J Comput Aided Mol Des (2014) 28:169–173
DOI 10.1007/s10822-013-9701-3

Prediction of free energies of hydration with COSMO-RS on the SAMPL4 data set

Jens Reinisch · Andreas Klamt

Received: 4 November 2013 / Accepted: 16 December 2013 / Published online: 14 January 2014

Abstract The COSMO-RS method has been used for the prediction of free energies of hydration on a dataset of 47 complex multifunctional compounds considered in the SAMPL4 challenge. Straight application of the COSMO*therm* software with the parameterization C21_0108 yields a predictive accuracy of 1.46 kcal/mol root mean square error overall and 1.18 kcal/mol if a single dominant outlier is removed.

Keywords Free energy of solvation · Hydration free energy · COSMO-RS · SAMPL4 · Molecular modeling · Molecular simulation

Introduction

The COSMO*therm* implementation of the COSMO-RS theory has been applied to the free energy of hydration part of all three previous SAMPL challenges. On all three datasets it proved to deliver robust predictions [1–3]. Here we report the results of the COSMO-RS method for the prediction of the free energies of hydration considered in the SAMPL4 challenge. The SAMPL4 dataset [4, 5] mainly consists of compounds with one or more hydrogen bonding acceptor or donor functions including hydroxy-, amine-, ester-, ether- and keto-groups. In addition, aromatic compounds, alkenes, and some less common groups like nitro or nitryloxy substituents are present.

Since the COSMO-RS [6–8] method has been used in the publications for SAMPL1 [1], SAMPL2 [2] and SAMPL3 [3] it shall not be described here in detail. The general accuracy of the free energy of solvation for small and medium sized neutral organic compounds is in the range of 0.5 kcal/mol as was shown on the large training dataset of the SM8-model containing overall 2,346 solvation free energies [9], 284 of these being hydration free energies. The overall accuracy for the 36 chlorinated compounds in SAMPL3 was 1.05 kcal/mol (RMSE) and it was below 0.5 kcal/mol for the 22 dioxins and chloroethanes.

Methods

Only a brief overview on COSMO-RS can be given here, since the details of the method have been described in previous papers [6–8]. COSMO-RS starts with a quantum chemical continuum solvation calculation (i.e. a COSMO calculation) for all relevant conformations of each molecule. The main information taken from these calculations is the surface polarization charge density σ, which basically provides the charge distribution on the molecular surface. The next step is to use σ to calculate the interaction energies as pair contact energies E(σ, σ') and the resulting contact probabilities of all possible surface contacts with the COSMO-RS equations by neglecting the 3D geometry of the molecules and considering a liquid as an ensemble of molecular surface patches. From the contact probabilities the chemical potentials of the surface patches and of the molecules in the solution can be calculated. The additional ingredients for the free energy of solvation are the quantum

J. Reinisch (✉) · A. Klamt
COSMOlogic GmbH&CoKG, Imbacher Weg 46,
51379 Leverkusen, Germany
e-mail: reinisch@cosmologic.de

A. Klamt
Institute of Physical and Theoretical Chemistry, University of Regensburg, 93040 Regensburg, Germany

Table 1 Common data points of parameterization C21_0108 with SAMPL4 data set, units are kcal/mol

#	Name	SAMPL4 exp. data		Internal exp. data		NIST webbook		
		G	Error	G	Dev.	min	max	average
42	Tetrahydropyran	−3.13	0.1	−3.12	0.01	−3.12		−3.12
43	Cyclohexene	0.14	0.1	0.37	0.23	0.27	0.37	0.32
44	1,4-dioxane	−5.08	0.1	−5.05	0.03	−5.09	−5.06	−5.06
50	Anthracene	−4.14	0.1	−3.46	0.68	−4.28	−2.09	−3.29

Additional data from the NIST webbook [10] has been calculated from Henry Law constant data. For the NIST values only those marked as direct measurement data, or as values computed from vapour pressure measurements, or as taken from a literature review have been considered (i.e. data denoted as “M”, “V”, or “L”)

chemical energies for the gas phase and for the COSMO state.

The workflow of our COSMO-RS predictions is the same as in the SAMPL2 evaluation and is described in detail in the respective publication [2]. The COSMO*therm* software with the parameterization BP-TZVP_C21_0108 (2008) was used to calculate the free energies of hydration without any adjustments or additional fitting. The parameterization was chosen because it has only 4 data point for free energy of hydration in common with the SAMPL4 data set. Later parameterizations make use of more data points of the SAMPL4 set. The 4 data points used for parameterizing C21_0108 are listed in Table 1. In addition Table 1 provides some data from the NIST webbook [10].

