DOI 10.1007/s10822-014-9729-z

Extended solvent-contact model approach to SAMPL4 blind prediction challenge for hydration free energies

Hwangseo Park

Received: 7 November 2013 / Accepted: 10 February 2014 / Published online: 20 February 2014

Abstract Extended solvent-contact model was applied to the blind prediction of the hydration free energies of 47 organic molecules included in the SAMPL4 data set. To obtain a suitable prediction tool, we constructed a hydration free energy function involving three kinds of atomic parameters. With respect to total 34 atom types introduced to describe all SAMPL4 molecules, 102 atomic parameters were defined and optimized with a standard genetic algorithm in such a way to minimize the difference between the experimental hydration free energies and those calculated with the hydration free energy function. In this parameterization, we used a training set comprising 77 organic molecules with varying sizes and shapes. The estimated hydration free energies for the SAMPL4 molecules compared reasonably well with the experimental results with the associated squared correlation coefficient and root mean square deviation of 0.89 and 1.46 kcal/mol, respectively. Based on the comparative analysis of experimental and computational hydration free energies of the SAMPL4 molecules, the methods for further improvement of the present hydration model are suggested.

Keywords Hydration free energy · SAMPL4 · Solvent-contact model · Genetic algorithm · Atomic parameters

H. Park (✉)
Department of Bioscience and Biotechnology, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 143-747, Korea
e-mail: hspark@sejong.ac.kr

Introduction

The hydration free energy, i.e. the free energy change for the transfer of a solute molecule from the gas phase to aqueous solution, is of fundamental importance to characterize a variety of equilibria and kinetics in material, biological, and pharmaceutical sciences. For example, it can affect the protein–ligand binding affinity because the desolvation cost for complexation can make a significant contribution to the total binding free energy [1]. The exact determination of hydration free energy is also important in drug discovery because the bioactivities of drug candidates depend on the stability in aqueous solution. This makes the measurement of hydration free energies of drug candidates be a necessary step at the early stage of drug discovery [2]. In particular, the exact determination of the differences among the hydration free energies of structurally similar compounds has become very important with the advent of combinatorial chemistry [3]. Because the experimental measurement of hydration free energy is a time-consuming procedure, however, it has been very difficult to screen a large chemical library from which the active compounds can be identified.

It is also difficult to predict the hydration free energy with theoretical methods due to the complexity of water-solute interactions [4]. Nonetheless, a variety of computational methods for estimating the hydration free energy have been proposed and explored since the earlier work of Onsagar [5]. Included in these methods are the dielectric continuum model [6], Poisson-Boltzmann equation approach [7], all-atom model calculations based on molecular dynamics and Mote Carlo simulations [8–13], and quantum mechanical models [14–17]. Although the high-level quantum mechanical and all-atom models have been able to estimate the hydration free energies with accuracy, they are difficult to be employed widely in practical applications due to the high computational costs [18, 19]. To reduce the computational burden, therefore, a number of efficient computational methods have also been proposed based on various theoretical frameworks such as solvent-accessible surface area model [20, 21], three dimensional reference interaction site model [22], cellular automata based algorithm [23], quantitative structure–property relationship (QSPR) model [24], linear interaction energy method [25], and quantum mechanical continuum solvation models [26].

In the early 1990s, Stouten et al. proposed a hydration free energy function based on the solvent-contact model developed by Colonna-Cesari and Sander [27, 28]. Assuming that the hydration free energy of a molecule could be obtained by the sum over the individual atomic contributions, they obtained a hydration free energy function that included the atomic parameters for the six atom types (C, N, O, N^+, O^-, and S) only. Although this simple hydration model proved to be useful in estimating the structural properties of proteins [28], its applicability could not be extended to organic molecules because the number of atom types was insufficient to discriminate the atoms with various chemical environments. In the previous studies, therefore, we improved Stouten et al.'s hydration model by extending the atom types to cope with various organic molecules [29, 30]. This modification made it possible to estimate the hydration free energies of small organic molecules with reasonable accuracy.

In the present study, we report a test of our extended solvent-contact model through the participation in the hydration free energy category of SAMPL4 blind prediction challenge. To make the hydration free energy function suitable for the molecules included in the SAMPL4 data set (Fig. 1), we augmented the number of atom types to describe a variety of chemical environments present in the data set. This modification seems to have an effect of increasing the accuracy in the estimation of hydration free energies because the subdivision of atom types according to the chemical environments is necessary to cope with organic molecules with varying shapes and atomic compositions. Some hypotheses behind the present hydration model are presented and discussed. We also address the problems that limit the applicability of our extended solvent-contact model, and propose the methods for further improvement.

Computational methods

Construction of the hydration free energy function

As described in the previous papers [29, 30], we defined the molecular hydration free energy (ΔG_{sol}) as a function of three atomic parameters and interatomic distances (r_{ij}'s) between solute atoms.

$$\Delta G_{sol} = \sum_{i}^{atoms} S_i \left(O_i^{\max} - \sum_{j \neq i}^{atoms} V_j e^{-\frac{r_{ij}^2}{2\sigma^2}} \right) \tag{1}$$

Here, we place an emphasis on the fact that an organic solute molecule can be stabilized in aqueous solution as a consequence of the coordination between the intermolecular water-solute interactions and the intramolecular interactions between solute atoms. It should be noted that this extended solvent-contact model excludes the stability of solute in the gas phase and considers the solute-water interactions in liquid phase only. The neglect of gas-phase stability may limit the accuracy of the present hydration model because the gas phase chemical potential of the solute molecules affects their hydration free energies. In the hydration free energy function, we used the gaussian-type envelope function to model the effects of all surrounding atoms in a molecule on the hydration of one atom in distance-dependent manner. The key atomic parameters in the hydration free energy function include the maximum atomic occupancy (O_i^{max}), the atomic fragmental volume (V_i), and the atomic hydration (S_i) energy per unit volume. The negative and positive values of S_i parameter indicate the stabilization and destabilization of the solute atom i, respectively, due to the combined effects of intermolecular interactions with water molecules and intramolecular interactions with the rest of solute atoms. The optimization of three different atomic parameters for all possible atom types is thus prerequisite for the prediction of hydration free energies of the SAMPL4 molecules.

