Prediction of free energies of hydration with COSMO-RS on the SAMPL3 data set

Jens Reinisch[*1], Andreas Klamt[1,2], Michael Diedenhofen[1]*

1) COSMOlogic GmbH&CoKG, Burscheider Str. 515, 51381 Leverkusen, Germany

2) Institute of Physical and Theoretical Chemistry, University of Regensburg, 93040 Regensburg, Germany

*) CORRESPONDING AUTHOR: +49-2171-363664; reinisch@cosmologic.de

ABSTRACT: The COSMO-RS method has been used for the prediction of free energies of hydration on the dataset of 36 chlorinated ethanes, biphenyls and dioxins considered in the SAMPL3 challenge. Straight application of the latest version of the COSMO*therm* software yields an overall predictive accuracy of 1.05 kcal/mol (RMSE). The predictions for the chlorinated ethanes and dioxins are much better with 0.38 kcal/mol and 0.49 kcal/mol RMSE respectively. The predictions for the chlorinated biphenyls show a systematic shift of approximately 1 kcal/mol, but the large RMSE of 1.59 kcal/mol mainly arises from two exceptional outliers. Possible reasons for this observation are discussed.

Introduction

Blind tests, like the SAMPL challenges or the Industrial Fluid Property Simulation Challenge, are valuable instruments for the evaluation of computational simulation methods. We therefore participate in as many blind prediction contests as possible to provide independent data for the assessment of COSMO-RS [1-4] in its COSMO*therm* implementation to the scientific community. In the SAMPL2 contest the COSMO-RS had the smallest RMSE of all submissions [5]. Unfortunately we missed the SAMPL3 challenge and can therefore only present our predictions after the experimental data have been published. Nevertheless, because no special adjustments to experimental data were made, the predictions are still appropriate to evaluate the performance of COSMO-RS versus other simulation methods. The data set is well described in the papers of Geballe and Guthrie [6] and Beckstein and Iorga [7] including 2D structures. The latter contribution also presents the best results of all participants in the SAMPL3 blind prediction challenge. We will compare the quality of our predictions with the experimental results and also the best data from Beckstein and Iorga.

The COSMO-RS method is well established and only some important aspects of the method will be described. The basis of COSMO-RS are quantum chemical calculation of the molecules using the conductor like screening model (COSMO), which is available in several quantum chemistry software suites as COSMO [8] or C-PCM [9]. In addition to the dielectric continuum solvation approach, COSMO-RS adds other interaction contributions, e.g. hydrogen bonding, and combines it with a statistical thermodynamics treatment. To predict the free energy of solvation an estimate for the chemical potential in gas phase (μ_{gas}) is needed in addition to the chemical potential in the liquid state (μ_{liq}). Within the COSMO*therm* software μ_{gas} is based on the quantum chemical energy of the molecule in gas phase, which is not a native part of COSMO-RS theory. The prediction of the free energy of hydration$_{=}$ ΔG^{X}_{hydr} is thus based on quantum chemical calculations in the liquid phase (COSMO) and the gas phase, and the COSMO-RS post-processing. This allows for the prediction of arbitrary organic compounds in any kind of solvent or solvent mixtures without additional adjustments or fitting. The accuracy of ΔG^{X}_{solv} for small and medium sized, neutral organic compounds is generally in the range of 0.5 kcal/mol as was shown on the large training dataset of the SM8-model containing overall 2346 solvation free energies [10], 284 of these being hydration free energies.

Methods

The workflow of our COSMO-RS predictions is the same as in the SAMPL2 evaluation and is described in detail in the respective publication [6]. The COSMO*therm* software, with the current parameterization BP-TZVP_C30_1201 was directly used to calculate the free energies of hydration. Many of the compounds have been already available in our databases. The databases contain a set of gas-phase and COSMO conformations for each compound. In addition some experimental values for vapour pressure, melting and boiling point, free energy of fusion but generally no experimental data for ΔG_{solv}^{X} are available. The predictions have been conducted without any experimental data and only the conformations have been taken from the databases to conduct the COSMO*therm* calculations. Only 16 compounds (see table 1) had to be calculated in COSMO state and gas phase, which took a few CPU days on a typical 2.4 GHz processor. A conformational search has been conducted with our COSMO*conf* workflow [4], though the most of the compounds of this dataset can be expected to have only one relevant conformation. Only for two of the chlorinated ethanes two conformations needed to be taken into account.