For all compounds the COSMO*conf* workflow [1] has been applied to identify relevant conformations. 15 of the final 47 compounds have been taken from our precalculated COSMO file database, which is based on the same procedure. The databases contain a set of gas-phase and COSMO conformations for each compound. For 11 of the compounds at least one conformation that has been generated manually was added to the conformer set.

Results and discussion

Table 2 and Fig. 1 present the predicted versus experimental data. The original set provided for the blind predictions consisted of 52 molecules of which three have been removed because the wrong compound structure was provided in the contest briefings. The remaining 49 compounds have been reduced to the final set of 47 compounds by discarding 4,5-dichloroguaiacol and 5-chloroguaiacol, which after taking a deeper look into the experimental data turned out to be not reliable enough to be kept within the set. Moreover, new and/or improved experimental data values have been assigned to 9 of the data points by P. Guthrie and D. Mobley of the SAMPL steering committee during the evaluation process. The errors provided for the COSMO-RS predictions are not statistical errors, which would be essentially zero for COSMO-RS. Instead they show the general accuracy as compared to experimental data (0.5 kcal/mol) with some enlarged values according to the experience of the authors. The error values do not have a strict mathematical meaning.

The root mean square error of the COSMO*therm* predictions for the remaining 47 compounds is finally computed to 1.46 kcal/mol and thus is higher than was expected from previous SAMPL challenges. The main reason for the larger error is a single strong outlier, the mannitol molecule. Without this data point the RMSE drops to 1.18 kcal/mol, which is consistent with previous observations.

It should be noted that most of the major corrections of experimental data (mannitol, 4-propylguaiacol, 4,5-chloroguaiacol, 5-chloroguaiacol and 1-benzylimidazol) have been stimulated by an outlier analysis. In the above cases we could prove or at least give plausible arguments that the experimental data was wrong or inconsistent. Checking experimental data points which deviate more than expected (e.g. >2 kcal/mol) from the COSMO-RS prediction seems to be a viable approach for data curation. This could be especially effective if several different (not adjusted) methods show the same outliers.

As mannitol is the most significant outlier, we checked for possible reasons. From the experimental side it has the most negative of all hydration free energies and is therefore extremely hard to measure. The given value is derived from a temperature extrapolation, and it has been largely corrected compared to the initially provided experimental value of −27.79 kcal/mol. To check our predictions we calculated water solubility data and octanol–water partition and compared the results with experimental values. The results are shown in Table 3.

From the deviations to experimental data it is straightforward to calculate the deviation in the chemical potential prediction. The error of COSMO*therm* predictions for the chemical potential difference of the two shown properties lies between 1 and 2 kcal/mol, which is quite a lot compared with the typical accuracy (0.5–1 kcal/mol). It should

