J Comput Aided Mol Des (2010) 24:621–625
DOI 10.1007/s10822-010-9332-x

Some conclusions regarding the predictions of tautomeric equilibria in solution based on the SAMPL2 challenge

Andreas Klamt · Michael Diedenhofen

Received: 5 March 2010 / Accepted: 12 March 2010 / Published online: 8 April 2010

Abstract The COSMO-RS method, a combination of the quantum chemical dielectric continuum solvation model COSMO with a COSMO based statistical thermodynamics of surface interactions, has been used in its COSMOtherm implementation for the direct, blind prediction of tautomeric equilibria within the SAMPL2 challenge. Since the quantum chemical level underlying COSMOtherm, i.e. BP/TZVP DFT-calculations, is known to be of limited accuracy with respect to reaction energies, we tested MP2 reaction energy corrections in addition. As expected, the straight application of the latest version of COSMOtherm yielded a poor predictive accuracy of $\sim$4 kcal/mol (RMSE) for the eight compounds of the blind prediction data set, and the MP2-corrected predictions reduced the average error considerably to $\sim$1.2 kcal/mol. But a more detailed analysis shows that this improvement is not systematic and mostly a lucky coincidence on the small data set. The systematic results of COSMOtherm allow for an efficient empirical correction with an RMSE of 0.61 kcal/mol. This allows for systematic predictions for the most important case of generalized keto-enol tautomerism.

A. Klamt (✉) · M. Diedenhofen
COSMOlogic GmbH&CoKG, Burscheider Str. 515, 51381
Leverkusen, Germany
e-mail: klamt@cosmologic.de

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

Keywords Tautomerization · Solvation · COSMO-RS

Introduction

Blind tests are important for the evaluation of the predictive power of computational models. Therefore, we take every opportunity to participate in blind tests and the SAMPL2 challenge [1] was a welcome opportunity to validate the predictive capabilities of our COSMO*therm* [2] implementation of the COSMO-RS method [3–5]. Our main focus was the XFER part of SAMPL2 which was on the prediction of free energy of hydration $\Delta G^{X}_{\mathrm{hydr}}$ of a number of demanding compounds with very large solvation energies. Our results and findings for that part have been reported in a previous paper [6].

Despite our limited experience in the area of reaction and tautomerization modelling, but since tautomerization is a very important phenomenon often going along with solvation, we decided to also participate in the tautomerization part TAUT09 of SAMPL09 on a more experimental basis. We decided to investigate two levels of calculation. The first level, abbreviated as CT-BP-TZVP further on, was just the straight forward usage of COSMOtherm based on BP [7, 8] DFT calculations with a TZVP basis set [9, 10]. Since DFT in general is known to be not very reliable for reaction energies, it was to be expected that this level would not be very accurate, but we wanted to consider it as a starting point for further improvements, especially since it is the basis for reliable predictions of solvation effects with COSMO-RS. Therefore, as a second level, we corrected the CT-BP-TZVP free energy differences by single point MP2 corrections for the gas energy differences, and for zero-point and thermal vibrational (and rotational) free energy contributions. This level will be abbreviated as MP2+vib-CT-BP-TZVP.

Data

The data sets for both parts of the SAMPL09 challenge, XFER09 and TAUT09, have been collected and assembled by the organizers of the challenge are described in detail in the introductory article of this special issue [1]. The tautomerization part TAUT09 in special is based on a data collection by Taylor.

The TAUT09 data set is split into three parts. Eight tautomeric equilibria were given as obscure. These were the core blind prediction examples for later comparison with experimental data known to the organizers and disclosed after the deadline of the challenge. Thirty three tautomerization equilibria were given as investigatory examples, but no experimental data exist for these examples. Since no comparison with experiment is possible for that part, we therefore will omit the investigatory examples in the present paper. Finally, the explanatory data set consisted of 12 tautomeric pairs with given experimental free energy differences, and one compound with 15 tautomeric forms together with experimental data of their relative free energies. The exploratory data set may have been, but in our study was not, used for calibration of the method. All tautomeric equilibria were considered at 298 K in aqueous solution.

Because the data for the one compound with 15 tautomers are less simple to compare and the free energies of the higher tautomers are less reliable, we will focus on the 8 plus 12 tautomeric equilibrium pairs of the obscure and explanatory data sets in this paper.

For the structures of the compounds see ref. [1]. The naming has been used as introduced by Taylor. Sixteen of the twenty tautomeric systems can be considered as generalized keto-enol equilibria, three as keto-keto equilibria, and one is a neutral-zwitterionic equilibrium (see Table 1). Thirteen of the sixteen keto-enol systems are written as enol to keto equibria in the SAMPL09 data set, three in keto to enol direction.

Methods

Standard BP/TZVP DFT geometry optimizations using the BP86 functional [7, 8] in combination with a TZVP basis set [9, 10] in vacuum and with the COSMO solvation model [11] have been performed for each of the involved compounds, combined with a conformational search for the lowest energy conformations in gas phase and solution with the COSMOconf procedure [12].

