J Comput Aided Mol Des (2014) 28:211–219
DOI 10.1007/s10822-014-9725-3

Predicting hydration free energies with chemical accuracy: the SAMPL4 challenge

Lars Sandberg

Received: 15 November 2013 / Accepted: 30 January 2014 / Published online: 19 February 2014

Abstract An implicit solvent model described by a non-simple dielectric medium is used for the prediction of hydration free energies on the dataset of 47 molecules in the SAMPL4 challenge. The solute is represented by a minimal parameter set model based on a new all atom force-field, named the liquid simulation force-field. The importance of a first solvation shell correction to the hydration free energy prediction is discussed and two different approaches are introduced to address it: either with an empirical correction to a few functional groups (alcohol, ether, ester, amines and aromatic nitrogen), or an ab initio correction based on the formation of a solute/explicit water complex. Both approaches give equally good predictions with an average unsigned error <1 kcal/mol. Chemical accuracy is obtained.

Keywords Hydration free energy predictions · SAMPL4 · Implicit solvent model · Modified Langevin–Debye model · LSFF · Molhydro

Introduction

Accurate predictions of molecular solvation free energies are essential in describing the many processes that take place in solutions, e.g. equilibrium partition ratios of compounds in two immiscible solvents, chemical reactions and protein-ligand binding [1]. The solute molecule is usually represented at atomic resolution, while the solvent is represented either by an explicit or an implicit solvent model. This corresponds to either a particle or a field-theoretic modelling approach to the complex many-body problem of a solute in solution. Today implicit solvent models are widely used with great success to predict thermodynamic equilibrium properties of solubility processes. However, implicit solvent models are usually heavily parameterized and the parameters differ, sometimes significantly, from those used in explicit solvent models, and maybe more seriously, from those found experimentally. Indeed this questions the basis for applying a physical representation as the one used in implicit solvent modelling. For most approaches is it difficult to pin-point where the model uncertainties arise from. Is it from the solvent representation, that is the dielectric continuum approximation, or from the representation of the solute molecule, e.g. the atomic parameters, or from both? In this paper we will apply an approach which will make this much clearer, with the aim of having an interpretable model that is fully transferable and independent of the choice of solvent representation.

As a rule the implicit solvent is portrayed as a *simple medium* (i.e. a linear responding, homogenous and isotropic dielectric medium), for which the thermodynamic state is described by the dielectric constant ε_r. For instance, the PB/SA method (Poisson–Boltzmann and surface-area solvation) and its heuristic next of kin, the GB/SA method (generalised Born and surface-area solvation), are both examples of a simple medium model. However, the simple medium approximation is rather coarse and not exactly valid for nanoscale molecular systems, only for macroscopic bulk systems at standard conditions. If we go beyond the simple medium approximation and extend the theory to include non-linear response effects, e.g.dielectric saturation and electrostriction, predictions will become more robust and accurate [2]. Furthermore, model

L. Sandberg (✉)
Division of Biological Chemistry and Drug Discovery College of Life Sciences, University of Dundee, Dundee, UK
e-mail: l.h.sandberg@dundee.ac.uk

parameters become transferable between implicit and explicit solvent models [3]. In a non-simple dielectric medium the dielectric constant is replaced by a relative permittivity ε_r that is a function of the electric field strength, i.e. $\varepsilon_r(\|\mathbf{E}\|)$. We have previously developed and implemented a non-linear dielectric continuum model which solves the stationary Maxwell's equations rigorously for the solute-solvent interaction term [2, 3].

The aim for the present study is to apply an implicit solvation model that meet the following requirements for calculating the hydration free energy of the SAMPL4 example molecules [4, 5], i.e. the method

1. includes all physical effects liable to hydration free energy calculations
2. is applicable to molecules found in biological and organic chemistry
3. introduces as few parameters and additive corrections, derived from experimental hydration free energy data, as possible
4. parameters are fully transferable and calculations are comparable to, or better than, the accuracy of the golden standard of explicit solvent free energy calculations
5. predictions are made within chemical accuracy, commonly defined as thermochemical predictions within $\pm$ 1 kcal/mol.

Methods

The transfer process of a solute molecule between an ideal gas phase and a water solution phase is described by the hydration free energy. Free energies of solvation are calculated under the convention that the gas and solvent standard states have equal solute molar concentration. The hydration free energy of a solute molecule M is calculated using a thermodynamic cycle scheme:

$$\Delta G_{\text{hyd}}(\mathsf{M}) =: \Delta U_{\text{ind}}(\mathsf{M}) + \Delta G_{\text{hyd}}(\mathsf{M}^*), \quad (1)$$

where M^* denotes the polarized molecule, and the induction energy is the intramolecular reorganization energy of polarising the solute molecule in the local reaction-field $\mathbf{F}$ of the solvent. Hence

$$\Delta U_{\text{ind}} =: \langle\psi(\mathbf{F})|\mathcal{H}|\psi(\mathbf{F})\rangle - \langle\psi(\mathbf{0})|\mathcal{H}|\psi(\mathbf{0})\rangle, \quad (2)$$

where $\mathcal{H}$ is the Hamiltonian, $\psi(\mathbf{0})$ is the wave function in the gas phase and $\psi(\mathbf{F})$ is in the solution.