Table 2 Experimental and predicted hydration free energies in kcal/mol

#	Name	Generation	Orig. exp. data		Corrected exp. data		Predicted		
			G	Error	G	Error	G	Error	Deviation
1	Mannitol	CC	−27.79	0.32	−23.62	0.32	−17.60	0.5	6.02
2	Linalyl acetate	CC	−2.49	0.85			−3.27	0.5	−0.78
3	Nerol	CC	−4.78	0.25			−4.61	0.5	0.17
4	Geraniol	CC	−4.45	0.24			−5.02	0.5	−0.57
5	1,2-Dimethoxybenzene	DB + M	−5.33	0.1			−5.09	1.5	0.24
6	4-Propylguaiacol	DB + M	−6.44	0.18	−5.26	0.18	−4.96	1	0.30
7	4,5-Dichloroguaiacol	CC	Removed, unreliable exp. results						
8	5-Chloroguaiacol	CC	Removed, unreliable exp. results						
9	2,6-Dichlorosyringaldehyde	CC	−8.24	0.76			−9.80	1.5	−1.56
10	3,5-Dichlorosyringol	CC	−6.24	0.38			−6.36	1	−0.12
11	2-Chlorosyringaldehyde	CC	−7.78	0.77			−9.68	1.5	−1.90
12	Dihydrocarvone	CC	−4.46	0.21	−3.75	0.21	−5.12	0.5	−1.37
13	Carveol	CC	−5.03	0.43	−4.44	0.43	−5.48	0.5	−1.04
14	l-Perillaldehyde	CC	−4.09	0.17			−5.48	0.5	−1.39
15	Piperitone	CC	−4.48	0.1	−4.51	0.1	−6.25	0.5	−1.74
16	Menthol	CC	−3.36	0.27	−3.2	0.28	−4.53	1	−1.33
17	Menthone	CC	−2.51	0.25	−2.53	0.25	−3.97	0.5	−1.44
18		CC	Removed: wrong compound						
19	9,10-Dihydroanthracene	CC + M	−3.78	0.1			−3.66	0.5	0.12
20	1,1-Diphenylethene	CC	−2.78	0.1			−2.94	0.5	−0.16
21	1-Benzylimidazole	CC	−3.9	0.12	−7.63		−10.09	0.5	−2.46
22	Mefenamic acid	CC	−6.78	0.1			−8.13	0.5	−1.35
23	Diphenhydramine	CC	−9.34	0.62			−6.75	0.5	2.59
24	Amitriptyline	CC	−6.92	0.6	−7.43	0.6	−6.01	0.5	1.42
25	1-Butoxy-2-propanol	CC	−5.73	0.15			−4.29	0.5	1.44
26	2-Ethoxyethyl acetate	CC	−5.31	0.1			−6.16	1.5	−0.85
27	1,3-*bis*-(nitrooxy)propane	CC	−4.8	0.39			−3.15	0.5	1.65
28	1,3-*bis*-(nitrooxy)butane	CC	−4.29	0.39			−2.70	0.5	1.59
29	Hexyl nitrate	CC	−1.66	0.1			−0.20	0.5	1.46
30	Hexyl acetate	CC	−2.29	0.12			−3.15	1	−0.86
31		CC	Removed: wrong compound						
32	3,4-Dichlorophenol	DB + M	−7.29	0.1			−7.73	0.5	−0.44
33	2,6-Dimethoxyphenol	DB + M	−6.96	0.1			−7.54	2	−0.58
34	4-Methyl-2-methoxyphenol	DB + M	−5.8	0.1			−5.11	1	0.69
35	2-Hydroxybenzaldehyde	DB + M	−4.68	0.1			−3.86	0.5	0.82
36	2-Ethylphenol	DB + M	−5.66	0.1			−5.14	0.5	0.52
37	2-Methoxyphenol	DB + M	−5.94	0.1			−5.17	1	0.77
38	2-Methylbenzaldehyde	DB + M	−3.93	0.1			−4.09	0.5	−0.16
39	1-Ethyl-2-methylbenzene	DB + M	−0.85	0.1			−0.62	0.5	0.23
40		CC	Removed: wrong compound						
41	Piperidine	DB + M	−5.05	0.1			−3.63	1	1.42
42	Tetrahydropyran	DB + M	−3.13	0.1			−2.93	0.5	0.20
43	Cyclohexene	CC	0.14	0.1			0.36	0.5	0.22
44	1,4-Dioxane	CC	−5.08	0.1			−5.65	0.5	−0.57
45	2-Amino-9,10-anthraquinone	CC	−11.53	0.29			−14.14	0.5	−2.61
46	1-amino-9,10-anthraquinone	CC	−9.44	0.74			−9.88	0.5	−0.44
47	1-(2-Hydroxyethylamino)-9,10-anthraquinone	CC	−14.21	1.1			−13.15	0.5	1.06

Table 2 continued

#	Name	Generation	Orig. exp. data		Corrected exp. data		Predicted		
			G	Error	G	Error	G	Error	Deviation
48	1,4-Diamino-9,10-anthraquinone	CC	−11.85	0.35			−13.70	0.5	−1.85
49	Dibenzo-p-dioxin	CC	−3.16	0.1			−3.33	0.5	−0.17
50	Anthracene	DB	−4.14	0.1			−3.64	0.5	0.50
51	1-Amino-4-hydroxy-9,10-anthraquinone	CC	−9.53	0.28			−10.10	0.5	−0.57
52	Diphenyl ether	DB	−2.87	0.69			−3.25	0.5	−0.38
	Root mean square error (RMSE)								1.46
Data for removed compounds 7 and 8									
7	4,5-Dichloroguaiacol	CC	−4.37	0.62	−3.58		−6.31		
8	5-Chloroguaiacol	CC	−8.21	0.37	−5.26		−5.82		

The column "Generation" provides information about the source of the conformations: "CC" = COSMO*conf* workflow, "DB" = COSMO*logic* databases, "M" = one or more conformations were added manually. The "Orig. exp. data" column contains the experimental data given initially, while the "Corrected exp. data" show those values corrected by P. Guthrie and D. Mobley after revisiting experimental sources

