J Comput Aided Mol Des (2016) 30:1079–1086
DOI 10.1007/s10822-016-9951-y

Calculation of distribution coefficients in the SAMPL5 challenge from atomic solvation parameters and surface areas

Diogo Santos-Martins[1] · **Pedro Alexandrino Fernandes**[1] · **Maria João Ramos**[1]

Received: 21 June 2016 / Accepted: 21 August 2016 / Published online: 1 September 2016

Abstract In the context of SAMPL5, we submitted blind predictions of the cyclohexane/water distribution coefficient (D) for a series of 53 drug-like molecules. Our method is purely empirical and based on the additive contribution of each solute atom to the free energy of solvation in water and in cyclohexane. The contribution of each atom depends on the atom type and on the exposed surface area. Comparatively to similar methods in the literature, we used a very small set of atomic parameters: only 10 for solvation in water and 1 for solvation in cyclohexane. As a result, the method is protected from overfitting and the error in the blind predictions could be reasonably estimated. Moreover, this approach is fast: it takes only 0.5 s to predict the distribution coefficient for all 53 SAMPL5 compounds, allowing its application in virtual screening campaigns. The performance of our approach (submission 49) is modest but satisfactory in view of its efficiency: the root mean square error (RMSE) was 3.3 log D units for the 53 compounds, while the RMSE of the best performing method (using COSMO-RS) was 2.1 (submission 16). Our method is implemented as a Python script available at https://github.com/diogomart/SAMPL5-DC-surface-empirical.

Keywords SAMPL5 · Drug design data resource · D3R · Solvent accessible area · Free energy of solvation · Distribution coefficient

Introduction

The free energy of solvation ΔG^{solv} can be separated in (1) cavitation free energy and (2) solute–solvent interaction free energy. The cavitation free energy corresponds to the cost of disrupting solvent–solvent interactions in order to create a cavity that accommodates the solute. Solute–solvent interactions include van der Waals interactions and electrostatic interactions. Hydrogen bonds can be treated separately or within the general framework of electrostatic interactions. The assumption that cavitation and solute–solvent interaction free energies are additive provides a simple framework where the balance between these two terms rationalizes observed phenomena. For example, the hydrophobic effect observed for apolar solutes in water results from the high cost of forming a cavity (which includes the entropic penalty associated with constrained water molecules) and lack of counterbalancing strong solute–water interactions.

The solute–solvent interaction energy is mostly determined by the first layer of solvent molecules and by exposed solute atoms, simply because atoms in close proximity make the largest vdW and electrostatic contribution (charged buried atoms, such as in transition metal complexes, may be exceptions to this general rule). Moreover, if the solvent has hydrogen bond donors/

Electronic supplementary material The online version of this article (doi:10.1007/s10822-016-9951-y) contains supplementary material, which is available to authorized users.

✉ Diogo Santos-Martins
diogom@fc.up.pt

Pedro Alexandrino Fernandes
pafernan@fc.up.pt

Maria João Ramos
mjramos@fc.up.pt

[1] UCIBIO, REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, 4169-007 Porto, Portugal

acceptors, only exposed solute atoms can participate in hydrogen bonds with solvent molecules. For this reason, computational definitions of surface area have found application in the calculation of solute–solvent interactions, either by applying rigorous electrostatic formalisms as in the Poisson–Boltzman equation, or simply to estimate the contribution of different solute atoms in empirical models, as is the case of the present work.

The free energy of solvation of a molecule can be predicted as the sum of the individual contribution of each solute atom, weighted by its exposed surface area and by an atomic solvation parameter associated with its atom type [1]. Despite its simplicity, this formalism has been reported multiple times in scientific publications. Table 1 provides a comparison of published models used to calculate the free energy of solvation in water. Most studies employ a large number of parameters allowing the model to adhere very well to experimental data. In the work of Boyer et al. [2] a total of 84 parameters were fitted, leading to a mean absolute error (MAE) of 1.41 kcal/mol. Other publications report extremely low MAE's achieving 0.54 kcal/mol by Wang et al. [3] and 0.65 kcal/mol by Hou et al. [4]. However, using a large number of parameters relatively to the size of the training set makes the model susceptible to overfitting. Ooi et al. [5] fitted 7 parameters using only 22 molecules for training, and reported an extremely low root mean square error (RMSE) for compounds in the training set (RMSE = 0.32 kcal/mol) but a significantly larger error when the model was tested on molecules outside the training set (RMSE = 2.0 kcal/mol). For this reason, in this work we used a reduced number of parameters and a large training set.