According to classical thermodynamics and the isothermal-isobaric ensemble the Gibbs free energy of solvation ΔG_{hyd} is equal to the change in hydration enthalpy minus the change in entropic contribution, i.e.

$$\Delta G_{\text{hyd}}(\mathsf{M}^*) \equiv \Delta H_{\text{hyd}} - T\Delta S = \Delta U_{\text{int}} + p\Delta V - T\Delta S. \quad (3)$$

The thermodynamic internal energy ΔU_{int} is the intermolecular (solute-solvent) interaction potential energy term, and $p\Delta V$ is the pressure-volume work. The last two terms are often combined into one new term which describes the reversible work of creating a cavity to host the solute in the solvent, the cavity formation term, and hence

$$\Delta G_{\text{hyd}}(\mathsf{M}^*) =: \Delta U_{\text{int}} + \Delta G_{\text{cav}}. \quad (4)$$

We will begin by looking more closely at the first term and later return to the cavity formation term.

The intermolecular potential energy term

The intermolecular potential energy of a classical fluid can be further decomposed into three independent terms

$$\Delta U_{\text{int}}^{\text{solute-solvent}} =: \Delta U_{\text{elec}} + \Delta U_{\text{rep}} + \Delta U_{\text{disp}}. \quad (5)$$

representing the electrostatic energy, the repulsion energy due to exchange interactions, and the intermolecular energy due to London dispersion interactions. The electrostatic energy term is the difference between the electrostatic work of charging the molecule in the gas phase and the water solution

$$\Delta U_{\text{elec}} = W_{\text{e}}(\text{water}) - W_{\text{e}}(\text{vacuum}), \quad (6)$$

where the electrostatic work of charging is calculated as

$$W_{\text{e}} = \int_V d^3r \int_0^E \varepsilon_0\varepsilon_r(E)EdE, \quad (7)$$

where V is the volume of the solvent (i.e. $V = \mathbb{R}^3 \backslash V_{\text{solute}}$) and ε_0 is the permittivity of free space. The electric field E is found by solving the stationary Maxwell's equations describing the solvent dielectric response to the applied electric field of the polarised solute molecule M^* [2, 3]. The relative permittivity ε_r is constant and equal to one in the gas phase. In the water solution we go beyond the simple medium approximation and include two non-linear response effects that are important when the local electric field becomes stronger. First we allow for the lowering of the relative permittivity due to normal dielectric saturation which occur when the thermally fluctuating water dipoles align in the electric field of the solute. Secondly, we include the electrostriction volume change that arises from the compression of the dielectric medium due to the same electric field. The relative permittivity becomes a function of the applied electric field and is described by the modified Langevin–Debye equation[2] in combination with the electrostriction volume change described by the model of Desnoyers, Verrall and Conway[6].

Finally, the repulsion energy and the dispersion energy terms are described by a (12, 6) Lennard-Jones potential function $u_{LJ}(\mathbf{r})$ using Lorentz–Berthelot's mixing rules

$$\Delta U_{rep} + \Delta U_{disp} = \int_V d^3r \sum_{i=1}^{N} n(\mathbf{r}) u_{LJ}(\mathbf{r}), \quad (8)$$

where n is the water number density and N is the number of solute atoms. The Lennard–Jones potential is tail corrected beyond the cut-off distance of 9 Å.

The cavity formation term

At the mesoscopic scale the hydrostatic pressure-volume work contribution is negligible and the cavity formation free energy is proportional to the solvent accessible surface area of the cavity

$$\Delta G_{cav} = \gamma_\infty A_{ASA} \quad (9)$$

where $\gamma_\infty = 43.35\ \mathrm{kJ/nm^2mol}$ is the surface tension of water for a cavity of zero curvature, i.e. a planar vapor-liquid interface [7]. However, at the microscopic scale cavity corrections become important due to a more pronounced cavity curvature for smaller radii. The microscopic cavity formation free energy is described by extrapolating thermodynamic fluctuation theory of the appearance of spontaneously forming cavities. According to scaled-particle theory (SPT) it becomes a third order polynomial [8, 9]

$$\Delta G_{cav} = \kappa_0 + \kappa_1 r_c + \frac{1}{2}\kappa_2 r_c^2 + \frac{1}{6}\kappa_3 r_c^3, \quad (10)$$

where r_c is the cavity radius. The four κ coefficients are fitted to experimental hydration free energy data of inert gases, small diatomic compounds, carbonoxides and normal alkanes (methane to n-tetradecane).

The liquid simulation force-field

The water parameters for the implicit solvent model used in the SAMPL4 challenge come from experimental data at room temperature (T = 298 K), e.g. the gas phase dipole moment of the water molecule, the bulk density, and the dielectric constant [3]. However, the Lennard-Jones model parameters of water come from the TIP5P-E model, which is known to describe the dielectric properties of water well [10].

One of the basic ideas in this paper is to have a solute description which is transferable, that is independent of which representation we choose for the solvent (an implicit or an explicit solvent model). Solute geometry and atomic partial charges are found by quantum chemistry calculations. The Lennard-Jones parameters come from a newly derived all atom force-field named the liquid simulation force-field (LSFF) presented below.