Results and discussion

Figure 1 and Table 1 show the predicted versus experimental data. The predictions for chlorinated ethane and dioxins show a very good correlation with no exceptional outliers and a low root mean square error (RMSE) of 0.38 kcal/mol and 0.49 kcal/mol respectively. In contrast the predictions for PCBs show a generally weaker agreement (1.59 kcal/mol RMSE) and two severe outliers (> 3kcal/mol). The overall error for the 36 compounds is 1.05 kcal/mol. In contrast to molecular dynamics simulations (MD) COSMO-RS will yield the exact same numbers in every run as long as the same molecules and parameterization are used. Thus when comparing COSMO-RS results with MD results only the statistical error of the MD result has to be considered and all other deviations are due to systematic errors in the methods (e.g. force fields vs. COSMO-RS). To give the reader an impression of the quality of the presented predictions a comparison to the best value from Beckstein and Iorga [7], i.e. 1.21 kcal/mol, is adequate. We would like to emphasize that Beckstein and Iorga made a true blind prediction, whereas our results have been calculated after the publication of the SAMPL3 results. As we did not make any adjustments to any experimental data, every user, using the same method, would get the same results and we therefore believe that comparability is still given.

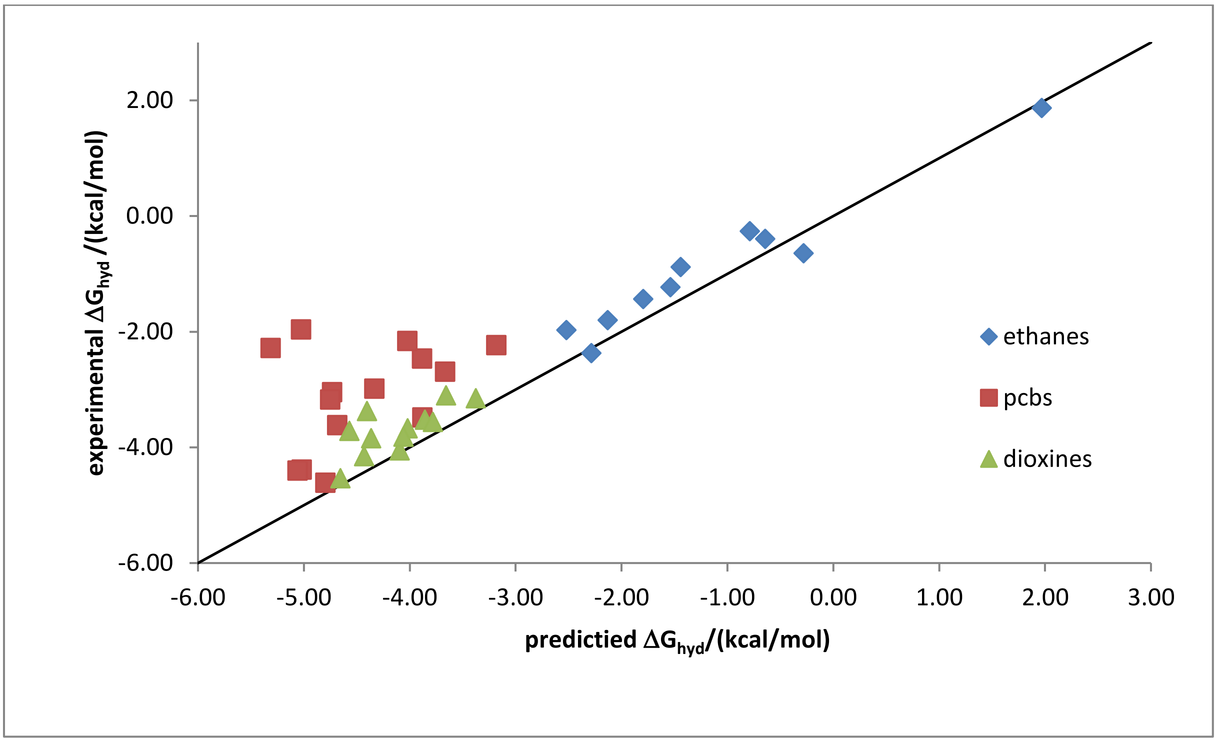

Fig. 1: COSMO*therm* predictions vs. experimental data.