These models have been used to predict the solvation free energy of different solute conformations [5]. This is possible because surface areas effectively capture the solvent exposure of each solute atom, preventing shielded atoms (e.g. after intramolecular hydrogen bonding) from contributing to the hydration free energy of the conformer. Moreover, atomic solvation parameters and surface areas have also been used to calculate partition coefficients [6] and aqueous solubilities [7], and have been integrated into both molecular dynamics [8] and molecular docking [9].

The existence of approximate but computationally inexpensive methods enables the large scale prediction of free energies of solvation. The question is: how does the performance of empirical models compare to more physical models? In the previous edition of the SAMPL challenge (SAMPL4), one of the blind predictions of the hydration free energies was an empirical model that performed almost as well as the more physically grounded methods [10, 11]. Instead of using surface areas to estimate solvent exposure of solute atoms, the proximity of other solute atoms from the atom of interest was taken into account. A total of 34 atom types were defined, but the total number of fitted parameters was 102 (68 parameters were used to describe shielding effects and quantify solvent exposure of solute atoms).

In this work we built empirical models to predict the free energy of solvation of organic compounds in water and cyclohexane, where the contribution of each solute atom is weighted by its exposed surface area and an atomic solvation parameter specific to its atom type. Then, we used these models to predict the cyclohexane/water distribution coefficient of 53 SAMPL5 molecules (depicted in Figure S2) for which experimental log D values have been calculated [12, 13]. Despite the reduced number of atomic solvation parameters (10 for the free energy of solvation in water and 1 for the free energy of solvation in cyclohexane), our method performed reasonably. Unsurprisingly, more physical methods made better log D predictions (see the SAMPL5 overview paper [14]), but the computational efficiency of our approach makes it valuable for large scale applications.

Methods

Cyclohexane/water distribution coefficients (D) were calculated from the free energies of solvation in each solvent, according to the following equation:

Table 1 Comparison of quality of fit (training errors) for ΔG^{solv}_{water} for several models found in the literature and for the one proposed in this work

	Solvent radius (Å)	Type of surface	Partial charges	Fitted parameters	Dataset size	MAE	RMSE
Ooi [5]	1.4	SAS	–	7	22	–	0.32
Wang [3]	0.6	SAS	–	54	401	0.54	0.79
Hou [4]	0.5	SAS	–	58	415	0.65	0.75
Boyer [2]	1.4	SAS	RESP	84	596	1.41	–
This work	1.5	SES	Gasteiger–Marsili	10	642	1.25	1.69

SAS solvent accessible surface, *SES* solvent excluded surface, *MAE* mean absolute error and *RMSE* root mean squared error, both presented in kcal/mol

$$log\, D = \frac{\Delta G^{solv}_{water} - \Delta G^{solv}_{cyclohexane}}{2.303RT} \quad (1)$$

where T is the temperature (293 K) and R is the ideal gas constant (1.9872 cal/mol K). The chosen temperature value is approximate: some training molecules had their solvation free energies determined at 298 K while others were studied at 293 K. The following sections describe the calculation of free energies of solvation in water and in cyclohexane.

Throughout this work, the following simplifications were adopted: (1) molecules were used in the conformation provided by SAMPL5 organizers and no conformational sampling was performed; (2) only a single protonation state was considered for each molecule, corresponding to a neutral state, and ignoring different tautomeric states.

Free energy of solvation in water (hydration)

The free energy of hydration (ΔG^{solv}_{water}) was calculated as the sum of atomic contributions over all solute atoms. The contribution of an individual atom depends on the atomic solvation parameter, and on its solvent exposure:

$$\Delta G^{solv}_{water} = \sum_{i}^{N} W_i \times S_i \quad (2)$$

where N is the number of solute atoms, W_i is the atomic solvation parameter of the ith atom and S_i is the solvent exposure of the ith atom. Solvent exposure was calculated either as the solvent accessible surface (SAS) area or the solvent excluded surface (SES) area, computed with MSMS [15] using a solvent probe radius of 1.5 Å. The van der Waals radii for solute atoms are listed in Table 2. The difference between SAS and SES is illustrated in Fig. 1.