Preparation of training set

To obtain the O_i^{max}, V_i, and S_i parameters present in the hydration free energy function, we prepared a training set with which they could be optimized. This training set comprised 77 organic molecules shown in Fig. 2 for which the experimental hydration free energy data were available [30–34]. These reference molecules were selected in such a way that the training set they made could include all the atom types needed to describe the SAMPL4 molecules that had to serve as the test set in predicting the molecular hydration free energies. It is a drawback of the present hydration model that the training set constructed to calculate the hydration free energies for a given set of molecules cannot be used extensively for a wide range of molecules. If one needs to predict the hydration energy of a molecule whose structure or atomic composition is very different from the existing ones, the atomic parameters should be reoptimized with a new training set for the maximum accuracy of prediction. It would thus be difficult

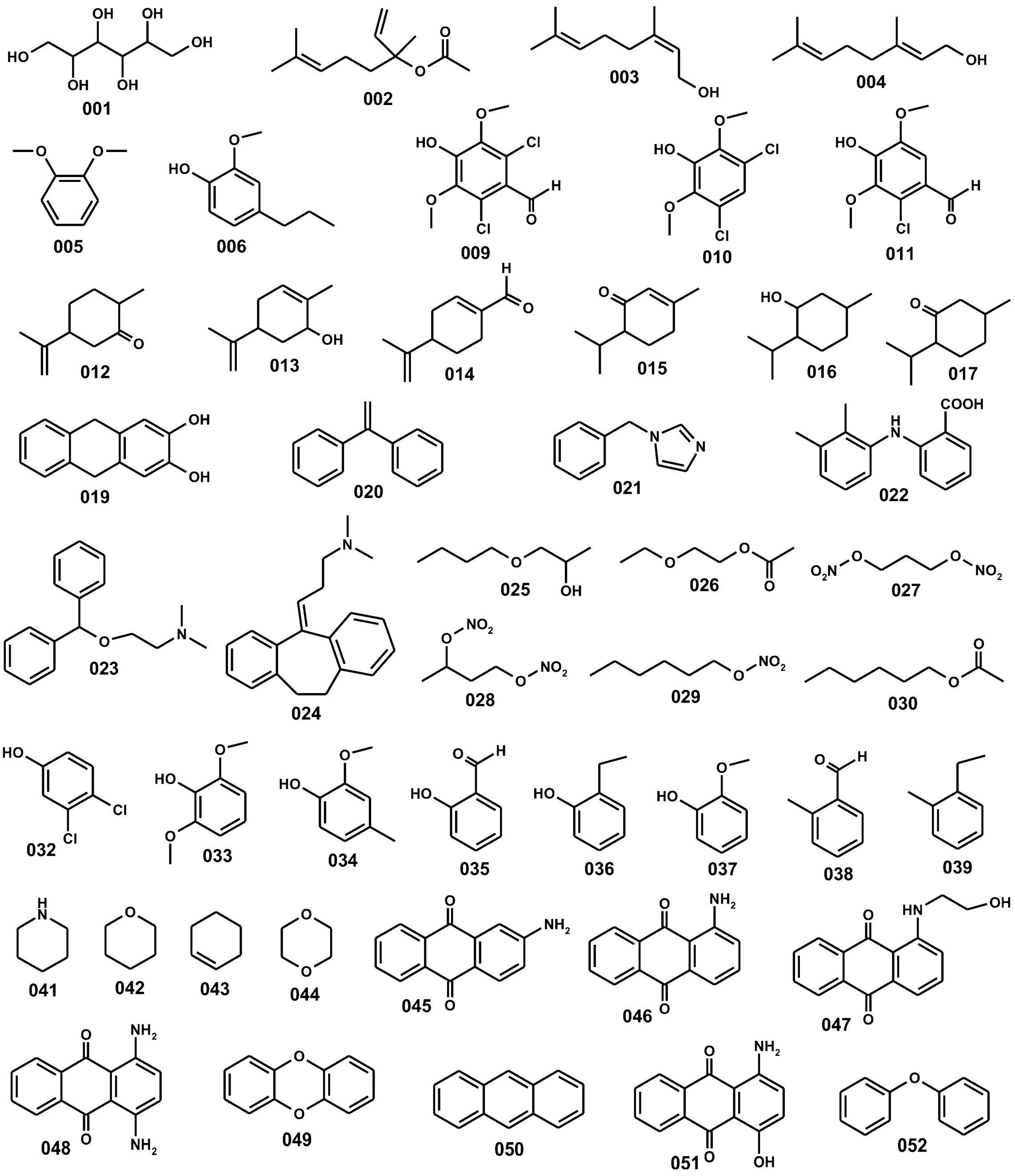

Fig. 1 Chemical structures of the molecules included in the SAMPL4 data set. All molecules are numbered according to the SAMPL4 ID numbering scheme

to estimate the hydration free energy of a molecule with abnormal structure and chemical bonds using the present hydration model.

To obtain the three dimensional atomic coordinates of all molecules in the training and test sets, we used the CORINA program [35] with which a stable molecular conformation could be generated using the conformational parameters derived from the X-ray crystal structures of about 230,000 small molecules contained in the Cambridge Structural Database. The atomic coordinates were then refined through the quantum chemical geometry optimizations at B3LYP/6-31G** level of theory with polarized continuum model (PCM) for hydration to obtain the final structures with which the hydration free energies were calculated. For simplicity, only a single molecular conformation was used in calculating the hydration free energy although the use of multiple conformations for a molecule seemed to be desirable to improve the accuracy in prediction.

