenol and acetic acid are 10^6- and 10^{11}-fold more acidic than ethanol in water, respectively. The OH groups in alcohol, phenol, and carboxylic acid groups are thus expected to exhibit different patterns in interacting with solvent molecules, which should be reflected in the optimized atomic parameters. The hydrogens attached to nitrogen (H.N) and oxygen (H.O) were also subdivided according to the polarities of X–H bonds because the strength of interactions with solvent depends on the acidity and basicity of the solute atom. As a consequence, a total of 41 atom types were defined to describe all the molecules in the data set. For simplicity to implement the atom type classifications, all atom types were designated in the similar way to those in Sybyl MOL2 format.

2.4. Optimization of the atomic parameters with genetic algorithm

The determination of three atomic parameters for all atom types was required to calculate the molecular solvation free energies with respect to water and 1-octanol solvents. Among them, the V_i parameter represents the fragmental volume of atoms with type i in

molecules. Because V_i parameters exhibited a bad convergent behavior in the simultaneous optimization of three kinds of atomic parameters, they were optimized separately with a standard genetic algorithm as detailed in our previous papers [18,19]. The S_i and $O_{i,max}$ parameters of each atom type were then optimized through the standard genetic algorithm to make the solvation free energy functions complete. This started with the definition of a generation comprising 100 vectors whose elements were V_i, S_i, and $O_{i,max}$ parameters for all the defined atom types. In the next step, 50 of 100 vectors were removed with a bias toward preserving the best fit with the lowest error. The empty 50 vectors were then filled with the new ones constructed from the top 50. These new vectors were generated with point mutations to alter the values of S_i and $O_{i,max}$ parameters with probability 0.01, and with cross breeds with probability 0.6 to select some S_i and $O_{i,max}$ parameters from one vector to replace the corresponding elements of another vector of the top 50. The 50 new vectors created in these ways were then evaluated together with the top 50. This cycle was iterated until satisfying the convergence criterion. To evaluate the 100 vectors, we used the error hypersurface (F_s) defined by the sum of the absolute values of the differences between the molecular solvation free energies measured from experiment ($\Delta G^i_{\exp}$) and those estimated with the energy functions (ΔG^i_{calc}). This fitness function can be written as follows.

$$F_s = \sum_{i}^{molecules} \left| \Delta G^i_{\exp} - \Delta G^i_{calc} \right| \tag{6}$$

During the operation of genetic algorithm, the σ value in Eqs. (3) and (4) was set equal to 3.5 Å. The atomic parameters converged to their final values after around 10,000 iterations in the optimizations for both water and 1-octanol solvents.

3. Results and Discussion

As the starting point to propose an accurate computational method for estimating the molecular LogP, we assumed that it would be proportional to the difference between the solvation free energy of a molecule for water and that for 1-octanol. To validate the suitability of this approximation, we examined the similarity of the LogP values calculated with Eq. (2) using the experimental solvation free energy data for both solvents to those measured directly from the experiments. Only 118 and 30 molecules in the training and test set were considered respectively in this validation study because the experimental data for LogP values were unavailable for the rest of 30 molecules. As shown in Fig. 4, the squared linear correlation coefficient (R^2) between the two LogP data amounts to 0.992 with the associated slope and intercept values of 0.988 and 0.031, respectively. This nearly complete agreement indicates that the precise estimations of molecular solvation free energies for water and 1-octanol are sufficient to obtain the experimental LogP values. Hereafter, the LogP values calculated with Eq. (2) and experimental solvation free energies will be referred to as the experimental ones.

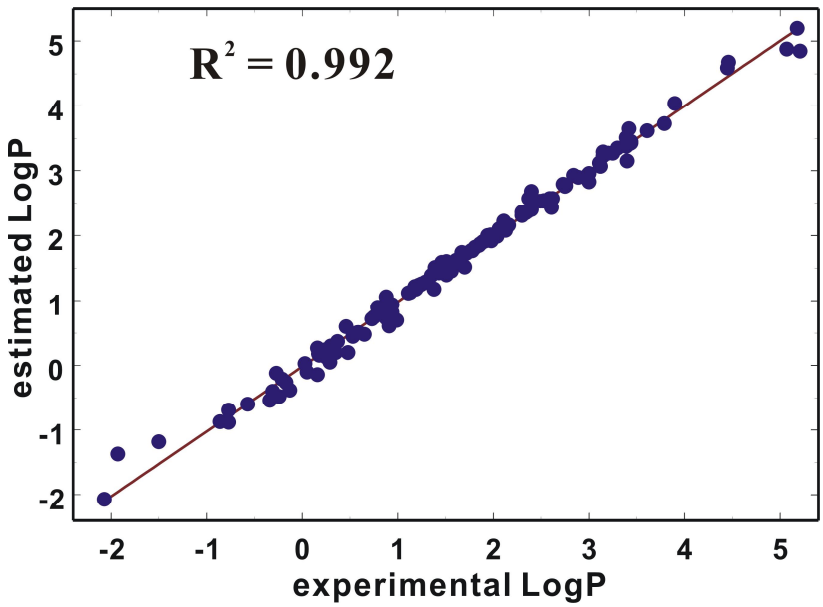


Fig. 4. Correlations between the LogP values measured directly from experiments and those estimated with experimental solvation free energies for water and 1-octanol.