In an alternative formalism, we included atomic partial charges for the calculation of free energies of hydration, using the Gasteiger–Marsili [16] method implemented in Openbabel 2.3.2 [17, 18]. The contribution of partial charges is also weighted by the solvent exposure of each solute atom, and is implemented by an additional term relatively to Eq. 2:

Table 2 Van der Waals radii used in this work

Element	vdW radius (Å)	Element	vdW radius (Å)
H	1.20	P	1.80
C	1.70	S	1.80
N	1.55	Cl	1.75
O	1.52	Br	1.85
F	1.47	I	1.98

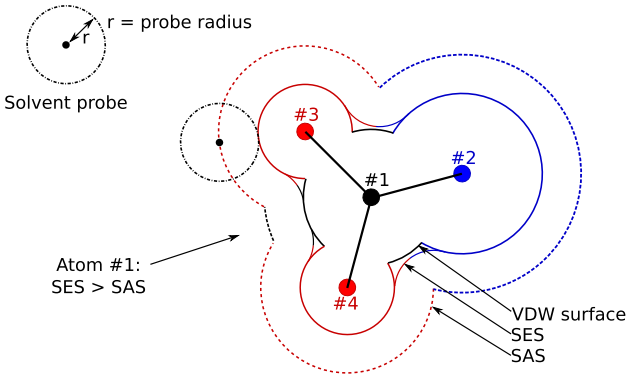

Fig. 1 Solvent accessible surface (SAS) and solvent excluded surface (SES). SES and SAS are both computed by rolling the probe sphere over the van der Waals surface of the molecule. The SAS is determined by the center of the probe, while the SES is determined by the surface of the probe. The SAS is generally larger than the SES, but the SES of buried atoms can be larger that the SAS. In this example, atom #1 is only solvent accessible on the left side between atoms #3 and #4, where its SES is larger than its SAS

$$\Delta G^{solv}_{water} = \sum_{i}^{N} W_i \times S_i + Q \sum_{i}^{N} |q_i| S_i \quad (3)$$

where q_i is the partial charge of the ith atom and Q is the weight factor for the contribution of partial charges to hydration free energy. Equations 2 and 3 provide two alternative models, including or excluding atomic charges.

Training procedure

Atomic solvation parameters (W_i and Q) were fitted by the least squares method to reproduce the experimental free energy of hydration of 642 compounds in the FreeSolv-0.32 database [19, 20], using the R software package [21]. Some atom types displayed poor statistical significance and were manually set to zero. This is either because there are few molecules in the training set containing these atom types or because they are often buried (e.g. phosphorous in phosphate groups). In different models (SES or SAS, with or without partial charges), the excluded atom types varied. The formalism presented in equations 2 and 3 lacks a term to explicitly describe the cost of creating a cavity in water to accommodate the solute. However, since the atomic solvation parameters are fitted to experimental free energies of solvation, the cost of cavity formation is implicitly incorporated into the atomic solvation parameters.

Atom types

We devised a simple atom typing scheme that resulted in a reduced number of atomic solvation parameters. Atom types indicate three attributes: (i) the element, (ii) aromaticity (iii) the possibility of making hydrogen bonds

with solvent waters. Both the aromaticity and hydrogen bonding were predicted by Openbabel 2.3.2. The resulting atom types are shown in Table 3. There are two types for hydrogen atoms (polar and apolar hydrogens), two types of carbon (aromatic and non-aromatic carbon), four types of nitrogen (aromatic / non-aromatic, able / unable to accept hydrogen bonds). Oxygen has a single atom type, thus all oxygens are typed as O. All oxygens in the FreeSolv-0.32 database and in the SAMPL5 set are considered H-bond acceptors by Openbabel 2.3.2. The remaining elements are composed of a single atom type each. Chemical groups which are H-bond donors, such as hydroxyl groups and amines, rely on the presence of polar hydrogens (HD) to describe their H-bond donor properties.