Table 1 The elemental (12, 6) Lennard-Jones parameters of the liquid simulation force-field (LSFF) used in the SAMPL4 challenge

Element		σ (Å)	ε/k_B (K)
Hydrogen	H	–	–
(mid-bond position)		2.646	4.4
Carbon	C	3.682	46.9
Nitrogen	N	3.302	35.5
Oxygen	O	3.031	46.5
Chlorine	Cl	3.394	162.3

LSFF is set up as a minimal parameter set model with one atom type per element only, i.e. without any dependence on hybridization. The LSFF all-atom parameters for the elements relevant to bioorganic chemistry are derived from the symmetrical two-center Lennard-Jones plus point quadrupole pair potential (2CLJQ) model for vapor-liquid equilibria (VLE) simulations of pure substances [11]. Hydrogens are treated in the bond centric representation. The LSFF Lennard-Jones parameters are tabulated in Table 1. In this way we achieve transferability since the LSFF parameters are universal and apply equally well to both implicit and explicit solvent representations. Furthermore, the atomic radii of the LSFF also agree well with intermolecular non-bonded contact distances found in organic crystal structures (c.f. Bondi radii) [12, 13].

Molecule preparation and computational details

The SAMPL4 molecules are prepared by LigPrep [14]. The molecules are subsequently energy minimized using Macromodel and the MMFF94s force-field [15], and the five lowest energy conformers per compound are saved. These conformers are then further minimised using quantum mechanics and re-ranked by their calculated gas phase energy using Jaguar and the hybrid density functional PBE1PBE/6-311++G(d,p) [16]. To speed up the protocol no potential conformational change of the molecule when transferred between the gas and water phase is considered. The gas phase geometry is used since the principal component to the electrostatic energy in equation (6) arises from the gas phase, and not from the water phase. Hence, it is essential to allow intramolecular hydrogen bonds to be formed when searching for low energy conformers in the gas phase. Only the lowest energy conformer per molecule is kept for the induction energy calculation. The PBF water solvation model is employed together with PBE1PBE/aug-cc-pVTZ to calculate the induction energy ΔU_{ind} and the atomic partial charges are fitted to the molecular electrostatic potential (ESP).

With the induction energy, atomic partial charges and the molecular geometry as input, the hydration free energy is calculated in a separate implicit solvent programme called Molhydro which solves the Poisson equation using a finite difference method with a lattice spacing of 0.2 Å [3]. The non-linear response effects of dielectric saturation and electrostriction are both implemented in Molhydro [2, 6].

Results and discussion

Validating the implicit solvent model

How accurate is the representation of water by the above introduced implicit solvent model, and how well does the calculated energies compare with energies of explicit solvent models? There is a way to make an independent test of the accuracy of the dielectric continuum model without introducing model uncertainties from the cavity formation term or the way the solute is represented (geometry, atomic partial charges, induction energy etc.).

The implicit solvent model is in part based on the TIP5P-E water model. The validity of the dielectric continuum model is tested by comparing the calculated intermolecular potential energy of water in the implicit and the explicit solvent representation. The intermolecular potential energy of the classical fluid is obtained from the heat of vaporization of water, and it has been determined to $\Delta U_{\text{int}}^{\text{class}} = -9.92$ kcal/mol [17]. We obtain a TIP5P-E value of $\Delta U_{\text{int}} = -9.74$ kcal/mol for the implicit solvent model. This compares very well with the TIP5P-E explicit solvent value of $\Delta U_{\text{int}} = -9.78$ kcal/mol obtained in a MD simulation [10]. This precision, that is the small difference in energy between the two solvent representations, validates our implicit solvation model of water. The TIP5P-E water model is shown to be transferable and the results are independent of the solvent representation.

First solvation shell correction

Predicted hydration free energies of small organic molecules are in general very good, see Table 2. However, there are evident systematic errors for certain functional groups, e.g. alcohols, ethers, and amines. A popular way to adjust for these anomalies, especially for implicit solvent models, has in the past been to introduce more atom types and/or to reparameterise the solvent model. This has led to improved results for the existing molecular sets with known experimental hydration free energy data. However, the outcome of previous SAMPL blind challenges has clearly shown the risk of overfitting model parameters and the value of having a simple and robust model approach to describe solvation. Solvation models should be transparent and physically sound, i.e. include all physical effects that are liable to affect molecular solvation. Here we do not

Table 2 Correction terms for functional groups represented in the SAMPL4 molecule set