Table 1 Experimental and predicted hydration free energies in kcal/mol. Exp. data from [6]. The column DB marks geometries taken from our databases.

Id	Name	DB	# Chlor	Exp.	Pred.	Error
1	Ethane	y	0	1.87	1.97	0.10
2	Chloroethane	y	1	-0.39	-0.64	-0.25
3	1,1-dichloroethane	y	2	-0.88	-1.44	-0.56
4	1,2-dichloroethane	y	2	-1.80	-2.13	-0.33
5	1,1,1-trichloroethane	y	3	-0.26	-0.79	-0.53
6	1,1,2-trichloroethane	y	3	-1.97	-2.52	-0.55
7	1,1,1,2-tetrachloroethane	y	4	-1.43	-1.80	-0.37
8	1,1,2,2-tetrachloroethane	y	4	-2.37	-2.29	0.08
9	Pentachloroethane	y	5	-1.23	-1.54	-0.31
10	Hexachloroethane	y	6	-0.64	-0.28	0.36
11	Biphenyl	y	0	-2.23	-3.18	-0.95
12	2-chlorobiphenyl	y	1	-2.69	-3.67	-0.98
13	2,5-dichlorobiphenyl	n	2	-2.46	-3.88	-1.42
14	2,4,6-trichlorobiphenyl	n	3	-2.16	-4.02	-1.86
15	2,3,4,5-tetrachlorobiphenyl	n	4	-3.48	-3.88	-0.40
16	2,2',6,6'-tetrachlorobiphenyl	n	4	-2.28	-5.31	-3.03
17	2',3,4,5,5'-pentachlorobiphenyl	n	5	-3.61	-4.69	-1.08
18	2,2',4,6,6'-pentachlorobiphenyl	n	5	-1.96	-5.03	-3.07
19	2,3,3',4',5,6-hexachlorobiphenyl	n	6	-4.38	-5.02	-0.64
20	2,3,3',4,4',5-hexachlorobiphenyl	n	6	-3.04	-4.73	-1.69
21	2,2',3,3',4,4',5-heptachlorobiphenyl	n	7	-4.40	-5.06	-0.66
22	2,3,3',4,4',5,5'-heptachlorobiphenyl	n	7	-3.17	-4.75	-1.58

23	2,2',3,3',4,4',5,6'-octachlorobiphenyl	n	8	-4.61	-4.80	-0.19
24	Decachlorobiphenyl	n	10	-2.98	-4.33	-1.35
25	Dibenzo-p-dioxin	y	0	-3.15	-3.38	-0.23
26	1-chlorodibenzo-p-dioxin	y	1	-3.52	-3.86	-0.34
27	2-chlorodibenzo-p-dioxin	y	1	-3.10	-3.66	-0.56
28	2,3-dichlorodibenzo-p-dioxin	y	2	-3.56	-3.78	-0.22
29	2,7-dichlorodibenzo-p-dioxin	y	2	-3.67	-4.02	-0.35
30	1,2,4-trichlorodibenzo-p-dioxin	y	3	-4.05	-4.10	-0.05
31	1,2,3,4-tetrachlorodibenzo-p-dioxin	n	4	-3.81	-4.06	-0.25
32	1,2,3,7-tetrachlorodibenzo-p-dioxin	n	4	-3.84	-4.36	-0.52
33	2,3,7,8-tetrachlorodibenzo-p-dioxin	y	4	-3.37	-4.41	-1.04
34	1,2,3,4,7-pentachlorodibenzo-p-dioxin	n	5	-4.15	-4.43	-0.28
35	1,2,3,4,7,8-hexachlorodibenzo-p-dioxin	n	6	-3.71	-4.57	-0.86
36	Octachlorodibenzo-p-dioxin	y	8	-4.53	-4.66	-0.13