Free energy of solvation in cyclohexane

For training the model to predict $\Delta G^{solv}_{cyclohexane}$ we used a total of 18 compounds with experimental values [22]. These compounds (Figure S3) are sidechain analogues of the 20 naturally occurring aminoacids except glycine and proline. Due to the reduced number of compounds used for training the model, we opted to fit only a single parameter: the SES area of the molecule. The free energy of solvation in cyclohexane is then calculated as:

$$\Delta G^{solv}_{cyclohexane} = W_c \times A \quad (4)$$

where W_c is the fitted parameter (using the least squares method) and A is the SES area of the solute, using a solvent probe radius of 1.5 Å. The value of W_c and the quality of the model are discussed in the results section.

Results and discussion

In the following sections, we discuss (1) the model for predicting free energies of solvation in water, (2) the model for predicting free energies of solvation in cyclohexane and (3) the blind prediction of cyclohexane/water distribution coefficients (log D) for SAMPL5 compounds.

Prediction of free energy of solvation in water

Atomic solvation parameters were fitted using the least squares method to reproduce the experimental free energies of hydration of 642 molecules in the FreeSolv-0.32 database [19, 20]. In order to evaluate the benefit of using partial charges calculated by a fast method (Gasteiger–Marsili) and also to test different surfaces (SES and SAS) to quantify solvent exposure of solute atoms, four sets of atomic solvation parameters were derived. The resulting parameters are reported in Table 3. The quality of

Table 3 Atomic solvation parameters used in the calculation of the free energy of hydration, fitted to experimental data in the FreeSolv-0.32 database using the least squares approach

Atom type	Description	Atomic solvation parameters (W_i) (cal/mol Å^2)			
		SES (Eq. 3)	SAS (Eq. 3)	SES (Eq. 2)	SAS (Eq. 2)
H	Apolar hydrogen	+11.2 (±1.6)	+3.2 (±0.7)	+8.1 (±2.3)	−2.0 (±0.5)
HD	Polar hydrogen	−193.5 (±13.3)	−45.7 (±5.1)	−303.3 (±13.0)	−95.5 (±4.4)
C	Carbon	0	+21.2 (±5.8)	−44.6 (±9.8)	0
A	Carbon (arom.)	−12.6 (±3.1)	0	−40.7 (±3.0)	−22.4 (±2.2)
N	Nitrogen	0	0	+144.9 (±38.4)	0
NA	Nitrogen (arom.)	−626.6 (±65.4)	−465.2 (±52.5)	−621.5 (±71.5)	−497.6 (±59.0)
NH	Nitrogen (H-bond acc.)	−128.7 (±22.4)	−24.6 (±8.3)	−130.9 (±24.4)	−47.7 (±9.1)
NHA	Nitrogen (arom./H-bond acc.)	−185.6 (±21.7)	−98.6 (±11.9)	−244.3 (±22.8)	−124.9 (±13.3)
O	Oxygen (H-bond acc.)	−42.2 (±7.2)	−10.9 (±2.8)	−108.8 (±4.5)	−46.8 (±2.2)
F	Fluorine	+72.4 (±6.5)	+38.9 (±2.8)	+31.7 (±5.7)	+11.3 (±2.6)
P	Phosphorus	0	0	0	0
S	Sulfur	0	0	0	−15.7 (±5.1)
Cl	Chlorine	+13.9 (±2.8)	+8.4 (±1.3)	−7.1 (±2.1)	−5.1 (±1.0)
Br	Bromine	0	0	0	0
I	Iodine	0	0	0	0
Weight factor for Gasteiger charges (Q)		−246.9 (±22.1)	−162.0 (±9.2)	–	–
Training RMSE (kcal mol^{-1})		1.69	1.76	1.80	1.99

Atomic solvation parameters correspond to W_i in Eqs. 2 and 3 and to Q in Eq. 3. Zeroed parameters were set manually due to poor statistics

prediction using SES areas and including atomic charges is depicted in Fig. 2.

The magnitude and sign of atomic solvation parameters quantifies the contribution of each atom type to the free energy of hydration. More negative values indicate a larger favorable contribution to the hydration free energy. However, the contribution of an individual atom is weighted by its surface area, explaining the larger magnitude of solvation parameters obtained using SES areas (the SAS is always larger than the SES except for highly buried atoms, as is exemplified for atom #1 in Fig. 1).