Prior to the calculation of solvation free energies with respect to water and 1-octanol for 139 and 39 molecules in the training and test set, respectively, their geometries were fully optimized at B3LYP/6-31G** level of theory. These structural refinements could be accomplished without a significant computational burden because all geometry optimizations were completed in fifteen steps at the most. Although some molecules contain an ionizable group such as carboxylic acid and amine moiety, only neutral form was considered for all molecules to be consistent with the definition of solvation free energy. With the energy-minimized structures of the molecules in the training set and their experimental solvation free energies for water and 1-octanol, we obtained the atomic parameters in the solvation free energy function by the operation of a standard genetic algorithm as described in the previous section. Because V_i parameters exhibited an oscillatory behavior without convergence to the proper values, they were allowed to vary even among the atoms of the same type. By virtue of this additional flexibility, V_i parameters could be determined so as to successfully reproduce the van der Waals volumes of all the molecules in the training and test sets.

Listed in Table 1 are the $O_{i,max}$ and S_i parameters optimized for 41 atom types with respect to water and 1-octanol solvents. For carbon atoms, thirteen atom types were defined to represent sp^3, sp^2, sp, aromatic, and carbonyl carbons with varying number of substituents. Similarly, seven atom types were introduced for nitrogen to discriminate sp^3, aromatic, planar, and nitro groups. Oxygens were also divided into eight atom types to describe sp^3, sp^2, planar, carboxylic acid, ester, and nitro groups. In case of hydrogen, we defined six atom types according to the property of the adjacent heavy atom. The atom types for the majority of carbon, nitrogen, oxygen, and hydrogen atoms were thus subdivided according to the number of substituents as well as to basic atomic properties. This extension of the atomic parameter space seems to makes it possible to discriminate the atoms with different solvent accessibilities, which would culminate in the improvement of solvation free energy functions.

Table 1

Maximum atomic occupancy ($O_{i,max}$) and atomic solvation (S_i) parameters of various atom types optimized for water and 1-octanol solvent. Numbers in parenthesis indicate the number of occurrences of each atom type in the training set.

atom type	description (# of occurrence in training set)	$O_{i,max}$ (Å^3)		S_i (kcal/molÅ^3)	
		water	1-octanol	water	1-octanol
C.3_1	sp^3 carbon with 1 substituent (145)	354.8	361.0	1.000	-6.635
C.3_2	sp^3 carbon with 2 substituents (183)	354.6	350.0	0.746	-5.698
C.3_3	sp^3 carbon with 3 substituents (12)	331.0	333.8	0.571	-6.127
C.3_4	sp^3 carbon with 4 substituents (11)	393.0	350.0	2.873	-4.071
C.2_1	sp^2 carbon with 1 substituent (8)	383.0	397.1	0.937	-3.571
C.2_2	sp^2 carbon with 2 substituents (10)	341.0	363.5	1.159	-4.095
C.2_3	sp^2 carbon with 3 substituents (3)	340.3	342.5	0.905	-3.540
C.ar_2	aromatic carbon with 2 substituents (176)	355.6	380.5	-1.429	-4.048
C.ar_3	aromatic carbon with 3 substituents (62)	375.4	374.1	0.032	-4.714
C.1_1	*sp* carbon with 1 substituent (4)	342.2	369.7	-0.905	-3.714
C.1_2	*sp* carbon with 2 substituents (7)	371.4	379.5	1.238	-3.048
C.CO_1	carbonyl carbon with 1 substituent (3)	362.5	399.2	-2.794	-10.000
C.CO_2	carbonyl carbon with 2 substituents (21)	347.9	351.6	-1.190	-8.889
N.3_1	sp^3 nitrogen with 1 substituent (6)	384.3	318.3	-11.667	-1.619
N.3_2	sp^3 nitrogen with 2 substituents (4)	308.7	299.8	-13.397	4.000
N.3_3	sp^3 nitrogen with 3 substituents (3)	377.8	368.4	-11.238	0.238
N.ar	aromatic nitrogen (6)	345.9	306.3	-12.667	-0.429
N.1	*sp* nitrogen (5)	320.6	354.4	-12.064	-5.460
N.pl_1	planar nitrogen with 1 substituent (3)	329.7	318.9	-9.762	4.683
N.no2	nitrogen in nitro group (4)	304.3	320.0	-2.937	-1.000
O.3_1	sp^3 oxygen with 1substituent (11)	337.6	311.0	-13.524	0.032
O.3_2	sp^3 oxygen with 2 substituents (25)	330.0	334.3	-7.841	-2.000
O.2	sp^2 oxygen (48)	316.7	328.4	-14.000	-0.476
O.pl_1	planar oxygen with 1 substituent (4)	293.7	298.6	-12.095	-2.000
O.pl_2	planar oxygen with 2 substituents (6)	330.5	337.1	-5.698	-2.190
O.es_1	sp^3 oxygen in carboxylic acids (4)	337.9	340.2	-11.238	-0.127
O.es_2	sp^3 oxygen in esters (7)	340.8	341.3	-0.571	0.762
O.no2	oxygen in nitro group (8)	325.6	327.5	-0.937	-3.000
F	fluorine (24)	474.6	503.5	-4.127	-2.683
Cl	chlorine (46)	442.5	423.3	-2.143	2.714
Br	bromine (15)	351.4	382.9	1.365	0.667
I	iodine (5)	420.3	420.6	-2.000	-2.175
S.3	sp^3 sulfur (4)	516.7	454.8	-1.905	-3.921
S.2	sp^2 sulfur (5)	535.7	619.0	-2.159	-4.254
P	phosphorus (8)	600.0	479.4	2.857	-8.333
H.C	hydrogen bonded to carbon (1027)	255.2	243.3	2.889	-0.270
H.N3	hydrogen bonded to sp^3 nitrogen (14)	255.7	260.0	-1.810	-9.143
H.Np	hydrogen bonded to planar nitrogen (4)	201.1	192.9	-2.000	-12.571
H.O3	hydrogen bonded to sp^3 oxygen (12)	229.4	212.5	-6.341	-18.492
H.Op	hydrogen bonded to planar oxygen (4)	253.0	259.5	-8.937	-18.413
H.Oa	hydrogen in carboxylic acid group (4)	214.4	229.0	-4.095	-15.556