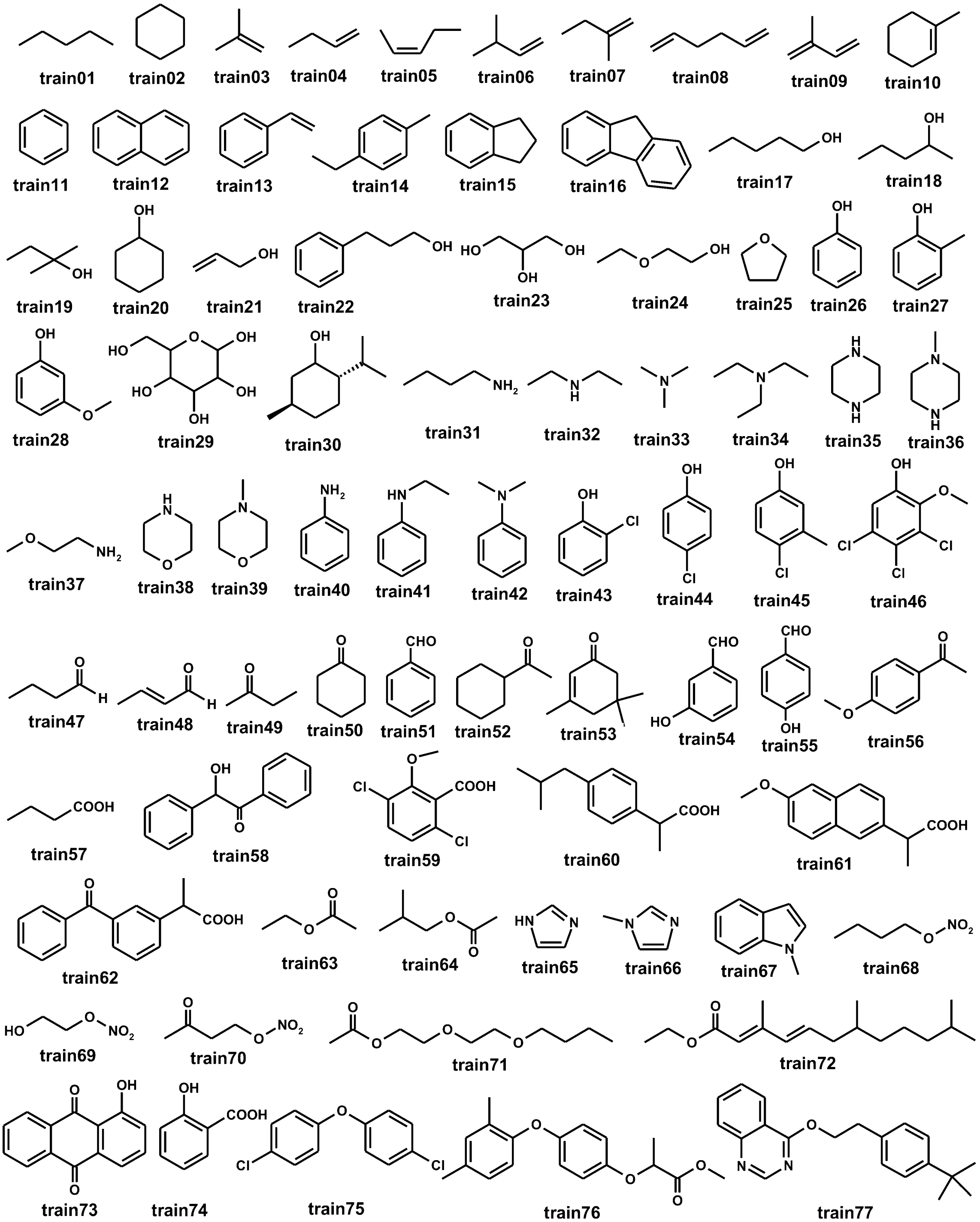

Fig. 2 Chemical structures of 77 molecules in the training set used in the parameterization

Definition of atom types

Because the individual atoms in a molecule have different chemical environments, the extents of their contributions to the molecular hydration free energy can vary significantly. Therefore, the atom type of each atom should be specified under consideration of the element, hybridization state, charge distribution, and the other atomic properties. In the previous study, we defined the atom types present in various molecules on the basis of the element, hybridization state, chemical bond, and number of substituents. We extended the number of atom types in this work to enhance the accuracy in estimating the molecular hydration free energies. For example, carbonyl carbons were distinguished from the normal sp^2 carbons to reflect the significant positive atomic charge developed due to the adjacent electronegative carbonyl oxygen. Similarly, specific atom types were assigned to the oxygen atoms in phenolic and carboxylic acid groups because their physicochemical properties are very different from those of normal sp^3 oxygens. For example, phenol is 10^6-fold more acidic than ethanol in aqueous solution. The oxygen and hydrogen atoms in alcohol and phenol groups are therefore expected to exhibit different patterns for the interaction with water molecules, which necessitates the subdivision of atom types for alcohol and phenol groups. The hydrogens attached to nitrogen (H.N) and oxygen (H.O) were also subdivided according to the polarities of X–H bonds because the extent of interactions with water can be dependent on the acidity and basicity of a solute. As a consequence, 34 atom types were defined in the present hydration model to describe the molecules in SAMPL4 data set whereas only 25 atom types were required in the previous model. For simplicity to implement the atom type classifications, all atom types were designated in similar way to those in Sybyl MOL2 format.

Optimization of atomic parameters with genetic algorithm

The determination of three atomic parameters for all atom types was required to calculate the hydration free energies of the SAMPL4 molecules. 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 the previous papers [29, 30]. The S_i and O_i^{max} parameters were then determined for each atom type to complete the hydration free energy function based on the standard genetic algorithm. 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 atom types under consideration. 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 as many times as desired. 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 experimental molecular hydration free energies ($\Delta G^i_{\exp}$) and those estimated with the energy function (ΔG^i_{calc}). This fitness function can be written as follows.

$$F_s = \sum_i^{molecules} \left| \Delta G^i_{\exp} - \Delta G^i_{calc} \right| \qquad (2)$$

During the operation of genetic algorithm, the σ value in Eq. (1) was set equal to 3.5 Å. The atomic parameters exhibited convergent behavior after 10,000 iterations.

Results and discussion

Prior to the calculation of hydration free energies of 77 and 47 organic molecules in the training and test set, respectively, their geometries were fully optimized at B3LYP/6-31G** level of theory with PCM model for hydration. With the energy-minimized structures of the molecules in the training set and their experimental hydration free energies, we obtained the atomic parameters in the hydration free energy function through the operation of a standard genetic algorithm. Listed in Table 1 are the optimized V_j, O_i^{max}, and S_i parameters for 34 atom types under consideration. For carbon atoms, eleven atom types are defined to represent sp^3, sp^2, aromatic, and carbonyl carbons with varying number of substituents. Similarly, eight atom types are introduced for nitrogen to discriminate sp^3, aromatic, planar, and nitro groups. Oxygen atoms are also divided into eight atom types to represent sp^3, sp^2, planar, carboxylic acid, ester, and nitro groups. In case of hydrogen, we define six atom types according to the property of the atom that forms a covalent bond with the hydrogen. It is thus a characteristic feature of the present hydration model that the atom types for the majority of carbon, nitrogen, oxygen, and hydrogen atoms are subdivided according to the number of substituents and the nature of neighboring atoms. This extension of the atomic parameter space makes it possible to discriminate the atoms with different water accessibilities, which may lead to the improvement in the accuracy of hydration free energy function.