Atom types capable of making Hydrogen bonds with water have the most negative solvation parameters (HD, O, NH, NHA). This means that solute–solvent hydrogen bonds can be captured by atomic solvation parameters. The coefficient Q from Eq. 3 scales the contribution of partial charges and also has a large negative value, indicating that the contribution of electrostatic interactions have also been incorporated in the parameters. This observations are consistent with a physically meaningful model, with a straightforward interpretation of atomic solvation parameters. It is important to note that inclusion of partial charges in the model (Eq. 3) decreases the magnitude of the atomic solvation parameter of atoms involved in hydrogen bonds (HD, O, NH, NHA) by about fourfold if SES areas are used and threefold if SAS areas are used. This means that Gasteiger–Marsili charges are able to describe a significant part of solute–solvent hydrogen bonds.

One particular atom type, NA (aromatic nitrogen that does not accept H-bonds) displays a more negative solvation parameter than atom types involved in hydrogen bonds, which is hard to rationalize. This is partially explained by the low exposure of NA atoms, which are shielded by three substituent groups in a planar geometry, making solvent contacts possible only in small surface patches above and below the plane of the aromatic ring. However, even considering their low solvent exposure, NA atoms can make significant contributions: for cyanuric acid (the molecule from FreeSolv-0.32 database with the largest SES area associated with NA atoms), NA atoms contribute with almost −8 kcal/mol to the hydration free energy. For comparison, the contribution from hydrogen bond donors/acceptors and from partial charges is about −15.3 kcal/mol for cyanuric acid, and the free energy of hydration is overestimated by −5.6 kcal/mol. These observations suggest that the parameter for NA has overfitted. We'll return to this discussion in view of the results obtained in blind log D prediction for SAMPL5 compounds containing NA atoms.

Among the four sets of parameters derived to predict ΔG^{solv}_{water}, the quality of the fit was slightly better (lower RMSE) with the use of SES areas and the inclusion of partial charges. Thus, this model was used to make blind predictions of the cyclohexane/water log D for compounds in the SAMPL5 set.

Prediction of free energy of solvation in cyclohexane

Free energies of solvation in cyclohexane were predicted using Eq. 4, in which a single parameter W_c is multiplied by the SES area of the solute to obtain $\Delta G^{solv}_{cyclohexane}$. From a physical point of view, we are assuming that the free energy of solvation is directly proportional to the solute area. This assumption is reasonable because the dielectric constant of cyclohexane is very low ($\epsilon = 2.02$), and van der Waals interactions constitute the largest contribute to intermolecular stabilization. Using a set of 18 molecules, W_c was fitted to −36 cal mol^{-1} Å^{-2}. The quality of the fit is depicted in Fig. 3, and has a RMSE of 1.02 kcal/mol. On an additional set of 91 molecules from Ref. [23], the RMSE is 1.07 kcal/mol (see Figure S1 and Table S1). The model systematically underestimates free energies of solvation for more negative values and overestimates more

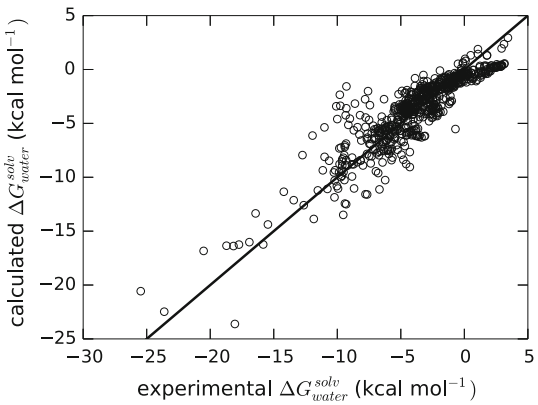

Fig. 2 Prediction of hydration free energies for molecules in the training set using SES areas and including partial charges

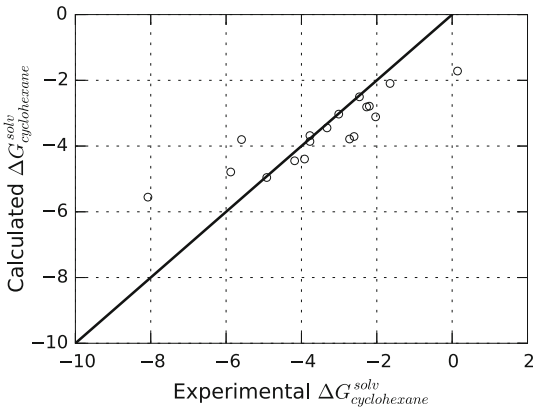

Fig. 3 Prediction of solvation free energies in cyclohexane for molecules in the training set