Name	$\Delta G_{\text{hyd}}^{\text{Exp}}$	$\Delta G_{\text{hyd}}^{(0)}$	$\Delta G_{\text{hyd}}^{(1)}$	$\Delta\Delta G_{\text{corr}}^{(1)}$	$\Delta\Delta G_{\text{corr}}^{\text{empirical}}$
Water	−6.31	−4.95	−7.74	−2.79	−1.36
Methanol	−5.06	−3.07	−5.13	−2.06	−1.99
Ethanol	−5.00	−2.82	−4.31	−1.49	−2.18
1-Propanol	−4.85	−3.07	−4.76	−1.69	−1.78
2-Propanol	−4.76	−2.98	−5.07	−2.09	−1.78
2-Methyl-2-propanol	−4.50	−3.07	−5.24	−2.17	−1.43
Phenol	−6.27	−4.48	−4.87	−0.38	−1.79
Methoxymethane	−1.89	−0.08	−1.96	−1.88	−1.81
Methoxyethane	−1.77	0.26	−1.43	−1.69	−2.03
Ethoxyethane	−1.93	0.74	−0.74	−1.48	−2.67
14-Dioxane	−5.06	−2.86	−4.94	−2.08	−2.20
Methoxybenzene	−2.46	−1.29	−1.84	−0.55	−1.17
Ethanal	−3.51	−3.50	−4.82	−1.32	−0.01
Propanal	−3.43	−2.64	−3.46	−0.82	−0.79
Butanal	−3.18	−2.64	−3.98	−1.35	−0.54
Benzaldehyde	−4.02	−3.73	−4.10	−0.37	−0.29
Propanone	−3.82	−4.42	−5.79	−1.37	0.60
2-Butanone	−3.59	−3.30	−4.28	−0.98	−0.29
2-Pentanone	−3.40	−2.47	−3.51	−1.04	−0.93
3-Pentanone	−3.29	−2.15	−2.93	−0.78	−1.14
Acetophenone	−4.58	−3.87	−3.11	0.76	−0.71
Formic acid	−7.02	−6.87	−6.53	0.35	−0.15
Acetic acid	−7.02	−7.04	−7.18	−0.14	0.02
Propanoic acid	−6.78	−6.25	−5.83	0.42	−0.53
Benzoic acid	−7.90	−6.90	−7.52	−0.62	−1.00
Methyl methanoate	−2.74	−3.38	−3.61	−0.23	0.64
Methyl ethanoate	−3.14	−4.06	−4.26	−0.20	0.92
Ethyl ethanoate	−2.99	−4.00	−4.38	−0.38	1.01
Methyl benzoate	−3.93	−3.33	−4.00	−0.67	−0.60
Methylamine	−4.63	−1.62	−4.30	−2.68	−3.01
Ethylamine	−4.50	−2.55	−5.26	−2.71	−1.95
2-Proylamine	−3.70	−2.50	−4.98	−2.47	−1.20
Tert-butylamine	−3.90	−2.66	−5.40	−2.74	−1.24
Aniline	−5.58	−5.38	−6.64	−1.26	−0.20
Dimethylamine	−4.29	−1.17	−4.91	−3.75	−3.12
Diethylamine	−4.08	−0.15	−3.69	−3.54	−3.93
Trimethylamine	−3.24	−0.43	−4.98	−4.55	−2.81
Triethylamine	−3.22	1.57	−2.77	−4.34	−4.79
Pyridine	−4.70	−2.28	−5.24	−2.95	−2.42
N-methylimidazole	–	−5.84	−8.39	−2.55	–
Chlorobenzene	−1.12	−0.45	−1.16	−0.71	−0.67
Ethyl nitrate	–	−2.65	−1.80	0.85	–
AE		1.28	−0.24		
AUE		1.44	0.70		
RMSE		1.80	0.88		

The experimental hydration free energies are from Plyasunova et al. [18] and Abraham et al. [19]. $\Delta\Delta G_{\text{corr}}^{(1)} := \Delta G_{\text{hyd}}^{(1)} - \Delta G_{\text{hyd}}^{(0)}$. The statistical summary at the bottom of the table: *AE* average error, *AUE* average unsigned error, *RMSE* root-mean-square error. Energies in kcal/mol

introduce more parameters but firmly believe in the LSFF all-atom parameters.

So why do these systematic errors exist? Let us define the empirical correction factor as the difference between the experimental hydration free energy and the model prediction, i.e.

$$\Delta\Delta G_{\text{corr}}^{\text{empirical}} := \Delta G_{\text{hyd}}^{\text{Exp}} - \Delta G_{\text{hyd}}^{(0)}. \quad (11)$$

The needed correction factors for hydroxyl and primary amines are about −2.0 kcal/mol. Also, due to neighbouring electron donating group addition, ethers, secondary and tertiary amines have even larger correction factors. However, a neighbouring electron withdrawing group reduces the correction factor, for example, compare ethers with esters or alcohols with carboxylic acids. It is clear that the corrections needed for esters, carboxylic acids, aldehydes and ketones are smaller than the ones for alcohols, ethers, and amines. Corrections seems to be primarily important for sp3 hybridised oxygen and nitrogen. This is further supported by the fact that sp2 hybridised nitrogen in aniline, and to a lesser degree the oxygen in phenol, does not need large corrections. Corrections seem to arise from the polar protic hydrogen bond donating ability of water to a hydrogen bond acceptor functionality of the solute. The solute forms an electron-donor-acceptor complex with the first solvation shell waters. The intermolecular electron delocalization of the donor ground state charge density into the screened nuclear potential of the electron acceptor leads to a stabilizing charge-delocalisation energy. This physical effect is not part of a classical implicit solvent model, nor any classical explicit solvent models, and that is why systematic errors arise and corrections are needed.