When the optimized V_i, O_i^{max}, and S_i parameters in Table 1 are compared with those obtained in the previous studies, it becomes apparent that the atomic parameters vary significantly with the change of training set and with the number of atom types. This is actually not surprising because each optimized atomic parameter should reflect all the contributions of the atoms with the same type and with different chemical environments to the hydration free energies of a variety of molecules in the training set. Despite such a complexity in parameterization, there is a trend in the optimized atomic parameters. For example, the

Table 1 The optimized atomic fragmental volume (V_i), maximum atomic occupancy (O_i^{max}), and atomic solvation (S_i) parameters for various atom types

Atom type	Description	V_i (Å^3)	O_i^{max} (Å^3)	S_i (kcal/mol Å^3)
C.3_1	sp^3 carbon with 1 substituent	10.123	363.5	1.476
C.3_2	sp^3 carbon with 2 substituents	13.516	356.3	0.381
C.3_3	sp^3 carbon with 3 substituents	17.731	341.1	1.000
C.3_4	sp^3 carbon with 4 substituents	7.778	396.8	2.048
C.2_1	sp^2 carbon with 1 substituent	11.151	385.7	0.650
C.2_2	sp^2 carbon with 2 substituents	9.431	380.2	1.603
C.2_3	sp^2 carbon with 3 substituents	15.913	352.4	0.652
C.ar_2	Aromatic carbon with 2 substituents	8.628	352.0	−0.873
C.ar_3	Aromatic carbon with 3 substituents	15.499	379.4	0.651
C.CO_1	Carbonyl carbon with 1 substituent	9.167	375.4	−3.794
C.CO_2	Carbonyl carbon with 2 substituents	10.556	350.0	−1.397
N.3_1	sp^3 nitrogen with 1 substituent	16.508	306.3	−13.667
N.3_2	sp^3 nitrogen with 2 substituents	8.571	369.2	−11.857
N.3_3	sp^3 nitrogen with 3 substituents	19.683	343.7	−12.429
N.ar	Aromatic nitrogen	23.651	346.8	−14.825
N.pl_1	Planar nitrogen with 1 substituent	7.381	323.0	−9.730
N.pl_2	Planar nitrogen with 2 substituents	30.000	312.7	−10.683
N.pl_3	Planar nitrogen with 3 substituents	17.698	350.6	−9.889
N.no2	Nitrogen in nitro group	6.961	304.0	−2.589
O.3_1	sp^3 oxygen with 1substituent	5.000	343.7	−13.841
O.3_2	sp^3 oxygen with 2 substituents	10.952	324.6	−7.778
O.2	sp^2 oxygen	6.975	344.3	−10.238
O.pl_1	Planar oxygen with 1 substituent	9.365	315.7	−12.571
O.pl_2	Planar oxygen with 2 substituents	7.778	294.0	−5.095
O.es_1	sp^3 oxygen in carboxylic acids	15.317	345.2	−13.508
O.es_2	sp^3 oxygen in esters	19.286	338.2	−1.033
O.no2	Oxygen in nitro group	6.943	325.4	−0.100
Cl	Chlorine	6.984	430.0	−2.730
H.C	Hydrogen bonded to carbon	2.339	242.9	−0.492
H.N3	Hydrogen bonded to sp^3 nitrogen	6.143	248.4	−1.333
H.Np	Hydrogen bonded to planar nitrogen	1.357	202.4	−1.873
H.O3	Hydrogen bonded to sp^3 oxygen	1.286	222.2	−5.863
H.Op	Hydrogen bonded to planar oxygen	5.214	249.2	−8.000
H.Oa	Hydrogen in carboxylic acid group	5.285	226.2	−5.556

O_i^{max} values increase with the increase in atomic radii from hydrogen to second-period atoms and to Cl in the present hydration model including 34 atom types (Table 1) whereas they converged to similar values irrespective of atomic radii in the previous method that required only 25 atom types to describe the molecules in the SAMPL4 data set [30].

In contrast to the similarities among the O_i^{max} values of varying atom types, the V_j parameters are found to change significantly with the variation of atom types even in the case of the same element. As can be seen in Table 1, for example, the V_j value of planar nitrogen with two substituents (N.pl_2) is larger than that of the nitrogen in the nitro group (N.no2) by more than fourfold as compared to only 3 % difference in their O_i^{max} values. Such a large difference between V_j parameters of the similar atoms is actually not surprising because each V_j value represents the average of atomic contributions with the atom type j to the van der Waals volumes of the molecules with various shapes and atomic compositions.

The optimized S_i parameters exhibit a trend consistent with general atomic properties. The overall interactions between the solute carbon atoms and water molecules are predicted to be repulsive in the present hydration model because most C.3, C.2, and C.ar atoms have positive S_i values. This is consistent with the immiscibility of hydrocarbons in water. However, we note that the S_i values become less positive in going from sp^3 to sp^2 and aromatic carbons: the average S_i values of four C.3 and three C.2 atom types amount to 1.226 and 0.968, respectively. This indicates that the interaction of carbon atoms with water becomes less unfavorable with the increase of the s-character in the hybridization state of the solute carbon atom. Such a dependence of S_i value on the degree of s-character can be understood because the increase in the s-character of hybrid orbitals of a central atom leads to the increase in its electronegativity, which culminates in the promotion of dipole–dipole interactions with water molecules. Both atom types for carbonyl carbons (C.CO_1 and C.CO_2) seem to have favorable interactions with bulk solvent because their S_i values are negative. This can be attributed to their partial positive charges due to the electron withdrawal by the neighboring carbonyl oxygen.