When interpreting the COSMO-RS results, the compounds show similar interaction schemes when solved in water. All compounds show no or low hydrogen bonding capacity. The chlorine atoms are not sufficiently polar to be acceptors and only a few of the hydrogen atoms are sufficiently polarized by the chlorine atoms in order to be weak donors. Even the oxygen atoms in the dioxins are rather non-polar and only weak acceptors. The dominating factors for the free energy of hydration are the van der Waals (vdW) interactions and the polarity. While the vdW interaction energy increases with the number of chlorine atoms in each class, the polarity is more complex. The initial chlorination always increases the polarity, but partial compensation of the local dipole moments sometimes goes along with further chlorination, often leading to almost completely non-polar per-chlorinated compounds. This is illustrated in Figure 2. While ethane and hexachloroethane show the same low polarization in the positive σ region (negative surface charges) chloroethane is more polar in this area. In the region of negative σ (positive charges) the pentachloroethane is much more polarized (σ around -0.015) and chlorethane is slightly more polarized (σ around -0.0075) than ethane or hexachloroethane, which is caused by polarized C-H bonds. Though the polar bump for pentachlorhexane looks rather tiny, it is very important as it corresponds to the strongest CH-donor of all ethanes with a polar area of 50% of two water donors. Though hydrogen bonding is not the dominating effect, it influences the free energy of solvation for some compounds. By simply looking at the sigma profiles, the ethane and hexachloroethane can be identified as the least polar compounds of the ethane series. The ΔG_{hydr}^{X} of ethane consequently has the highest value, i.e. is the most hydrophobic, in this series and hexachloroethane shows a rather high value as compared to pentachloro- or tetrachloroethanes. The presented qualitative analysis of the polarities is well resembled by the experimental data.

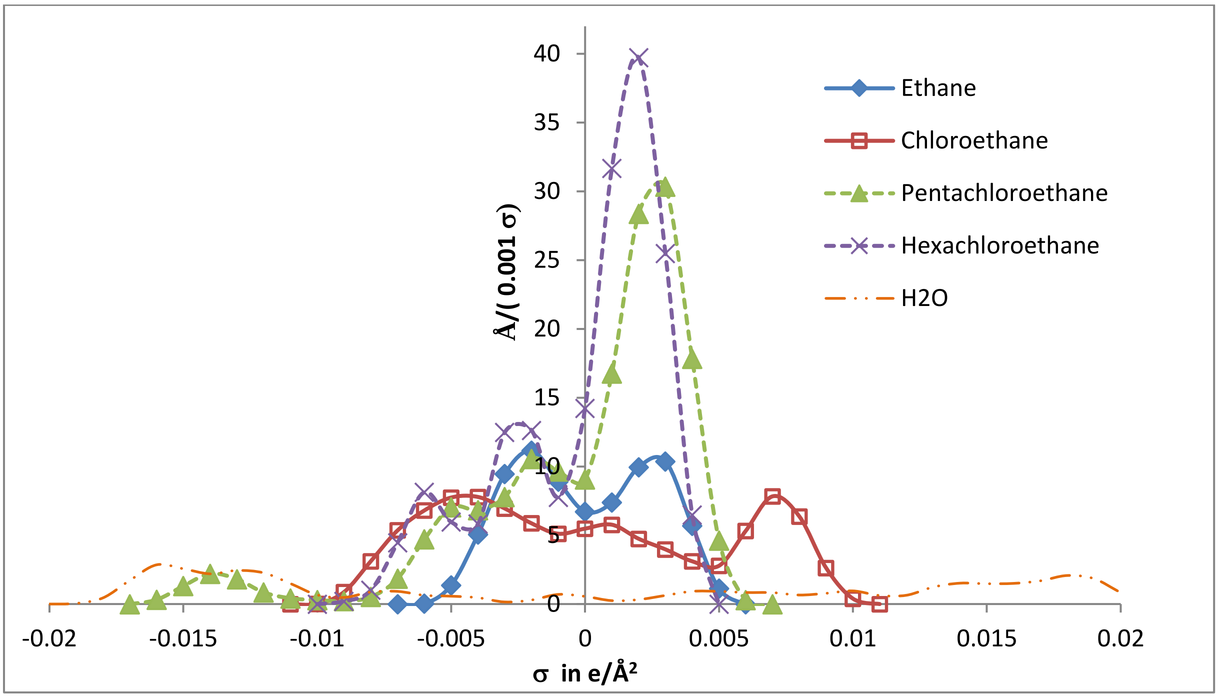

Fig. 2 σ profiles of ethane, chloroethane, pentachloroethane, hexachloroethane and water for comparison. The y-axis shows the amount of area of a certain polarity interval σ, where the integral of the curve is the total surface area of the molecule. The σ (surface screening charge) results from the quantum chemical COSMO calculations and has the opposite sign of the surface charge.