We can make a first order correction to the calculated hydration free energy of molecule M by introducing one explicit water molecule of the first solvation shell. Introducing more waters (higher order corrections) is possible but the need for adequate sampling of configuration space will make predictions increasingly more complicated and is not considered here. According to a thermodynamic cycle scheme

$$\Delta G_{\text{hyd}}^{(1)}(\mathsf{M}) = \Delta G_{\text{hyd}}^{(0)}(\mathsf{M}\cdot \text{H}_2\text{O}) + \Delta U_{\text{bind}}(\mathsf{M}\cdot \text{H}_2\text{O}) - \Delta G_{\text{hyd}}^{(0)}(\text{H}_2\text{O}), \quad (12)$$

where ΔU_{bind} is the hydrogen bond energy of the complex formation in the gas phase including changes to the zero-point vibrational energy. The experimental hydration free energy of a water molecule $\Delta G_{\text{hyd}}^{(0)}(\text{H}_2\text{O})$ is known to be −6.31 kcal/mol. The binding energy of the solute/single water molecular complex is extrapolated to the CCSD(T) energy in the complete basis set limit, as implemented in Jaguar [16]. Geometry optimisations are performed at the B1B95-D3/6-311++G(d,p) level with corrections included for the change in zero-point energy. Corrections for the basis set superposition error and energy extrapolation using local MP2 calculations in combination with two correlation-consistent basis sets are part of the Jaguar implementation.

The hydration free energies $\Delta G_{\text{hyd}}^{(1)}$ of Table 2 show a significant improvement compared to the original predictions. Let us define the one water first-order ab initio correction

$$\Delta\Delta G_{\text{corr}}^{(1)} := \Delta G_{\text{hyd}}^{(1)} - \Delta G_{\text{hyd}}^{(0)}. \quad (13)$$

The ab initio corrections correlate well with the corresponding empirical corrections, see Table 2. The average correction for each functional group is found in Table 3. Arylation of the functional groups with a sp3 hybridised nitrogen or oxygen will lead to a reduced correction, c.f. phenol, methoxybenzene and aniline. However, for benzoic acid and methyl benzoate the effect is opposite due to the phenyl ring electron effect on the carbonyl electron withdrawing group situated next to the sp3 hybridised oxygen of the carboxylic acid and the ester. The trends are similar for the empirical corrections. The correlation coefficient between the empirical corrections and the ab initio corrections is 0.95 (Tables 3 and 4).

The SAMPL4 blind challenge

There are 47 molecules in the final set of the SAMPL4 blind challenge [4, 5]. The predicted hydration free energies of the molecules in the SAMPL4 challenge are found in Table 5. Empirical corrections are introduced to improve on the prediction accuracy. However, in order to reduce the number of deployed empirical parameters,

Table 3 Average ab initio first solvation shell corrections defined by $\Delta\Delta G_{\text{corr}}^{(1)} := \Delta G_{\text{hyd}}^{(1)} - \Delta G_{\text{hyd}}^{(0)}$ for alkylated functional groups that are present in the SAMPL4 blind challenge

Functional group (alkylated)	Mean $\Delta\Delta G_{\text{corr}}^{(1)}$	SD
Alcohol	−1.90	0.30
Ether	−1.78	0.26
Aldehyde	−1.16	0.30
Ketone	−1.04	0.24
Carboxylic acid	0.21	0.30
Ester	−0.27	0.09
Primary amine	−2.65	0.12
Secondary amine	−3.64	0.15
Tertiary amine	−4.44	0.15
Aromatic nitrogen	−2.75	0.28

Energies in kcal/mol

Table 4 Average empirical first solvation shell corrections defined by $\Delta\Delta G_{\text{corr}}^{\text{empirical}} := \Delta G_{\text{hyd}}^{\text{Exp}} - \Delta G_{\text{hyd}}^{(0)}$ for the alkylated functional groups that are present in the SAMPL4 blind challenge

Functional group (alkylated)	Mean $\Delta\Delta G_{\text{corr}}^{\text{empirical}}$	SD
Alcohol	−1.79	0.26
Ether	−2.00	0.57
Aldehyde	−0.33	0.16
Ketone	−0.37	0.46
Carboxylic acid	−0.16	0.18
Ester	0.75	0.53
Primary amine	−1.88	0.12
Secondary amine	−3.70	0.56
Tertiary amine	−3.88	0.86
Aromatic nitrogen	−2.44	0.08

The corrections are derived from a large set of mono-functional group molecules with known hydration free energies. Example molecules are found in Table 2. Energies in kcal/mol

corrections <0.5 kcal/mol are ignored. Hence, only corrections for alcohols, ethers, esters, amines and aromatic nitrogens are included in the SAMPL4 challenge. It is found that arylation of a functional group will have a lesser need for functional group correction. The empirical correction is reduced by half, as a rule of thumb, when an aryl is added. Two aryls added to a functional group results in a nullified correction.

$$\Delta G_{\text{hyd}} = \Delta G_{\text{hyd}}^{(0)} + \sum_{i}^{\text{FG}} x_i \left(\Delta\Delta G_{\text{corr}}^{\text{empirical}}\right)_i, \tag{14}$$

where x_i is the number of occurrences of functional group i in the SAMPL4 molecule. This simple and transparent implicit solvent model, based on non-linear response dielectric theory with a few empirical post-processing corrections, results in high precision hydration predictions with an accuracy <1 kcal/mol, that is, chemical accuracy is obtained (see Table 5 and Fig. 1). The correlation coefficient is 0.98, the average error (AE) is −0.42 kcal/mol, the average unsigned error (AUE) is 0.85 kcal/mol and the root-mean-square error (RMSE) is 1.14 which is just above the 1 kcal/mol limit.