In accordance with the pivotal role of oxygen and nitrogen atoms in the stabilization of organic molecules in aqueous solution, their S_i values for most atom types are found to be highly negative. Besides the long-range electrostatic interactions with bulk solvent, they are capable of establishing the local hydrogen bonds with water molecules, which makes the water-solute interactions thermodynamically more favorable. In case of hydrogens, the S_i values become more negative with the increase in the electronegativity of the heavy atom to which the hydrogen of interest is attached. For example, the average S_i value of hydrogen atoms decreases from −0.492 to −1.603 as the central atom changes from carbon to nitrogen. Furthermore, all S_i values for the hydrogen atoms attached to oxygen (H.O3, H.Op, and H.Oa) are found to be less than −5. Thus, the S_i values of hydrogen atoms have a tendency to become more negative as the acidity of the central atom gets higher, which can be attributed to the effective stabilization of the negative charge developed on the central atom. Because the increase in the acidity of a donor facilitates the formation of a hydrogen bond with an acceptor, the highly negative S_i values of H.O atoms indicate their favorable interactions with water by the establishment of strong hydrogen bonds.

The correlations between the experimental hydration free energies and those calculated with the energy function and the associated atomic parameters are illustrated in Fig. 3. With the test set comprising 47 molecules in the SAMPL4 data set, we obtain the squared correlation coefficient (R^2) of 0.89, which is a little lower than that of the fitting with the training set including 77 molecules (0.93). With respect to the accuracy in the prediction of hydration free energies for the SAMPL4 data set, we see that R^2 value increases by 0.05 and 0.08 when the atom types of O.es and H.O are subdivided into O.es_1 and O.es_2, and H.O3, H.Op, and H.Oa, respectively. These results indicate the necessity for the definition of atom types in a specific fashion according to the chemical environment around each atom to improve the accuracy of the hydration free energy function. The root mean square deviation of the estimated hydration free energies for the SAMPL4 test set from the experimental ones amounts to 1.46 kcal/mol. Although this accuracy may be insufficient as compared to those of high-level quantum chemical methods and statistical simulations with all-atom model, the merit of our extended solvent-contact model lies in that one can produce the hydration free energies straightforwardly from a potential function without significant computational burden.

To address the effect of the extension of atom types on the accuracy of hydration free energy function, we evaluated the previous hydration model that required 25 atom types only to describe the molecules in SAMPL4 data set. The correlations between the hydration free energies measured from experiments and those obtained with Eq. (1) are illustrated in Fig. 4. With the test set comprising 47 SAMPL4 molecules, we obtain the R^2 value of 0.78, which is significantly smaller than that of the fitting with the training set including 77 molecules (0.91). The root mean square error (RMSE) and the average unsigned error (AUE) for the prediction of hydration free energies of 47 SAMPL4 molecules amount to 1.86 and 1.48 kcal/mol in this case, respectively, as compared to 1.46 and 1.12 kcal/mol in the present solvation model with 34 atom types. These large decreases in RMSE and AUE values indicate the necessity for the extension of atom types to better estimate the hydration free energies of SAMPL4 molecules. The large difference between the R^2 values for the training and test sets indicates that the atomic parameters should be over-trained with 25 atom types and the training set comprising the molecules shown in Fig. 2. Such a modest accuracy of the previous method can be attributed to the use of only 25 atom types for the solute atoms included in 47 SAMPL4 molecules. The increase of the R^2 value from 0.78 (Fig. 4b) to 0.89 (Fig. 3b) with the extension of atom types indicates that the hydration free energy function can become more accurate due to the subdivision of the atom types that may have the effect of enhancing the statistical fit for the atomic parameters with genetic algorithm.

A large difference in the predictive accuracy between the previous and the present hydration models is observed for **001** that contains six hydroxyl groups in the molecular structure: the difference between experimental and calculated hydration free energies amounts to 4.49 kcal/mol in the previous method with 25 atom types as compared to only 2.14 kcal/mol in the present model with 34 atom

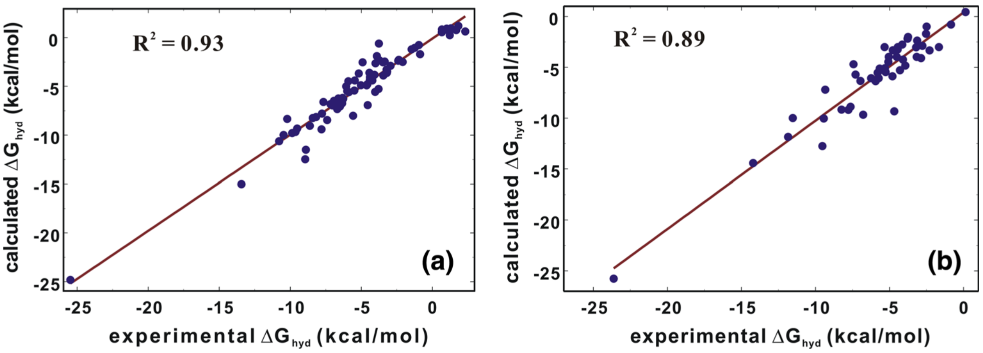

Fig. 3 Correlation diagrams for the experimental hydration free energies (ΔG_{hyd}) versus those calculated with the hydration free energy function for (**a**) 77 molecules in the training set and (**b**) 47 molecules in the SAMPL4 test set. The root mean square errors of the experimental and calculated hydration free energies amount to 1.17 and 1.46 kcal/mol for the training and test sets, respectively

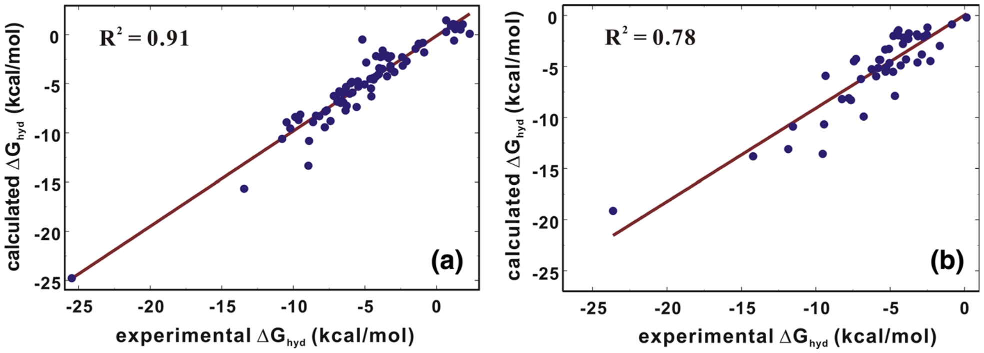

Fig. 4 Correlation diagrams for the experimental hydration free energies versus those estimated with the previous hydration model with 25 atom types for **a** 77 molecules in the training set and **b** 47 molecules in the SAMPL4 test set