The prediction quality for the chlorinated ethanes and dioxins shows that COSMO-RS in general catches the trends of chlorination problems quite well, even for the more complex dioxins. No trend can be observed with respect to the number of chlorine atoms in these two subsets. Nevertheless, a regression line of the experimental data vs. the COSMO-RS predictions has a slope of 0.925, indicating that COSMO-RS slightly overestimates the hydration energies of the more polar chlorinated compounds. This might be due to a slight overestimation of the polarity of the carbon-chlorine bond, or due to a slight overestimation of hydrogen bond acidity of the hydrogen atoms polarized by neighbouring chlorine atoms. The residual RMSD for these two classes with respect to this regression line is only 0.29 kcal/mol. For a further analysis of these outliers, Figure 3 shows the σ-profiles of 2,3,4,5-tetrachlorobiphenyl and 2,2',6,6'-tetrachlorobiphenyl and of 2',3,4,5,5'-pentachlorobiphenyl and 2,2',4,6,6'-pentachlorobiphenyl. All four pcbs show very little differences in the positive σ region and small differences in the negative σ region. Especially the two pentachloro compounds show the same profile in the polar regions. It can also be seen that the 2,3,4,5-tetrachlorobiphenyl is less polar than the 2,2',6,6'-tetrachlorobiphenyl. The experimental data, however, show a completely different picture, the less polar 2,3,4,5-tetrachlorobiphenyl has a 1.2 kcal/mol lower free energy of hydration than its counterpart and is thus less hydrophobic. The two pentachloro compounds differ by 1.65 kcal/mol despite the fact that the σ profiles show almost the same polarity. Given the fact that all of the chlorinated compounds have been treated in a systematic way in our COSMO-RS workflow, we cannot find any possible reason why the COSMO-RS predictions of ΔG_{hydr} should have such different errors for as similar compounds as the two pairs of PCBs considered here. Taking into account the fact that the different experimental data reported in the supplementary material [6] for some of the PCBs scatter by up to 3 kcal/mol, we tend to address at least the two extreme outliers in the PCB data set to experimental error rather than considering them as indications for mispredictions by COSMO-RS.

For a complete analysis, however, some of the possible COSMO-RS error sources need to be discussed. Finding the lowest energy conformations in gas phase and COSMO is crucial for the prediction of ΔG_{hydr} as the quantum chemical energies enter the calculation. This is a complex problem for many classes of flexible molecules, but not at all for the pcbs. The pcbs have only one rotatable bond and the minima are very easily found by standard geometry optimization.

The used DFT method has an approximate inaccuracy of 0.5 to 1 kcal/mol if one takes the energy difference between gas phase and COSMO state, but even this error would be systematic for all pcbs and should thus result in a trend or shift. The last aspect which is obvious in the framework of the SMAPL3 challenge is the parameterization. A COSMO-RS parameterization has parameters adjusted to a large set of data. The most important parameters are general for all molecules and elements, some parameters are element specific (i.e. vdW parameters) or even element pair specific. No parameters, however, are especially adjusted to chlorinated or polychlorinated compounds and none have been changed for this paper, which means the expected prediction accuracy of COSMO-RS is rather universal. The above mentioned aspects underline that COSMO-RS might either systematically fail or succeed but that big random deviations within a single class of molecules are not likely.