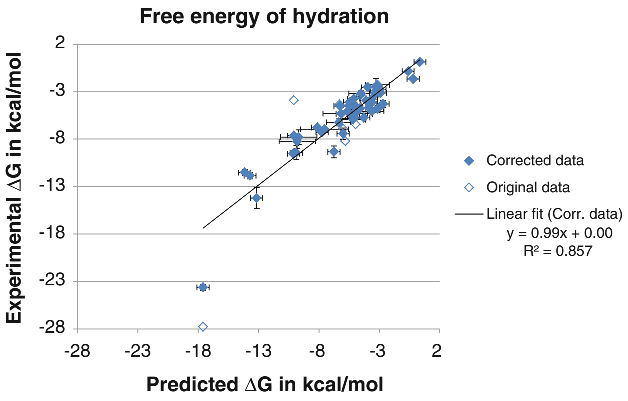

Fig. 1 Scatter plot of predicted versus experimental data. The *horizontal error bars* indicate the estimated errors for the submission, while the *vertical error bars* correspond to the errors provided in the SAMPL4 data set. Original data for the two removed guaiacols have been included in the original data plot

be noted that in both examples the prediction underestimates the solubility in water which is consistent with the too positively predicted value for the free energy of hydration. Though mannitol is predicted quite poorly in water, the missing 6 kcal/mol in free energy of hydration cannot be explained by wrong predictions of the chemical potential in water. Thus the origin for the exceptionally strong deviation remains unresolved. Therefore we do still have doubts with respect to the experimental value provided for mannitol.

Conclusion

COSMO-RS as implemented in the COSMO*therm* software is able to yield a RMSE of 1.46 kcal/mol for the SAMPL4 data set. Leaving out the problematic case of mannitol, the RMSE reduces to 1.18 kcal/mol. The experimental errors provided by the SAMPL organizers have an overall RMSE of 0.39 kcal/mol. This error was designated by reading and analyzing a large number of relevant publications. To acquire such good quality experimental data, a deeper knowledge of the measurement techniques and quite some experience is indispensable. Reconsidering the data collected in this paper, it becomes clear that just picking a single experimental data point from the first publication or measurement at hand will give a much larger uncertainty than the given 0.39 kcal/mol. In addition, for several cases of large deviations between the COSMO-RS prediction and the originally provided experimental data, inconsistencies or errors in the experimental data have been identified and corrected. A robust prediction can thus give guidance for the validation of

Table 3 Comparison of experimental and predicted values for octanol–water partitioning (logP) and water solubility of mannitol [11]

	Exp. data	Predictions	Dev. in ln-units	Dev. in kcal/mol
LogP	−2.20	−1.20	2.30	1.36
Ln(x) Solubility	−3.63	−7.06	−3.43	−2.02

Data for logP have been taken from a commercial logP database [12], where −2.20 is the recommended value and 6 other values between −2.1 and −3.10 have also been reported. The water solubility predictions have been conducted with experimental data for H_{fusion} = 56.1 kJ/mol and T_{melt} = 439 K [13] experimental data and provide approximate values for cases where measurements are difficult or inaccessible.

References

1. Klamt A, Eckert F, Diedenhofen MJ (2009) J Phys Chem B 113:4508–4510
2. Klamt A, Diedenhofen MJ (2010) J Comput Aided Mol Des 24:357–360
3. Reinisch J, Klamt A, Diedenhofen MJ (2012) J Comput Aided Mol Des 26:660–673
4. Guthrie JP. SAMPL4, A blind challenge for computational solvation free energies: the compounds considered. Department of Chemistry, University of Western Ontario, London
5. Mobley DL, Wymer K, Lim NM. Blind prediction of solvation free energies from the SAMPL4 challenge
6. Klamt A (1995) J Phys Chem 99:2224
7. Klamt A, Jonas V, Bürger T, Lohrenz JCW (1998) J Phys Chem 102:5074
8. Klamt A (2005) COSMO-RS: from quantum chemistry to fluid phase thermodynamics and drug design. Elsevier, Amsterdam
9. Klamt A, Mennucci B, Tomasi J, Barone V, Curutchet C, Orozco M, Luque FJ (2009) Acc Chem Res 42:489–492
10. Sander R. “Henry’s Law Constants”, NIST Chemistry WebBook, NIST Standard Reference Database Number 69. In: Linstrom PJ, Mallard WG (eds) National Institute of Standards and Technology, Gaithersburg, http://webbook.nist.gov. Retrieved 30 Oct 2013
11. Robinson RA, Stokes RH (1961) J Phys Chem 65(11):1954–1958
12. BioByte Corporation, Clairmont, CA, 2008
13. Domalski ES, Hearing ED. “Phase Change Data”, NIST chemistry webbook, NIST standard reference database number 69. In: Linstrom PJ, Mallard WG (eds) National Institute of Standards and Technology, Gaithersburg, http://webbook.nist.gov. Retrieved 30 Oct 2013