types. In contrast, both models predict the hydration free energy of train29 with accuracy that has five hydroxyl moieties. These different results for the molecules with similar functional groups can be understood in terms of the different distributions of interatomic distances in molecules. Within the framework of extended solvent-contact model, the hydration free energy of a molecule depends on the interatomic distances between solute atoms as well as on the atomic parameters, which is reflected in Eq. (1). In case of cyclic compounds such as train29 in which the interatomic distances are short, the contribution of atomic parameters becomes relatively less significant due to the large screening effect of neighboring atoms. This is the reason for the similarity in the hydration free energies of train29 calculated with the previous (−24.81 kcal/mol) and the present (−24.77 kcal/mol) hydration models. On the other hand, the effects of atomic parameters on hydration free energy become more important with the increase of the interatomic distances in chain molecules such as **001**. Therefore, the present hydration model with 34 atom types can give better result for the hydration free energy of **001** than the previous one with 25 atom types because the former can make it possible to perform the better statistical fitting for the atomic parameters than the latter.

Because the intramolecular interactions between solute atoms (self-solvation) had been found to have a significant effect on the hydration of a molecule, we also tried to optimize the hydration free energy function to which a proper atomic self-solvation term was added as in the previous study [30]. Despite the extensive search for the parameter space with genetic algorithm, however, we could not obtain the four kinds of atomic parameters successfully due to the bad convergent behavior of the fitness value. Because the extension of atom types were found to be more important than the inclusion of a self-solvation term to increase the accuracy of hydration free energy function [30], we selected the hydration model with 34 atom types and three kinds of atomic parameters to apply to the blind prediction for the hydration energies of SAMPL4 molecules.

Table 2 compares the experimental and computational hydration free energies of 47 molecules in the SAMPL4 data set. Overall, the estimated hydration free energies compare reasonably well with the experimental ones with the average unsigned error of 1.21 kcal/mol. The highest discrepancies between experimental and computational results are observed for **035**, **051**, and **022**. With respect to the experimental data, it is quite unexpected that that the change of hydrophobic ethyl group in **036** to hydrophilic

Table 2 Experimental (ΔG_{exp}) and calculated (ΔG_{calc}) hydration free energies (in kcal/mol) of 47 molecules in the SAMPL4 data set

Compound	ΔG_{exp}	ΔG_{calc}	Compound	ΔG_{exp}	ΔG_{calc}
001	−23.62	−25.76	**028**	−4.29	−5.28
002	−2.49	−0.99	**029**	−1.66	−3.03
003	−4.78	−3.31	**030**	−2.29	−3.32
004	−4.45	−3.23	**032**	−7.29	−5.71
005	−5.33	−3.00	**033**	−6.96	−6.33
006	−5.26	−5.46	**034**	−5.80	−5.53
009	−8.24	−9.15	**035**	−4.68	−9.32
010	−6.24	−6.07	**036**	−5.66	−5.12
011	−7.78	−9.18	**037**	−5.94	−6.36
012	−3.75	−2.04	**038**	−3.93	−4.82
013	−4.44	−3.01	**039**	−0.85	−0.78
014	−4.09	−4.23	**041**	−5.05	−3.96
015	−4.51	−3.96	**042**	−3.13	−2.36
016	−3.2	−3.02	**043**	0.14	0.45
017	−2.53	−1.69	**044**	−5.08	−4.45
019	−3.78	−2.17	**045**	−11.53	−9.98
020	−2.78	−2.88	**046**	−9.44	−10.03
021	−7.63	−8.87	**047**	−14.21	−14.41
022	−6.78	−9.66	**048**	−11.85	−11.84
023	−9.34	−7.2	**049**	−3.16	−3.96
024	−7.43	−4.69	**050**	−4.14	−2.75
025	−5.73	−6.07	**051**	−9.53	−12.74
026	−5.31	−5.08	**052**	−2.87	−4.08
027	−4.80	−5.88			

carbonyl group in **035** leads to the increase in hydration free energy from −5.66 to −4.68 kcal/mol. The absolute and relative errors in predicting the hydration free energy of **035** amount to 4.64 kcal/mol and 99 %, respectively, despite the inclusion of two structurally similar compounds in the training set (train54 and train55 in Fig. 2). This result indicates that the inclusion of structurally similar compounds in the training set would have little effect on the accuracy in hydration free energy predictions. It is also noteworthy that the experimental hydration free energy of **051** (−9.53 kcal/mol) is almost the same as that of **046** (−9.44 kcal/mol) despite the addition of –OH moiety. Our hydration model appears to underestimate the hydration free energy values of **035**, **051**, and **022** by 4.64, 3.21, and 2.88 kcal/mol, respectively. These large deviations indicate the necessity for further improvement of the hydration free energy function and the associated atomic parameters. Because the two volume parameters (V_j and O_i^{max}) could be fully adjusted with the atomic coordinates only, future modifications should be focused on the further extension of atom types and the reoptimization of S_i parameters in such a way to reflect the local electronic structures in molecules.

The high hydration free energies of **035**, **051**, and **022** imply that the substitution of a polar group in some molecules may have an effect of making the water-solute interactions thermodynamically unfavorable. A similar phenomenon was observed in some peptidic molecules that were shown to become more hydrophobic by the addition of two polar moieties to form an intramolecular hydrogen bond [36]. The hydration free energy value of a molecule with polar groups may thus increase unexpectedly in the presence of the intramolecular hydrogen bonds. Therefore, we examined the possibility of establishing the intramolecular hydrogen bonds for 47 SAMPL4 molecules based on quantum chemical calculations. To obtain the structures of energy minima, we carried out the geometry optimizations through the density functional calculations using the Gaussian program. Figure 5 shows the structures of **035**, **051**, and **022** optimized at B3LYP/6-31G** level of theory. We note that the phenolic groups of **035** and **051** establishes a strong hydrogen bond with the neighboring carbonyl group with the associated O–H···O distances of 1.76 and 1.65 Å, respectively. These intramolecular hydrogen bonds seem to have an effect of making **035** and **051** less hydrophilic by preventing the two polar groups from forming the intermolecular hydrogen bonds with water molecules. Thus, the weakening of interactions with water molecules due to the formation of intramolecular hydrogen bonds can be invoked to explain the unexpectedly high hydration free energies of **035** and **051**.