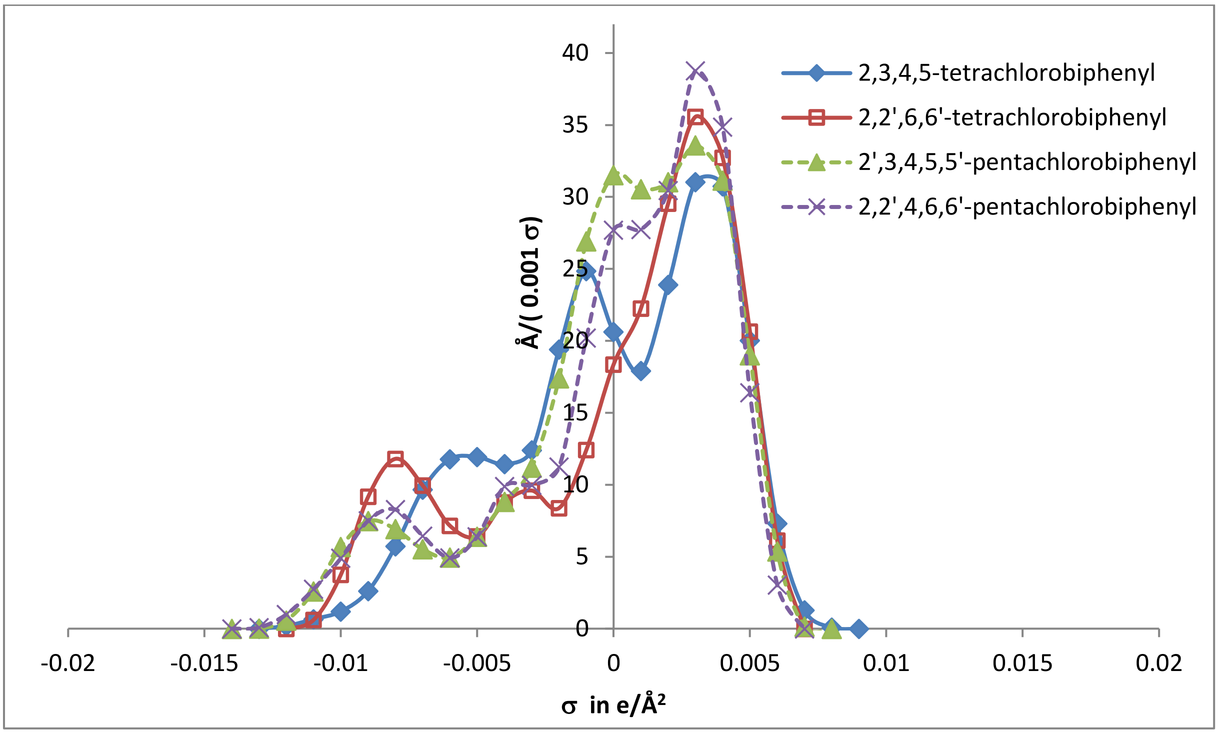

Fig. 3 σ profiles two tetrachloro pcbs and two pentachloro pcbs.

Conclusion

COSMO-RS as implemented in the COSMO*therm* software was able to yield a RMSE of 1.05 kcal/mol for the SAMPL3 dataset. Though the predictions were conducted after publication of the SAMPL3 results, no fitting or adjustments were made to the method and a limited comparison is therefore reasonable. In addition the computational effort was low, even though quantum chemical calculations were necessary. Apart from purely predicting the values, a qualitative analysis regarding the relevant interactions and their origin can be conducted. The chlorinated ethanes series is completely explainable within the picture of polarity, hydrogen bonding and vdW interaction as used by COSMO-RS.

Though giving good results, we are not satisfied as the error is much larger than typically observed for COSMO-RS. This large error can solely be attributed to mispredictions for the pcbs, whiles chlorinated ethane and dioxins stay within the typical COSMO-RS expectations. The analysis of the two teatrachloro and pentachloro pcbs showed large differences within the experimental data for similar compounds. The origin of this behaviour remains unclear as quantum chemically calculated properties do not indicate strong differences in polarity and hydrogen bonding and vdW forces should only give minor differences.

References

[1] Klamt A (1995) J. Phys. Chem. 99:2224

[2] Klamt A, Jonas V, Bürger T, Lohrenz J C W (1998) J. Phys. Chem. 102:5074

[3] Klamt A (2005) COSMO-RS: From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design, Elsevier, Amsterdam

[4] Klamt A, Eckert F, Diedenhofen M J, (2009) J. Phys. Chem. B 113:4508–4510

[5] Klamt A, Diedenhofen M J, (2010) J Comput Aided Mol Des, 24:357-360

[6] Geballe M T, Guthrie J P (2012) J Comput Aided Mol Des, DOI 10.1007/s10822-012-9568-8

[7] Beckstein O, Iorga B I, (2012) J Comput Aided Mol Des, DOI 10.1007/s10822-011-9527-9

[8] Klamt A, Schüürmann G J (1993) Chem. Soc. Perkins Trans. 2:799

[9] Barone V, Cossi M, Mennucci B, Tomasi J, (1998) J. Phys. Chem. A 102:1995

[10] Klamt A, Mennucci B, Tomasi J, Barone V, Curutchet C, Orozco M, Luque FJ. (2009) Acc Chem Res. 42:489-492