When the optimized $O_{i,max}$ and S_i parameters for water are compared to those for 1-octanol, it follows immediately that the atomic parameters vary significantly with the change of solvent. Actually this is not surprising because of the large differences in various physicochemical properties between water and 1-octanol. Although each atomic parameter was obtained with a complicated procedure under consideration of the atomic properties and the chemical environments present in a variety of molecules in the training set, some interesting tendencies are observed in the optimized atomic parameters. For example, we note that the $O_{i,max}$ values of the second-period atoms range from 300 to 400 in both solvents, and increase together with atomic radius from hydrogen to the third-period atoms. Whereas the $O_{i,max}$ values are relatively similar among the varying atom types, the S_i parameters appear to change significantly with the variation of atom types even in the case of the same element. For instance, the S_i value of carboxylate oxygen (O.es_1) is even more negative than that of the ester group (O.es_2) in water as compared to only 0.3% difference in their $O_{i,max}$ values. This may be understood in the sense that the carboxylic acid group can be ionized in aqueous solution while the ester group exists in the neutral form.

Despite the large fluctuations with respect to atom types, the optimized S_i parameters reveal the general trends consistent with the peculiarities of atomic and electronic structures in molecules. We note in this regard that the overall interactions between the solute carbon atoms and water molecules should be repulsive in the present solvation model because most carbon atoms have positive S_i values. This is consistent with the general notion that water is immiscible with hydrocarbons. On the other hand, it is also found that the S_i values of all carbon atoms become negative with the change of solvent from water to 1-octanol, which implies the attractive interactions between the carbon atoms with solvent molecules. Such a significant variation of S_i values in 1-octanol can be attributed to the increase in the hydrophobicity of solvent that facilitates the van der Waals interactions of nonpolar solute

carbon atoms with the hydrophobic alkyl chain of 1-octanol solvent molecules. It should also be noted that the S_i values of the two atom types for the carbonyl carbon (C.CO_1 and C.CO_2) are negative in both solvents, which indicates their favorable interactions with bulk solvent even in aqueous solution. Actually, the attractive interactions between carbonyl carbons and water can be anticipated because they have partially positive charges due to the electron withdrawal by the neighboring carbonyl oxygen.

Consistent with the critical role of nitrogen and oxygen atoms in the solubility of organic molecules in polar solvents, their S_i values appear to be highly negative for most atom types in water. Indeed they can be stabilized in aqueous solution through the local hydrogen bonds with water molecules as well as through the long-range electrostatic interactions with bulk solvent, which has the effect of making the interactions of the solute atoms with solvent thermodynamically favorable. As can be seen in Table 1, however, most S_i values for nitrogen and oxygen atoms become positive or much less negative in 1-octanol. This is not surprising because the majority of molecular volume of 1-octanol comprises the nonpolar hydrocarbon that is immiscible with the polar groups of a solute molecule.