positive values. This bias could be fixed by introducing an intercept term B in Eq. 4 and transforming it into $\Delta G^{solv}_{cyclohexane} = W_c \times A + B$. However, the presence of an intercept term B would mean that a molecule with no surface area would have an interaction with cyclohexane, which is physically unreasonable. For this reason, we decided to avoid the use of an intercept. Moreover, in a retrospective analysis, we used an intercept in the model to predict $\Delta G^{solv}_{cyclohexane}$ but this showed no improvement in the prediction of log D values for SAMPL5 compounds, and showed a systematic bias for larger molecules, indicating that Eq. 4 without intercept is more appropriate to calculate $\Delta G^{solv}_{cyclohexane}$.

Prediction of log D for SAMPL5 compounds

The prediction of log D values for compounds in the SAMPL5 challenge was based on the free energies of solvation in water and on cyclohexane (Eq. 1). The free energy of solvation in water was calculated using SES areas and partial charges (Eq. 3). The model that predicts the free energy of solvation in cyclohexane consists of a single coefficient multiplied by the SES area of the molecule (Eq. 4). Our predictions are reported in Table S2.

Figure 4 compares the calculated and experimental log D values for the 53 SAMPL5 molecules. While a correlation is readily observable the model exaggerates the magnitude of the predictions, both for negative and positive valued log D's. In other words, if the calculated log D was scaled by a factor of about 0.3, the predictions would approach the equality line. The key parameters to describe the quality of the prediction are the Pearson's correlation coefficient of 0.58, the Kendall rank order correlation coefficient of 0.42, a mean signed error of −1.06 (1.42 kcal/mol in ΔG^{solv} units), a mean absolute error (MAE) of 2.57 (3.45 kcal/mol) and a root mean square error (RMSE) of 3.27 (4.39 kcal/mol). These values correspond to modest prediction of log D values. The Kendall rank order correlation coefficient (0.42) also indicates modest performance in ranking the compounds.

The largest outlier is adenosine (ID: SAMPL5_074) which is predicted to have a log D value of −14.1 while the experimental value is −1.9. Analysis of the predictions submitted by other participants revealed a systematic bias towards more negative values. This may indicate a problem with the experimental value of this molecule, or the existence of a phenomena that lies outside the scope of the modeling techniques, such as the formation of adenosine dimers in cyclohexane, satisfying a significant number of hydrogen bond donors/acceptors. However, this is a merely speculative explanation for the systematic deviation of the predictions. If SAMPL5_074 is removed, the RMSE decreases from 3.27 (4.39 kcal/mol in ΔG^{solv} units) to 2.84 (3.80 kcal/mol), and the MAE reduced from 2.57 (3.45 kcal/mol) to 2.39 (3.20 kcal/mol).

Discussion of the NA atom type

The atomic solvation parameter for NA in the ΔG^{solv}_{water} model is the most negative among all fitted parameters, which is suspicious in view of the smaller magnitude of other parameters associated with strong interactions with water: hydrogen bonds and atomic charges. As is depicted in Fig. 5, our blind predictions on SAMPL5 compounds confirmed the suspicions: a larger contribution from NA atoms is indeed associated with a biased prediction of log D values towards distribution of the solute in water. In view of these results, we concluded that the atomic solvation parameter for atom type NA has overfitted. It is important to note that the aberrant NA parameter does not explain all errors in our model: molecules in which NA is absent still present large deviations from the experimental