As shown in the optimized structures of **051** and **022**, the intramolecular hydrogen bonds are also established between the amino group on the phenyl ring and the neighboring carbonyl oxygen with the associated N–H···O distance of 1.85 Å. As can be inferred from the increase in interatomic distances and the decrease in bond angles, the N–H···O hydrogen bonds are established in the weaker form than the O–H···O one. This hydrogen bond weakening can be attributed to the decrease in the polarity of X–H bond due to the change of the donor atom from oxygen to the less electronegative nitrogen. Therefore, the N–H···O intramolecular hydrogen bond is expected to be broken more easily than the O–H···O one in water, which would have an effect of increasing the possibility to form the intermolecular hydrogen bonds with water. This can be invoked to explain the significant decrease in hydration free energy from −9.44 in **046** to −11.85 in **048** due to the addition of –NH_2 group (Table 2), which could be predicted accurately with the hydration free energy function and the optimized atomic parameters. Related with the structural change from **046** to **048**, the hydration free energy can become more favorable by the addition of NH_2 group due to the increase in molecular polarity caused by the delocalization of non-bond electrons on the nitrogen to the phenyl ring. On the other hand, the deviation between

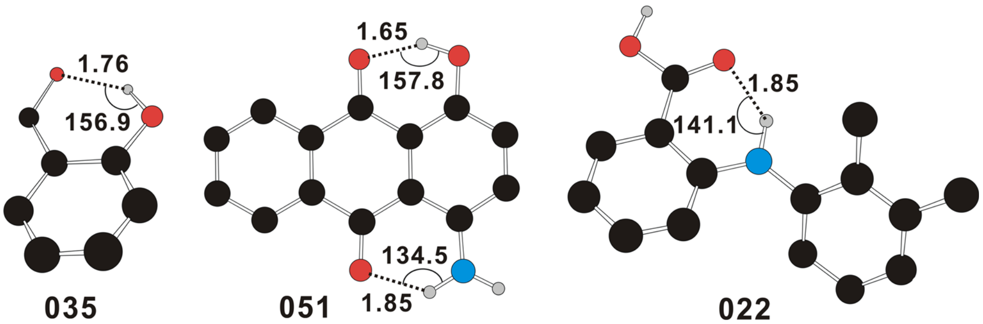

Fig. 5 The structures of **035**, **051**, and **022** optimized at B3LYP/6-31G** level of theory with PCM solvation model. Carbon, hydrogen, nitrogen, and oxygen atoms are indicated in *black*, *gray*, *blue*, and *red*, respectively. Each *dotted line* indicates a hydrogen bond. Hydrogen bond distances and angles are indicated in Å and degree, respectively. Hydrogen atoms attached to carbons are omitted for visual clarity

experimental and computational hydration free energies for **022** amounts to 2.88 kcal/mol. We note in this regard that it would be difficult for **022** to form an additional hydrogen bond with water because it includes only one anilinic hydrogen to establish the intramolecular hydrogen bond. Thus, the comparative analysis of experimental and computational results indicates that the present hydration free energy function may be improved through the introduction of new atom types for the atoms involved in the intramolecular hydrogen bond and the reoptimization of the associated atomic parameters.

Our extended solvent-contact model might be inefficient if the number of atomic parameters grows substantially as the patterns for establishing the intramolecular hydrogen bonds become complicated. However, this seems to be not the case because the effect of intramolecular hydrogen bonds on the hydration free energy is found to become significant only when they are established in a strong form such as the O–H···O hydrogen bond between phenolic and carbonyl groups. Although the intramolecular hydrogen bonds are also present in **001**, **033**, **034**, **037**, **046**, **047**, and **048**, their hydration free energies are estimated with accuracy without the additional atomic parameters (Table 2). Therefore, further extension of atomic parameters seems to be required only for the atoms involved in the intramolecular hydrogen bonds stronger than the O–H···O one between phenolic and carbonyl groups. The number of additional atomic parameters can thus be limited substantially due to the restriction for the strength of intramolecular hydrogen bonds.

Besides the implementation of intramolecular hydrogen bond effects, there are some additional methods for the improvement of the present hydration free energy function. First, the conformational diversity of organic molecules should be taken into account during the parameterization because the volumes of solvent-exposed and buried regions can vary with the conformational changes. For this purpose, molecular dynamics or Monte Carlo simulations can be applied to collect various local energy minima of the solute molecules. Second, the hydration free energy function needs to be decomposed into enthalpy and entropy terms. Because both thermodynamic quantities are experimentally accessible, the potential parameters in the enthalpic and entropic terms can be optimized independently using the corresponding experimental data. Apparently, this dual parameterization warrants the better correlation between the experimental and computational hydration free energies than the single parameterization because more diverse experimental data can be included in reference data set. Because the sign of hydration free energy is determined by the combination of enthalpic and entropic contributions, the decomposition analysis of hydration free energy can also provide thermodynamic insight into the hydration mechanism. Our future studies will focus on the further improvement of the hydration free energy function with the three above-mentioned points kept in mind.

Conclusions

We constructed a hydration free energy function through the extension of the solvent-contact model to apply to the blind prediction of the hydration free energies for organic molecules in the SAMPL4 data set. In this hydration model, we also defined 102 atomic parameters for 34 atom types to cope with a variety of chemical environments. All these parameters could be optimized with a standard genetic algorithm using the experimental hydration free energy data for 77 organic molecules and their atomic coordinates optimized at B3LYP/6-31G** level of theory. The hydration free energies estimated with the potential function and the optimized atomic parameters compared reasonably well with the experimental results with the associated R^2 value of 0.89 and the root mean square deviation of 1.46 kcal/mol for 47 SAMPL4 molecules. The comparative analysis of experimental and computational hydration free energies of the SAMPL4 molecules indicated that the present hydration model could be further improved through the extension of the atomic parameters to take into account the effect of intramolecular hydrogen bonds on the hydration free energy. Considering the simplicities in energy calculation and in model refinement, the present hydration free energy function is expected to be useful for modeling the water-solute interactions for organic molecules.