In case of hydrogens, S_i parameters tend to be more negative in both solvents as the heavy atom to which the hydrogen of interest is bonded becomes more electronegative. For example, the average S_i value for hydrogens decreases from 2.889 to -1.905 and from -0.270 to -10.857 with respect to water and 1-octanol solvent, respectively, with the change of the central atom from carbon to nitrogen. Furthermore, all S_i values for the hydrogens attached to the oxygen (H.O3, H.Op, and H.Oa) fall lower than -4 and -15 in water and in 1-octanol, respectively. The decrease in the S_i values of the hydrogens that belong to the more electronegative atoms can be attributed to the increase in their partial positive charges. Because the formation of hydrogen bonds with solvent molecules can be facilitated by the increased positive charge of polar hydrogens, the highly negative S_i parameters of H.O atom

types can be related with the strengthening of hydrogen-bond interactions with solvent molecules.

It seems to be quite unexpected that the S_i parameters of polar hydrogens (H.N and H.O) are more negative in 1-octanol than in water, which means that the attractive interactions between solvent and polar solute atoms get stronger with the increase of solvent hydrophobicity. This contradictory result may be understood in the context that the extent of screening of electrostatic interactions between the two point charges depends on the dielectric constant of bulk solvent. Because the dielectric constant of 1-octanol (10) is lower than that of water (78) [22], the change of solvent from the latter to the former has the effect of strengthening the long-range electrostatic interactions between the polar hydrogens of the solute and the oxygen atoms of solvent molecules. The highly negative S_i values of the solute polar hydrogens in 1-octanol can thus be attributed to the simultaneous presence of a small polar (–OH) moiety and a long hydrophobic alkyl (C_8H_{17}–) group that is responsible for the low dielectric constant.

With all the optimized V_i, S_i, and $O_{i,max}$ parameters, molecular solvation free energies could be calculated directly from Eqs (3) and (4). The linear correlation diagrams between the experimental and calculated ΔG_s^{wat} values are shown in Fig. 5. With the test set comprising 39 molecules in Fig. 2, we obtain the R^2 value of 0.893, which is similar to that of the fitting with the training set of 139 molecules (0.891). Root mean square error (RMSE) of the estimated ΔG_s^{wat} values from the experimental ones amounts to 0.868 kcal/mol for the test set. Although the error may not be negligible, the present solvation model exhibits a reasonably high predictive power because the experimental ΔG_s^{wat} values covered a wide range of ~14 kcal/mol. Therefore, it seems to be reasonable for the calculated ΔG_s^{wat} data to be used for predicting the LogP values.

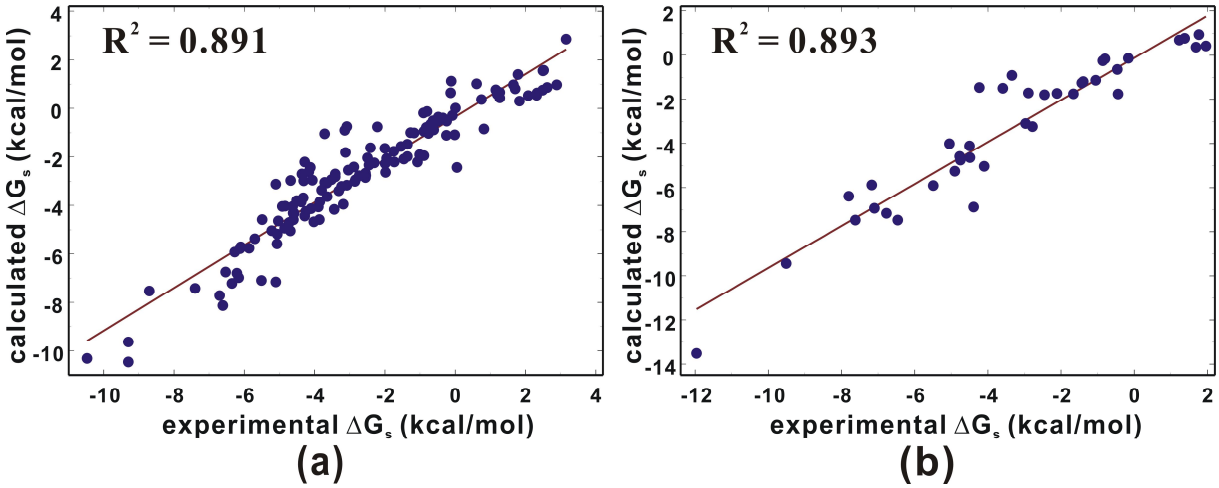


Fig. 5. Correlation diagrams for the experimental solvation free energies in water versus those calculated with the solvation free energy function for (a) 139 molecules in the training set and (b) 39 molecules in the test set. The root mean square errors of the experimental and calculated solvation free energies amount to 0.909 and 0.868 kcal/mol for the training and test set, respectively.