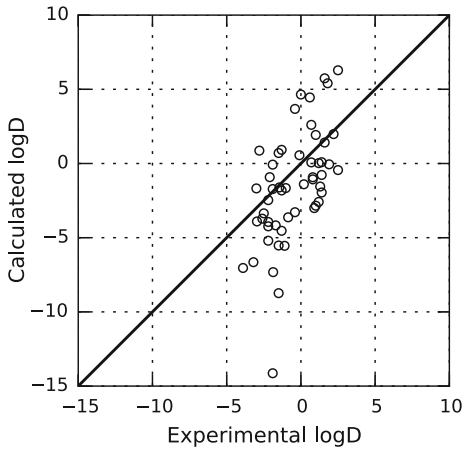

Fig. 4 Blind prediction of cyclohexane/water log D values for SAMPL5 compounds

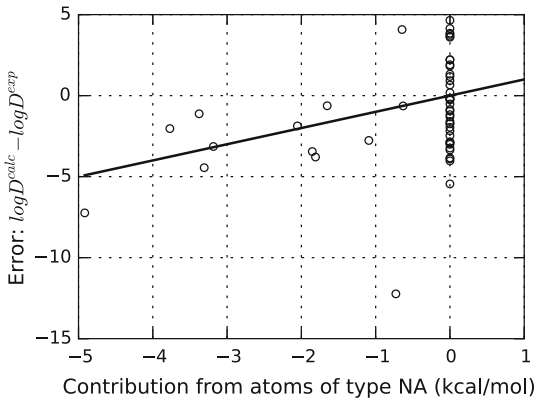

Fig. 5 Error in the blind prediction of cyclohexane/water log D values is associated with the contribution from NA atoms

value (see Fig. 5). The value of this analysis lies in the identification of an error created by a machine learning approach through interpretation of the physical meaning of solvation parameters.

In an attempt to explain the aberrant NA parameter, we performed three retrospective (non-blind) experiments, in which the cyclohexane/water log D was calculated for the 53 SAMPL5 molecules using a modified set of parameters to calculate the solvation free energy in water:

1. Set $W_{NA} = 0$ without change to any other atomic parameter.
2. Set $W_{NA} = 0$ and re-calibrate the remaining parameters using all 642 molecules in the FreeSolv-0.32 database.
3. Set $W_{NA} = 0$ and re-calibrate the remaining parameters using all molecules in the FreeSolv-0.32 database except those that contain NA (19 out of 642 molecules contain NA).

The performance of the new sets of solvation parameters were evaluated in (i) the full SAMPL5 set (n = 53), (ii) all compounds except SAMPL5_074 (n = 52), and (iii) the subset of SAMPL5 compounds that do not contain NA atoms (n=40). We note that SAMPL5_074 contains NA and is excluded in subset (iii).

The results are reported in Table 4. Overall, retrospective experiment 2 increased the error, while experiments 1 and 3 decreased the error relatively to the original set of parameters (Table 3). We speculate that NA containing molecules are implied in the origin of the aberrant NA parameter during the fitting process. Setting $W_{NA} = 0$ without change to other solvation parameters (experiment 1) lowers the error, but forcing $W_{NA} = 0$ while allowing other parameters to re-optimize (experiment 2) increases the error. Excluding NA containing molecules from the training set (experiment 3) also decreased the error. It is possible that specific chemical features in training set molecules containing NA introduce a bias in the NA parameter in order to reproduce the free energy of hydration. For exaple, the existence of multiple tautomeric states in NA containing molecules (e.g. cyanuric acid), or the induction of a dipole moment in aromatic rings by the presence of a nitrogen atom instead of a carbon atom are complex physical properties beyond the scope of the present model. Thus, the NA parameter optimizes to a meaningless value because it is prevalent in molecules that happen to contain chemical features that the present model is unable to describe.

Estimating the model error

In the submission of results to SAMPL5, participants were asked to estimate the uncertainty of the predictions. We estimated the uncertainty of our model based on the training RMSE of predictions of solvation free energy in water (1.68 kcal/mol) and cyclohexane (1.02 kcal/mol), which accumulate to 2.7 kcal/mol. We rounded up this value to 3 kcal/mol because the compounds in the SAMPL5 set are larger and chemically more diverse than those used for fitting parameters. According to Eq. 1, 3 kcal/mol correspond to 2.24 log D units. Our log D predictions displayed a RMSE of 3.27 and a mean absolute error (MAE) of 2.6, which is higher than the estimated error. If the compound with ID SAMPL5_074 is excluded (this compound was systematically predicted to have a lower log D by other SAMPL5 participants), the RMSE lowers to 2.84 and the MAE to 2.39. which is not far from our RMSE estimate of 2.24. Overall, the errors in the blind challenge were higher than we have anticipated, but the error estimate is reasonable.