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).

References

1. Zou X, Sun Y, Kuntz ID (1999) Inclusion of solvation in ligand binding free energy calculations using generalized-born model. J Am Chem Soc 121:8033–8043
2. 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
3. Corbett PT, Leclaire J, Vial L, West KR, Wietor JL, Sanders JKM, Otto S (2006) Dynamic combinatorial chemistry. Chem Rev 106:3652–3711
4. Jorgensen WL, Duffy EM (2002) Prediction of drug solubility from structure. Adv Drug Deliv Rev 54:355–366
5. Onsagar L (1936) Electric moments of molecules in liquids. J Am Chem Soc 58:1486–1493
6. Mehler EL, Solmajer T (1991) Electrostatic effects in proteins: comparison of dielectric and charge models. Protein Eng 4:903–910
7. Gilson MK, Sharp KA, Honig BH (1988) Calculating the electrostatic potential of molecules in solution: method and error assessment. J Comput Chem 9:327–335
8. Paluch AS, Mobley DL, Maginn EJ (2011) Small molecule solvation free energy: enhanced conformational sampling using expanded ensemble molecular dynamics simulation. J Chem Theory Comput 7:2910–2918
9. Mobley DL, Bayly CI, Cooper MD, Shirts MR, Dill KA (2009) Small molecule hydration free energies in explicit solvent: an extensive test of fixed-charge atomistic simulations. J Chem Theory Comput 5:350–358
10. Kehoe CW, Fennell CJ, Dill KA (2012) Testing the semi-explicit solvation model in the SAMPL3 community blind test. J Comput Aided Mol Des 26:563–568
11. Shivakumar D, Williams J, Wu Y, Damm W, Shelley J, Sherman W (2010) Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the OPLS force field. J Chem Theory Comput 6:1509–1519
12. Beckstein O, Iorga BI (2012) Prediction of hydration free energies for aliphatic and aromatic chloro derivatives using molecular dynamics simulations with the OPLS-AA force field. J Comput Aided Mol Des 26:635–645
13. Genheden S, Mikulskis P, Hu L, Kongsted J, Söderhjelm P, Ryde U (2011) Accurate predictions of nonpolar solvation free energies require explicit consideration of binding-site hydration. J Am Chem Soc 133:13081–13092
14. Anisimov VM, Cavasotto CN (2011) Hydration free energies using semiempirical quantum mechanical Hamiltonians and a continuum solvent model with multiple atomic-type parameters. J Phys Chem B 115:7896–7905
15. Gupta M, da Silva EF, Svendsen HF (2012) Modeling temperature dependency of amine basicity using PCM and SM8T implicit solvation models. J Phys Chem B 116:1865–1875
16. Marenich AV, Cramer CJ, Truhlar DG (2008) Perspective on foundations of solvation modeling: the electrostatic contribution to the free energy of solvation. J Chem Theory Comput 4:877–887
17. Klamt A, Eckert F, Diedenhofen M (2009) Prediction of the free energy of hydration of a challenging set of pesticide-like compounds. J Phys Chem B 113:4508–4510
18. Nicholls A, Mobley DL, Guthrie JP, Chodera JD, Bayly CI, Cooper MD, Pande VS (2008) Predicting small-molecule solvation free energies: an informal blind test for computational chemistry. J Med Chem 51:769–779
19. Corbeil CR, Sulea T, Purisima EO (2010) Rapid prediction of solvation Free Energy. 2. The first-shell hydration (FiSH) continuum model. J Chem Theory Comput 6:1622–1637
20. Eisenberg D, Mclachlan AD (1986) Solvation energy in protein folding and binding. Nature 319:199–203
21. Boyer RD, Bryan RL (2012) Fast estimation of solvation free energies for diverse chemical species. J Phys Chem B 116:3772–3779
22. Sergiievskyi VP, Fedorov MV (2012) 3DRISM multi-grid algorithm for fast solvation free energy calculations. J Chem Theory Comput 8:2062–2070
23. Setny P, Zacharias M (2010) Hydration in discrete water. A mean field, cellular automata based approach to calculating hydration free energies. J Phys Chem B 114:8667–8675
24. Bernazzani L, Duce C, Micheli A, Mollica V, Tine MR (2010) Quantitative structure − property relationship (QSPR) prediction of solvation Gibbs energy of bifunctional compounds by recursive neural networks. J Chem Eng Data 55:5425–5428
25. Almlöf M, Carlsson J, Åqvist J (2007) Improving the accuracy of the linear interaction energy method for solvation free energies. J Chem Theory Comput 3:2162–2175
26. Marenich AV, Cramer CJ, Truhlar DG (2009) Performance of SM6, SM8, and SMD on the SAMPL1 test set for the prediction of small-molecule solvation free energies. J Phys Chem B 113:4538–4543
27. Colonna-Cesari F, Sander C (1990) Excluded volume approximation to protein-solvent interaction. The solvent contact model. Biophys J 57:1103–1107
28. Stouten PFW, Frömmel C, Nakamura H, Sander C (1993) An effective solvation term based on atomic occupancies for use in protein simulations. Mol Simul 10:97–120
29. Kang H, Choi H, Park H (2007) Prediction of molecular solvation free energy based on the optimization of atomic solvation parameters with genetic algorithm. J Chem Inf Model 47:509–514
30. Choi H, Kang H, Park H (2013) New solvation free energy function comprising intermolecular solvation and intramolecular self-solvation terms. J Cheminformatics 5:8
31. Rizzo RC, Aynechi T, Case DA, Kuntz ID (2006) Estimation of absolute free energies of hydration using continuum methods: accuracy of partial charge models and optimization of nonpolar contributions. J Chem Theory Comput 2:128–139
32. Geballe MT, Skillman AG, Nicholls A, Guthrie JP, Taylor PJ (2010) The SAMPL2 blind prediction challenge: introduction and overview. J Comput Aided Mol Des 24:259–279
33. Marenich AV, Cramer CJ, Truhlar DG (2013) Generalized Born solvation model SM12. J Chem Theory Comput 9:609–620
34. 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

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

36. Rafi SB, Hearn BR, Vedantham P, Jacobson MP, Renslo AR (2012) Predicting and improving the membrane permeability of peptidic small molecules. J Med Chem 55:3163–3169