To prepare the calculated ΔG_s^{oct} data required to estimate the molecular LogP values, we also optimized the solvation energy function for 1-octanol using the same training and test sets employed for estimating the hydration free energies. The correlations between the experimental and calculated ΔG_s^{oct} values are displayed in Fig. 6. We obtain the R^2 value of 0.895 with the test set, which is a little lower than that of the fitting with the training set (0.918). The RMSE values for predicting the ΔG_s^{oct} values of the molecules in the training and test sets amount to 0.715 and 0.800 kcal/mol, respectively. R^2 and RMSE values for the estimated ΔG_s^{oct} values are similar to those for the comparison of the experimental and computational ΔG_s^{wat} values. Our extended solvent contact model expressed in Eqs (3) and (4) is thus found to be useful for predicting the molecular solvation free energies with reasonable accuracy irrespective of solvent. It is therefore most likely that the calculated molecular solvation free energies can be utilized for the estimation of the other physicochemical quantities that have a functional dependence on solvation free energy.

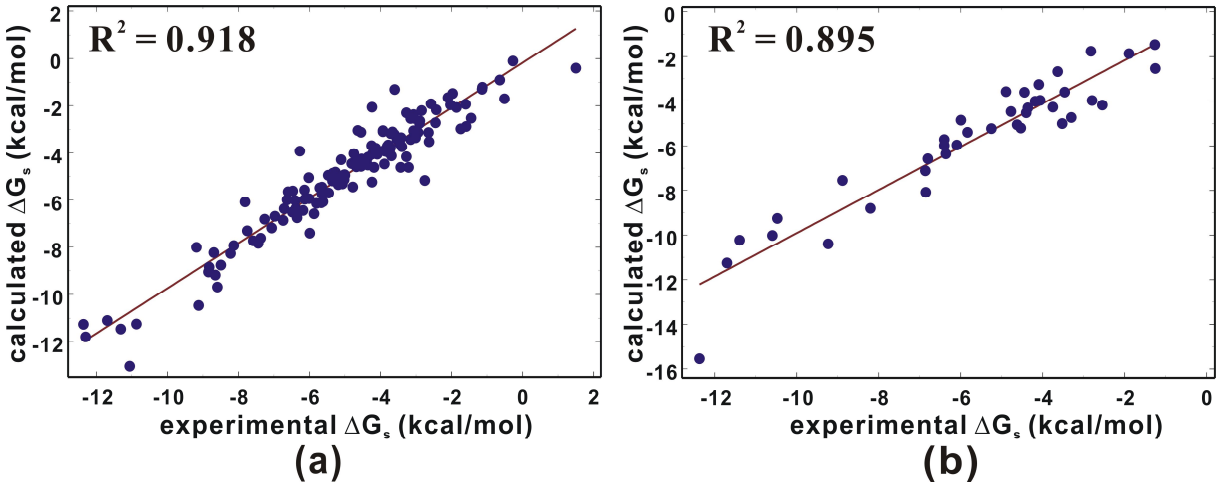


Fig. 6. Correlation diagrams for the experimental solvation free energies in 1-octanol versus those calculated with the solvation free energy function for (a) 139 molecules in the training set and (b) 39 molecules in the test set.

It is worth noting that some atom types such as O.3_2, N.3_3, and N.pl_1 are rare in the training set. Such a low occurrence of atom types in the training set may affect the accuracy of solvation free energy functions. For example, the differences between the experimental and calculated solvation free energies of [1,4]dioxane (test022) amount to more than 1 kcal/mol for both water and 1-octanol. These large deviations may be attributed to the incomplete optimization of atomic parameters for the ether moiety due to the low occurrence of the atom type (O.3_2) in the training set.

Using the solvation free energy data predicted for water and 1-octanol solvents, we finally calculated the LogP values of all molecules contained in the training and test sets to validate the effectiveness of our extended solvent-contact model in estimating the LogP values. The results for the linear regression between the experimental and calculated LogP data are illustrated in Fig. 7. We see that the calculated LogP values compare reasonably well with the experimental results with the associated R^2 values of 0.810 and 0.824 for the training

and test sets, respectively. RMSE values for 139 training set and 39 test set molecules amount to 0.633 and 0.697, respectively. Judging from the R^2 and RMSE values, the present method seems to be comparable in accuracy to the solvent-dependent conformational analysis [23] and to structural analogue approach [24] tested with 11 and 78 molecules, respectively. However, the accuracy seems to be insufficient when compared to the recently reported methods such as QSPR modeling with intramolecular interaction parameters [13] and the hybrid GB/SA model [15]. Such a relatively low performance may be attributed to the use of a small number of molecules (139) in the reference dataset in the optimization of solvation free energy functions. This was actually inevitable because the experimental solvation free energy data for 1-octanol were available only for a small number of molecules in publicly accessible chemical databases. Despite the relatively moderate accuracy, the present extended solvent-contact model has a merit in that one can compute the molecular LogP value in a straightforward way using the potential energy functions and the atomic coordinates of molecules without additional computational burden for calculating the molecular descriptors and parameters.