Conclusions

In this work, we employed an empirical model based on atomic solvation parameters and on the surface area of exposed solute atoms to predict the free energies of solvation in two solvents: water and cyclohexane. This approach was used to make blind predictions of the cyclohexane/water distribution coefficients of 53 molecules in the context of the SAMPL5 challenge. Our predictions were not among the best performing methods in the challenge, but can be considered satisfactory in view of its speed: it takes an average of 0.01 s per molecule.

The most striking feature of this work relatively to similar studies is the reduced number of atomic solvation parameters. Typically, the number of atom types ranges between 30 and nearly 100, but here we have fitted parameters for only 10 atom types (for predicting free

Table 4 RMSE (log D units) evaluated on SAMPL5 compounds using updated atomic solvation parameters from retrospective experiments

	Evaluation set		
	All SAMPL5 (n = 53)	Except 074 (n = 52)	NA free (n = 40)
Experiment 1	2.97	2.52	2.63
Experiment 2	3.33	2.82	2.92
Experiment 3	2.89	2.45	2.55
Submission #49	3.27	2.84	2.63

energies of hydration). Thus, our model can capture only simple features of the solute–solvent interaction, such as hydrogen bonds, but in compensation has a straightforward interpretation of the physical meaning of atomic solvation parameters and is less susceptible to overfitting. As a result, the error in the blind predictions is only slightly higher than the errors obtained in the fitting stage.

Acknowledgments We acknowledge European Union (FEDER funds POCI/01/0145/FEDER/007728) and National Funds (FCT/MEC, Fundação para a Ciência e Tecnologia and Ministério da Educação e Ciência) under the Partnership Agreement PT2020 UID/MULTI/04378/2013.UID/MULTI/04378/2013; NORTE-01-0145-FEDER-000024, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF) D.S.M thanks Fundação para a Ciência e Tecnologia for scholarship SFRH/BD/84922/2012.

References

1. Eisenberg D, McLachlan AD (1986) Nature 319(6050):199
2. Boyer RD, Bryan RL (2012) J Phys Chem B 116(12):3772
3. Wang J, Wang W, Huo S, Lee M, Kollman PA (2001) J Phys Chem B 105(21):5055
4. Hou T, Qiao X, Zhang W, Xu X (2002) J Phys Chem B 106(43):11295
5. Ooi T, Oobatake M, Nemethy G, Scheraga HA (1987) Proc Natl Acad Sci 84(10):3086
6. Pei J, Wang Q, Zhou J, Lai L (2004) Proteins Struct Funct Bioinform 57(4):651
7. Wang J, Krudy G, Hou T, Zhang W, Holland G, Xu X (2007) J Chem Inf Model 47(4):1395
8. Kleinjung J, Scott WR, Allison JR, van Gunsteren WF, Fraternali F (2012) J Chem Theory Comput 8(7):2391
9. Huang SY, Zou X (2010) J Chem Inf Model 50(2):262
10. Park H (2014) J Comput Aided Mol Des 28(3):175
11. Mobley DL, Wymer KL, Lim NM, Guthrie JP (2014) J Comput Aided Mol Des 28(3):135
12. Rustenburg AS, Dancer J, Lin B, Feng JA, Ortwine DF, Mobley DL, Chodera JD (2016) bioRxiv 063081. doi:10.1101/063081
13. Lin B, Pease JH (2013) Comb Chem High Throughput Screen 16(10):817
14. Bannan CC, Burley KH, Chiu M, Gilson MK, Mobley DL (2016) J Comput Aided Mol Des (in prep)
15. Sanner MF, Olson AJ, Spehner JC (1996) Biopolymers 38(3):305
16. Gasteiger J, Marsili M (1978) Tetrahedron Lett 34:3181
17. OLBoyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) J Cheminform 3:33
18. O'Boyle NM, Morley C, Hutchison GR (2008) Chem Cent J 2(5). doi:10.1186/1752-153X-2-5
19. Mobley DL, Guthrie JP (2014) J Comput Aided Mol Des 28(7):711
20. Mobley DL (2013) Experimental and calculated small molecule hydration free energies. Retrieved from http://www.escholarship.org/uc/item/6sd403pz
21. R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org. ISBN 3-900051-07-0
22. Villa A, Mark AE (2002) J Comput Chem 23(5):548
23. Marenich AV, Cramer CJ, Truhlar DG (2009) J Phys Chem B 113(18):6378