Most compounds have a low prediction error, with mannitol (SAMPL4_001) as an exception. Mannitol is used here as an example to illustrate how the hydration free energy prediction of a molecule is built-up. The submitted SAMPL4 prediction for mannitol was −28.4 kcal/mol. However, a new lowest energy conformer was kindly provided by Andreas Klamt after the SAMPL4 submission. This conformer is 2.3 kcal/mol lower in energy and will be used hereafter. The calculated hydration free energy is

$$\Delta G_{\text{hyd}}^{(0)} = \Delta U_{\text{ind}} + \Delta U_{\text{elec}} + \Delta U_{\text{LJ}} + \Delta G_{\text{cav}} = 4.6 + (-24.0) + (-10.9) + 13.8 = -16.5 \quad [\text{kcal/mol}]$$

Non-linear response effects (normal dielectric saturation and electrostriction) make a total contribution of +0.8 kcal/mol to $\Delta U_{\text{elec.}}$. The empirical correction consists of six alcohols with an additive functional group correction of −1.79 kcal/mol each, and the predicted hydration free energy for mannitol is

$$\Delta G_{\text{hyd}}^{\text{empirical}} = \Delta G_{\text{hyd}}^{(0)} + 6 \times (-1.8) = -27.3 \quad [\text{kcal/mol}].$$

This predicted hydration free energy is much lower than the experimental value of $\Delta G_{\text{hyd}}^{\text{Exp}} = -23.6$ kcal/mol. The reason for this is that the six hydroxyl groups of mannitol are not identical and do not have the same physical properties. Hence the correction is overestimated. This will become clear when we switch over to use ab initio corrections instead of empirical corrections.

If we assume that the first solvation shell correction is approximately additive, i.e. the total correction is equal to the sum of all functional group corrections present in the molecule, we have

$$\Delta G_{\text{hyd}}^{\text{approx}(1)} = \Delta G_{\text{hyd}}^{(0)} + \sum_{i}^{\text{FG}} x_i \left(\Delta\Delta G_{\text{corr}}^{(1)}\right)_i, \tag{15}$$

where x_i is the number of occurrences of functional group i in the SAMPL4 molecule. It is clear that the six hydroxyl groups are not identical and two variants of water/solute complexes are set up. In the first the water molecule makes a hydrogen bond to the hydroxyl group of carbon 1 (identical with carbon 6). In the second set up the water interacts with the hydroxyl group of carbon 2 (identical with carbon 3, 4 and 5). The difference between hydroxyl group 1 and 2 is that number 1 makes one intramolecular hydrogen bond, but hydroxyl group 2 makes two intramolecular hydrogen bonds. The ab initio corrections are found to be −1.39 and −1.07 kcal/mol respectively, which are both less than the average correction −1.90 kcal/mol of a non-intramolecular hydrogen bonded hydroxyl group. The ab initio prediction becomes

$$\Delta G_{\text{hyd}}^{\text{approx}(1)} = \Delta G_{\text{hyd}}^{(0)} + 2 \times (-1.4) + 4 \times (-1.1) = -23.6 \quad [\text{kcal/mol}],$$

which corresponds very well with the experimental value of $\Delta G_{\text{hyd}}^{\text{Exp}} = -23.6$ kcal/mol.

The implicit solvent model with ab initio correction also gives predictions with a good precision (see Table 5 and Fig. 2). The corrections for aryl substituted functional groups are found in Table 2. Arylation leads to a reduced

Table 5 The calculated hydration free energies for the set of 47 molecules in the SAMPL4 challenge [4]