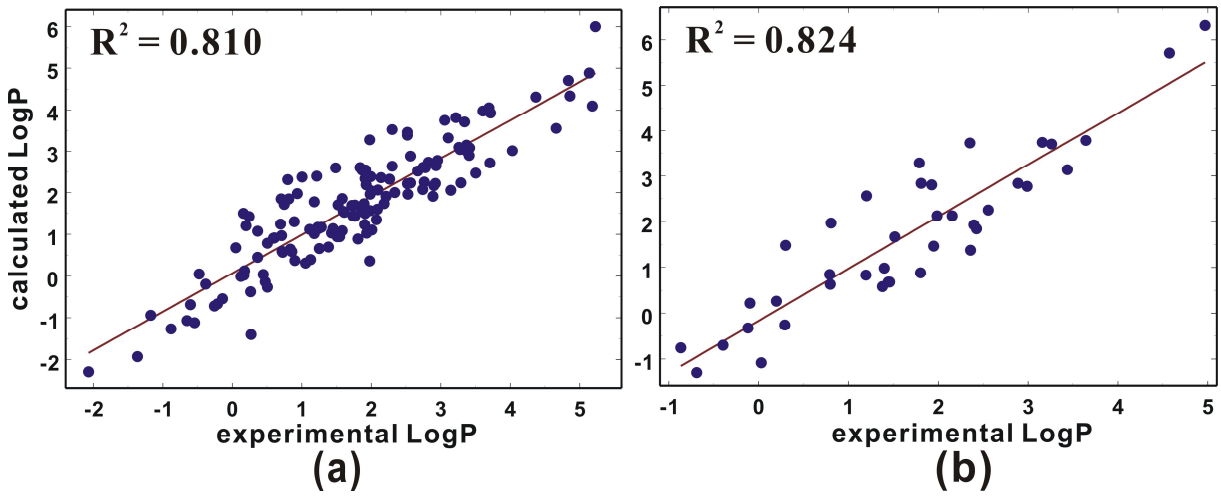


Fig. 7. Correlation diagrams for the experimental and calculated LogP values for (a) 139 molecules in the training set and (b) 39 molecules in the test set.

To address the dependence of the LogP prediction results on the contents of the training and test sets, we repeated the optimization of the atomic solvation parameters and the calculation of molecular LogP values using the two additional training/test set combinations. These second and third training/test sets were constructed in the same way as described in the previous section to include 39 representative molecules and the remaining 139 molecules in the test and training sets, respectively. As shown in Table 2, R^2 and RMSE values of the test set change from 0.824 and 0.697 to 0.817 and 0.711 for the second combination, and to 0.825 and 0.686 for the third one, respectively. These little changes in the results of prediction are actually not surprising because the molecules were divided into training and test sets based on the same criterion to avoid the structural redundancy in the test set.

Table 2

Performance measures of LogP prediction model for varying training/test set combinations.

	combination 1		combination 2		combination 3	
	training set	test set	training set	test set	training set	test set
R^2	0.810	0.824	0.806	0.817	0.818	0.825
RMSE	0.633	0.697	0.651	0.711	0.629	0.686

The largest errors in the prediction of molecular LogP values are observed for N,N-dimethylmethanamine (train111), 2-nitropropane (train126), and 3-bromopentane (test012) with the difference between experimental and calculated LogP values larger than 1.5. To find the source of such large errors, we compare the calculated and experimental values of ΔG_s^{wat} and ΔG_s^{oct} values for the three compounds in Table 3. A large discrepancy (2.29 kcal/mol) between experimental and calculated ΔG_s^{oct} values is observed for train111 as compared to almost no error in the prediction of hydration free energy. This indicates that the error in

estimating the LogP of train111 stems from the inaccuracy in calculating its ΔG_s^{oct} value. On the other hand, the imprecise prediction of ΔG_s^{wat} value of train126 seems to be responsible for the large error in the estimation of LogP. Judging from the significant differences between experimental and calculated results for both ΔG_s^{wat} and ΔG_s^{oct} of test012, the inaccuracy in predicting its LogP value can be attributed to the accumulation of errors in calculating the solvation free energies for both solvents. It is thus apparent that the improvement of solvation free energy function and the associated atomic parameters should be necessary for the precise prediction of molecular LogP values.