SAMPL4	Name	$\Delta G_{hyd}^{(0)}$	$\Delta G_{hyd}^{approx(1)}$	$\Delta G_{hyd}^{empirical}$	ΔG_{hyd}^{Exp}
001	Mannitol*	−16.53	−23.59	−27.28	−23.62 ± 0.32
002	Linalyl acetate	−3.61	−3.84	−2.87	−2.49 ± 0.85
003	Nerol	−3.61	−5.47	−5.40	−4.78 ± 0.25
004	Geraniol	−3.88	−5.74	−5.67	−4.45 ± 0.24
005	1,2-Dimethoxybenzene	−2.34	−4.58	−4.34	−5.33 ± 0.10
006	4-Propylguaiacol	−3.31	−4.24	−5.21	−5.26 ± 0.18
009	2,6-Dichlorosyringaldehyde	−6.90	−10.09	−9.80	−8.24 ± 0.76
010	3,5-Dichlorosyringol	−3.64	−6.46	−6.54	−6.24 ± 0.38
011	2-Chlorosyringaldehyde	−7.21	−9.73	−10.11	−7.78 ± 0.77
012	Dihydrocarvone	−4.18	−5.20	−4.18	−3.75 ± 0.21
013	Carveol	−4.17	−6.03	−5.96	−4.44 ± 0.43
014	l-Perillaldehyde	−4.45	−5.57	−4.45	−4.09 ± 0.17
015	Piperitone	−4.51	−5.53	−4.51	−4.51 ± 0.10
016	Menthol	−2.61	−4.51	−4.40	−3.20 ± 0.27
017	Menthone	−3.21	−4.25	−3.21	−2.53 ± 0.25
019	9,10-Dihydroanthracene	−2.85	−2.85	−2.85	−3.78 ± 0.10
020	1,1-Diphenylethene	−1.82	−2.28	−1.82	−2.78 ± 0.10
021	1-Benzylimidazole	−6.87	−9.42	−9.31	−7.63 ± 0.12
022	Mefenamic acid	−7.30	−8.12	−7.30	−6.78 ± 0.10
023	Diphenhydramine	−3.14	−9.36	−9.03	−9.34 ± 0.62
024	Amitriptyline	−2.30	−6.72	−6.19	−7.43 ± 0.60
025	1-Butoxy-2-propanol	−2.43	−6.11	−6.23	−5.73 ± 0.15
026	2-Ethoxyethyl acetate	−4.46	−6.51	−5.72	−5.31 ± 0.10
027	1,3-Bis-(nitrooxy)propane	−6.25	−4.55	−6.25	−4.80 ± 0.39
028	1,3-Bis-(nitrooxy)butane	−5.90	−4.20	−5.90	−4.29 ± 0.39
029	Hexyl nitrate	−1.71	−0.86	−1.71	−1.66 ± 0.10
030	Hexyl acetate	−3.42	−3.69	−2.67	−2.29 ± 0.12
032	3,4-Dichlorophenol	−5.18	−6.90	−6.98	−7.29 ± 0.10
033	2,6-Dimethoxyphenol	−5.09	−6.57	−7.99	−6.96 ± 0.10
034	4−Methyl−2−methoxyphenol	−3.86	−4.79	−5.76	−5.80 ± 0.10
035	2−Hydroxybenzaldehyde	−4.40	−5.15	−5.29	−4.68 ± 0.10
036	2-Ethylphenol	−3.52	−3.90	−5.31	−5.66 ± 0.10
037	2-Methoxyphenol	−3.85	−4.78	−5.75	−5.94 ± 0.10
038	2-Methylbenzaldehyde	−4.01	−4.38	−4.01	−3.93 ± 0.10
039	1-Ethyl-2-methylbenzene	−0.28	−0.28	−0.28	−0.85 ± 0.10
041	Piperidine	−1.16	−4.80	−4.86	−5.05 ± 0.10
042	Tetrahydropyran	−0.66	−2.44	−2.66	−3.13 ± 0.10
043	Cyclohexene	0.79	0.81	0.79	0.14 ± 0.10
044	1,4-Dioxane	−2.75	−6.31	−4.76	−5.08 ± 0.10
045	2−Amino-9,10-anthraquinone	−12.51	−12.25	−13.45	−11.53 ± 0.29
046	1-Amino-9,10-anthraquinone	−8.92	−8.66	−9.86	−9.44 ± 0.74
047	1-(2-Hydroxyethylamino)-9,10-anthraquinone	−10.56	−12.20	−13.29	−14.21 ± 1.10
048	1,4-Diamino-9,10-anthraquinone	−12.62	−13.62	−14.50	−11.85 ± 0.35
049	dibenzo-p-dioxin	−3.09	−4.19	−3.09	−3.16 ± 0.10
050	Anthracene	−2.99	−2.99	−2.99	−4.14 ± 0.10
051	1-Amino-4-hydroxy-9,10-anthraquinone	−9.85	−9.97	−11.69	−9.53 ± 0.28
052	Diphenyl ether	−2.41	−2.96	−2.41	−2.87 ± 0.69
	AE	1.15	−0.26	−0.42	
	AUE	1.55	0.94	0.85	
	RMSE	2.19	1.09	1.14	

* The predicted hydration free energies of mannitol are based on the new lowest energy conformer provided after the submission of SAMPL4, see text. The statistical summary at the bottom of the table: *AE* average error, *AUE* average unsigned error, *RMSE* root-mean-square error. Energies in kcal/mol

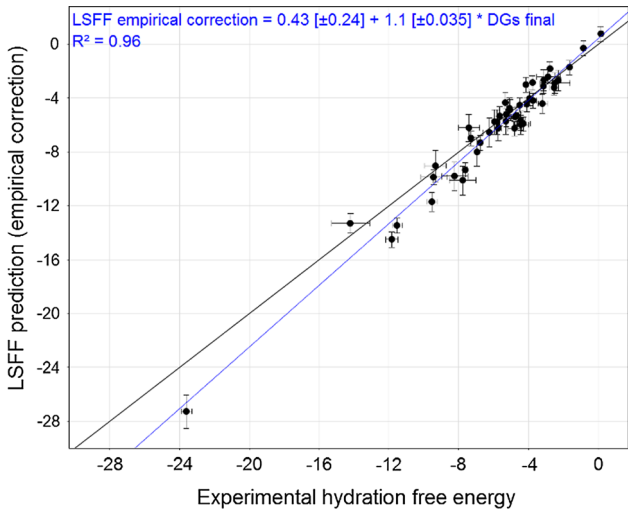

Fig. 1 Experimental hydration free energies versus Molhydro/LSFF predictions with empirical corrections. *Error bars* are estimated from the average error of molecules not needing an empirical correction and from the uncertainties of applied empirical corrections, assuming propagation of uncorrelated errors. Energies in kcal/mol

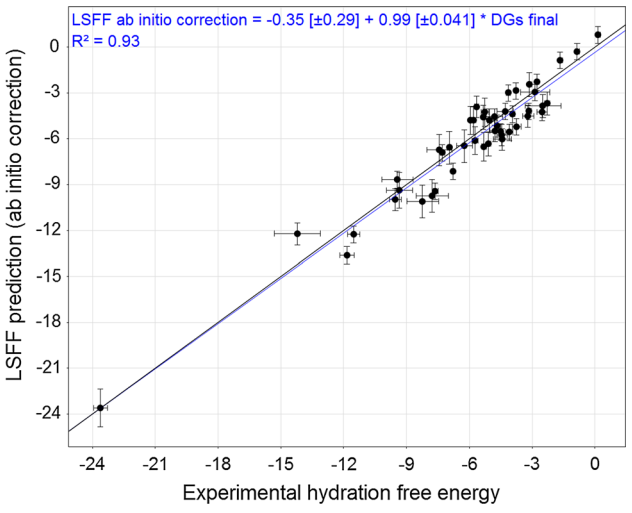

Fig. 2 Experimental hydration free energies versus Molhydro/LSFF predictions with ab initio corrections. *Error bars* are identical to the ones in Fig. 1. Energies in kcal/mol

need for corrections in much the same way as for the empirical corrections. The correlation coefficient for the SAMPL4 set is 0.96, the average error (AE) is −0.26 kcal/mol, the AUE is 0.94 kcal/mol and the RMSE is 1.09 which is just above the 1 kcal/mol limit. Chemical accuracy is obtained once more, but this time without any empirical correction terms. The implicit solvent model now only contains the four SPT coefficients in the cavity formation term, which are fitted to hydration free energy data. The remaining model parameters, e.g. the pair of LSFF atomic parameters per element, are not in any way associated with hydration free energy modelling.

Conclusion

In the introduction we outlined five requirements for calculating the hydration free energy of the molecules in the SAMPL4 challenge. (1) An implicit solvent model is implemented which is transparent and includes all physical effects relevant to solvation modelling. It comprises a non-simple medium description of the water solvent, and a solute represented by a minimal parameter set model, i.e. its quantum mechanical lowest energy conformer, atomic ESP charges and the all atom LSFF derived from VLE molecular models [11]. (2) This approach is generally applicable to molecules found in organic and biological chemistry. However, the released polarisation stress within the formed electron-donor-acceptor complex between the solute and the first solvation shell water molecules, introduces a need for a non-classical correction to calculated hydration free energies. Corrections can be derived from known experimental data, i.e. introducing empirical parameters, or be calculated from first principle by forming a solute/explicit water complex and add the binding energy contribution to the hydration free energy prediction. (3) If we use ab initio corrections the solvation model depends on only four fitted SPT coefficients linked to experimental hydration free energy data. (4) Made predictions are accurate and comparable to, or better than, the accuracy of the results of the golden standard of explicit solvent free energy calculations [5]. (5) The SAMPL4 challenge predictions has an AUE less than and a RMSE near 1 kcal/mol. Chemical accuracy is obtained.

Acknowledgments I would like to thank Ulf Ryde for fruitful discussions and valuable comments. I am also grateful to Andreas Klamt for providing me with his lowest energy conformer of mannitol (SAMPL4_001).

References

1. Mikulskis P, Genheden S, Rydberg P, Sandberg L, Olsen L, Ryde U (2012) J Comput Aided Mol Des 26:527–541
2. Sandberg L, Edholm O (2002) J Chem Phys 116:2935–2944
3. Sandberg L, Casemyr R, Edholm O (2002) J Phys Chem B 106:7889–7897
4. Guthrie JP (2014) SAMPL4, A blind challenge for computational solvation free energies: the compounds considered. J Comput Aided Mol Des (in press)
5. Mobley DL, Wymer K, Lim NM (2014) Blind prediction of solvation free energies from the SAMPL4 challenge. J Comput Aided Mol Des (in press)
6. Desnoyers JE, Verrall RE, Conway BE (1965) J Chem Phys 43:243–250
7. Chandler D (2005) Nature 437:640–647

8. Reiss H, Frisch HL, Lebowitz JL (1959) J Chem Phys 31:369–380
9. Reiss H, Frisch HL, Helfand E, Lebowitz JL (1960) J Chem Phys 32:119–124
10. Rick SW (2004) J Chem Phys 120:6085–6093
11. Vrabec J, Stoll J, Hasse H (2001) J Phys Chem B 105:12126–12133
12. Bondi A (1964) J Phys Chem 68:441–451
13. Rowland RS, Taylor R (1996) J Phys Chem 100:7384–7391
14. (2013) LigPrep, version 2.6, Schrödinger, LLC, New York
15. (2013) MacroModel, version 10.0, Schrödinger, LLC, New York
16. (2013) Jaguar, version 8.0, Schrödinger, LLC, New York
17. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) J Chem Phys 79:926–935
18. Plyasunova NV, Plyasunov AV, Shock EL (2004) Intern J Thermophys 25:351–360
19. Abraham MH, Andonian-Haftvan J, Whiting GS, Leo A, Taft RS (1994) J Chem Soc Perkin Trans 2:1777–1791