Table 3

Comparisons of the calculated and experimental results for ΔG_s^{wat} and ΔG_s^{oct} (in kcal/mol), and LogP values of the three most erroneous compounds.

compound	ΔG_s^{wat}		ΔG_s^{oct}		LogP	
	expt	calc	expt	calc	expt	calc
train111	-3.23	-3.21	-3.60	-1.31	0.37	-1.39
train126	-3.14	-0.90	-4.23	-4.05	0.80	2.31
test012	-0.86	-0.22	-3.30	-4.73	1.79	3.31

To address the usefulness of the present LogP prediction method in practical applications of drug discovery, we calculated the LogP values of a total of 214 known drug molecules with molecular weight ranging from 250 to 400 amu for which the experimental LogP values were available. As shown in Fig. 8, the R^2 value between the experimental and calculated LogP amounts to 0.808 with the associated RMSE value of 1.255. A little decrease and increase in the respective R^2 and RMSE values as compared to the results for the 178 small molecules can be understood in the context that the 41 atom types shown in Table 1 should be insufficient to describe all the atoms in the drug molecules that have more complex chemical environments than those shown in Figs. 2 and 3. Therefore, the accuracy of the

present LogP prediction method seems to be enhanced by subdividing the atom types according to the chemical environment for each atom in molecules.

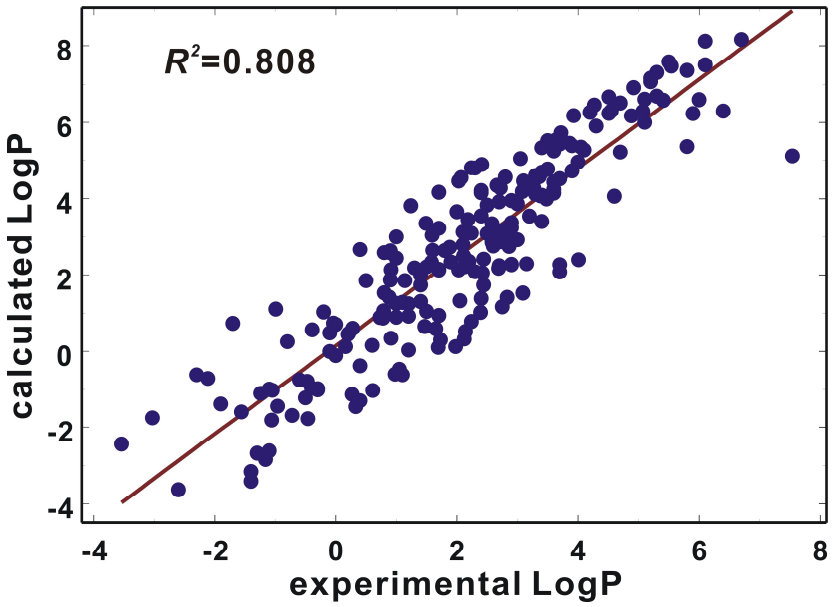


Fig. 8. Linear correlation diagram between the experimental and calculated LogP values for 214 drug molecules.

Despite the reasonably high accuracy of our extended solvent-contact model in predicting the ΔG_s^{wat} and ΔG_s^{oct} values, further improvements seem to be required for it to be useful for estimating the molecular partition coefficients as exemplified in Fig. 8. This is due to the possibility of the error amplifications during the calculation of LogP using the two estimated solvation free energy data. There should be some straightforward methods for the modification of the present solvation model. First, we expect that the better atomic parameters would be obtained by using the multiple conformations that the molecules can adopt instead of using a single representative conformation. This becomes apparent by noting the functional dependence of the molecular solvation free energy function on the atomic distributions in molecule. Monte Carlo and molecular dynamics simulations would be helpful for identifying the local energy minima that are energetically feasible by the solute molecule in solution.

Because both enthalpy and entropy are thermodynamically measurable, the solvation free energy function is likely to be further improved by the decomposition into the two terms. The atomic parameters of enthalpy and entropy terms can be optimized separately with the respective experimental data. It is apparent that molecular solvation free energies would be estimated with higher accuracy in the dual optimization of atomic parameters than in the single parameterization because the more experimental data can be referenced. In the future study, we will focus our interest on the modification of the solvation free energy function to the extent to be able to estimate the molecular partition coefficient with prominent accuracy.

4. Conclusions

We have addressed the applicability of the extended solvent-contact model to the prediction of the LogP values of various organic molecules. Using the potential energy function defined with three kinds of atomic parameters for 41 atom types, molecular solvation free energies for water and 1-ocatanol could be estimated successfully with RMSE value lower than 0.9 kcal/mol. All the atomic parameters were optimized with a standard genetic algorithm using the reference dataset for experimental molecular solvation free energies. Under the assumption that LogP is proportional to the difference between solvation free energy in water and that in 1-octanol, we obtained the LogP prediction model with the R^2 value of 0.824 and the RMSE value of 0.697 for the comparison between the experimental and calculated results. Due to the simplicities in the calculation and in method improvement, the present extended solvent-contact model is expected to serve as a useful tool for the estimations of molecular LogP values of organic molecules with reasonably high accuracy. In order for the present method to serve as a valuable computational tool for LogP prediction, however, further modification should be made in the future when the more experimental data for ΔG_s^{oct} will be available.

Acknowledgments

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2011-0022858).

Appendix A. Supplementary data associated with this article can be found in the online version.

References

1. R. Collander, Partition of organic compounds between higher alcohols and water, Acta Chem. Scand. 5 (1951) 774–780.
2. H. van de Waterbeemd, D. A. Smith, B. C. Jones, Lipophilicity in PK design: methyl, ethyl, futile, J. Comput.-Aided Mol. Des. 15 (2001) 273–286.
3. N. Schneider, G. Lange, S. Hindle, R. Klein, M. A. Rarey, 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 (2013) 15–29.
4. P. T. Corbett, J. Leclaire, L. Vial, K. R. West, J. L. Wietor, J. K. M. Sanders, S. Otto, Dynamic combinatorial chemistry, Chem. Rev. 106 (2006) 3652–3711.
5. C. Hansch, R. M. Muir, T. Fujita, P. P. Maloney, F. Geiger, M. Streich, The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients, J. Am. Chem. Soc. 85 (1963) 2817–2824.
6. R. F. Rekker, H. M. D. Kort, Hidrophobic fragmental constant—Extension to a 1000 data point set, Eur. J. Med. Chem. 14 (1979) 479–488.
7. A. K. Ghose, G. M. Crippen, Atomic physicochemical parameters for three-dimensional-structuredirected quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions, J. Chem. Inf. Comput. Sci. 27 (1987) 21–35.

8. L. Moriguchi, S. Hirono, Q. Liu, I. Nakagome, Y. Matsushita, Simple method of calculating octanol/water partition coefficient, Chem. Pharm. Bull. 40 (1992) 127–130.

9. D. Yaffe, Y. Cohen, G. Espinosa, A. Arenas, F. Giralt, 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 (2002) 162–183.

10. T. J. Hou, X. J. Xu, 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 (2003) 1058–1067.

11. J. K. Wegner, A. Zell, Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method, J. Chem. Inf. Comput. Sci. 43 (2003) 1077–1084.

12. E. S. Souza, L. Zaramello, C. A. Kuhnen, B. S. Junkes, R. A. Yunes, V. E. F. Heinzen, Estimating the octanol/water partition coefficient for aliphatic organic compounds using semi-empirical electrotopological index, Int. J. Mol. Sci. 12 (2011) 7250–7264.

13. P. W. Kenny, C. A. Montanari, I. M. Prokopczyk, $ClogP_{alk}$: a method for predicting alkane/water partition coefficient, J. Comput.-Aided Mol. Des. 27 (2013) 389–402.

14. A. A. Toropov, A. P. Toropova, I. Raska, E. Benfenati, QSPR modeling of octanol/water partition coefficient of antineoplastic agents by balance of correlations, Eur. J. Med. Chem. 45 (2010) 1639–1647.

15. A. Daina, O. Michielin, V. Zoete, iLogP: A simple, robust, and efficient description of n-octanol/water partition coefficient for drug design using GB/SA approach, J. Chem. Inf. Model. 54 (2014) 3284–3301.

16. F. Colonna-Cesari, C. Sander, Excluded volume approximation to protein-solvent interaction. The solvent contact model, Biophys. J. 57 (1990) 1103–1107.

17. P. F. W. Stouten, C. Frömmel, H. Nakamura, C. Sander, An effective solvation term based on atomic occupancies for use in protein simulations, Mol. Simul. 10 (1993) 97–120.

18. H. Choi, H. Kang, H. Park, New solvation free energy function comprising intermolecular solvation and intramolecular self-solvation terms, J. Cheminformatics 5 (2013) 8.

19. H. Park, Extended solvent-contact model approach to SAMPL4 blind prediction challenge for hydration free energies, J. Comput.-Aided Mol. Des. 28 (2014) 175–186.

20. J. Wang, W. Wang, S. Huo, M. Lee, P. A. Kollman, Solvation model based on weighted solvent accessible surface area, J. Phys. Chem. B 105 (2001) 5055–5067.

21. J. Gasteiger, C. Rudolph, J. Sadowski, Automatic generation of 3D-atomic coordinates for organic molecules, Tetrahedron Comput. Methodol. 3 (1990) 537−547.

22. D. R. Lide, CRC Handbook of Chemistry and Physics, 90th Ed., CRC Press, 2009.

23. F. J. Torrens, Universal organic solvent-water partition coefficient model, J. Chem. Inf. Comput. Sci. 40 (2000) 236−240.

24. A. Y. Sedykh, G. Klopman, A structural analogue approach to the prediction of the octanol-water partition coefficient, J. Chem. Inf. Model. 46 (2006) 1598−1603.