Table 1 Tautomeric free energy differences considered in this paper

System	Set	Class	Subclass	CT-BP-TZVP	MP2+vib-CT-BP-TZVP	Experiment
1A_1B	o	ek	6-ring-oxohetcyc.	−8.90	−4.01	−4.80
2A_2B	o	ek	6-ring-oxohetcyc.	−9.92	−5.74	−6.10
3A_3B	o	ek	6-ring-oxohetcyc.	−11.81	−7.71	−7.20
4A_4B	o	ek	6-ring-oxohetcyc.	−5.95	0.46	−2.30
5A_5B	o	ek	6-ring-oxohetcyc.	−8.04	−3.94	−4.80
6A_6B	o	ek	6-ring-oxohetcyc.	−13.20	−7.61	−9.20
10B_10C	e	ek	5-ring oxohetcyc.	−5.63	1.70	−2.90
11D_11C	e	ek	5-ring oxohetcyc.	−3.39	4.64	−0.50
12D_12C	e	ek	5-ring oxohetcyc.	−3.74	3.30	−1.80
13D_13C	e	ek	5-ring oxohetcyc.	−3.12	4.04	0.10
14D_14C	e	ek	5-ring oxohetcyc.	−2.85	1.91	0.30
15A_15B	e	ke	5-ring oxohetcyc.	3.85	−2.59	0.90
10D_10C	e	ek	5-ring oxohetcyc. (nnn)	−3.80	3.78	−1.20
15B_15C	e	ek	5-ring oxohetcyc. (nnn)	−5.34	1.83	−2.20
8A_8B	e	ke	diketo	−3.35	−1.59	−3.00
7A_7B (outl.)	e	ke	diketo	3.58	5.28	7.00
15A_15C	e	kk	5-ring oxohetcyc.	−1.49	−0.76	−1.20
16A_16C	e	kk	5-ring oxohetcyc.	−1.82	−0.26	0.50
5B_5C	o	kk	6-ring-oxohetcyc.	−0.67	−1.72	0.50
6A_6Z	o	ez	6-ring-oxohetcyc.	−8.26	−3.43	−2.40

For the structures see ref. [1]. Column 2 denotes the data sets obscure (o) and explanatory (e). In column 3 ek means enol to keto, ke means keto to enol, kk stands for a tautomerization between two generalized keto forms, and ez denotes a tautomerization from an enol to a zwitterionic form. Column 4 gives a subclassification, with nnn denoting systems where the hydrogen atom is shifted to the next-nearest ring position

For the MP2+vib-CT-BP-TZVP level single point MP2 calculations on the BP-TZVP gas phase geometries were performed with the QZVPP basis [13, 14] and gas phase zero-point and finite temperature (298 K, 1 bar) free energy contributions have been calculated based on the harmonic BP/TZVP gas phase frequencies. All quantum chemical calculations have been performed with the TURBOMOLE program package, version 6.0 [15].

The tautomerization free energy differences for the CT-BP-TZVP level were directly calculated as differences of the COSMO-RS [3–5] free energy differences in water, using only the COSMO files from the BP/TZVP/COSMO geometry optimizations. The COSMO-RS calculations have been performed with the COSMOtherm [2].

For the MP2+vib-CT-BP-TZVP level we considered a thermodynamic cycle, starting with the tautmeric form $T1_{aq}$ of a compound in solution, calculated as described above. From there we consider desolvation to the gas-phase state $T1_{gas}$, taking the desolvation free energy from COSMOtherm. Next we consider the gas-phase reaction from tautomeric form $T1_{gas}$ to form $T2_{gas}$ by the MP2/QZVP//BP/TZVP calculations augmented with the vibrational and thermal corrections, and finally we go to the aqueous state of $T2_{aq}$ using the COSMOtherm solvation energy of T2.

Results and discussion

Table 1 and Fig. 1 present our results for the 8 and 12 dual tautomeric free energy differences of the obscure and explanatory TAUT09 subsets for the CT-BP-TZVP level, submitted as SAMPL09 entry 319, and the MP2+vib-CT-BP-TZVP level, submitted as SAMPL09 entry 320, respectively.

As expected, we do see a large systematic deviation (mean deviation) of −3.8 kcal/mol for the eight compounds of the blind prediction (obscure) data set for the bare CT-BP-TZVP method, resulting in a root mean squared deviation (RMSD) of 4.0 kcal/mol. The mean deviation reduces to only 0.33 kcal/mol for MP2+vib-CT-BP-TZVP method, with a RMSD of 1.5 kcal/mol. Hence the MP2+vib-CT-BP-TZVP method was ranked much better in the SAMPL09-TAUT blind test, i.e. rank 7 out of 17 entries, compared to the CT-BP-TZVP method, which got rank 14. Nevertheless, already on this data set it should be noted that despite of the large shift the CT-BP-TZVP data show a better correlation with the experimental data ($r^2 = 0.90$) compared to the MP2+VIB-CT-BP-TZVP method ($r^2 = 0.75$).

If we extend the picture to all 20 tautomeric systems, we find a reduction of the mean deviation to −2.7 kcal/mol and a reduction of the RMSD to 3.35 kcal/mol for the CT-BP-TZVP method, while the mean deviation increases to 1.4 kcal/mol for the MP2+VIB-CT-BP-TZVP method, at essentially unchanged RMSD. Noticeably, the correlation coefficients reduce to 0.85 and 0.58, respectively. Hence the absolute advantage of the MP2+VIB-CT-BP-TZVP method compared to the CT-BP-TZVP method decreases on the full data set, and the correlation is much better for the CT-BP-TZVP method.

This observation motivated us to perform a more rigorous analysis of the data. Sixteen of the twenty tautomeric pairs can be classified as generalized keto-enol tautomerization, i.e. a reaction from R1–X(H)–C(=O)–R2 to R1–X=C(OH)–R2, where X may be a N or CH. Surprisingly, 13 of the 16

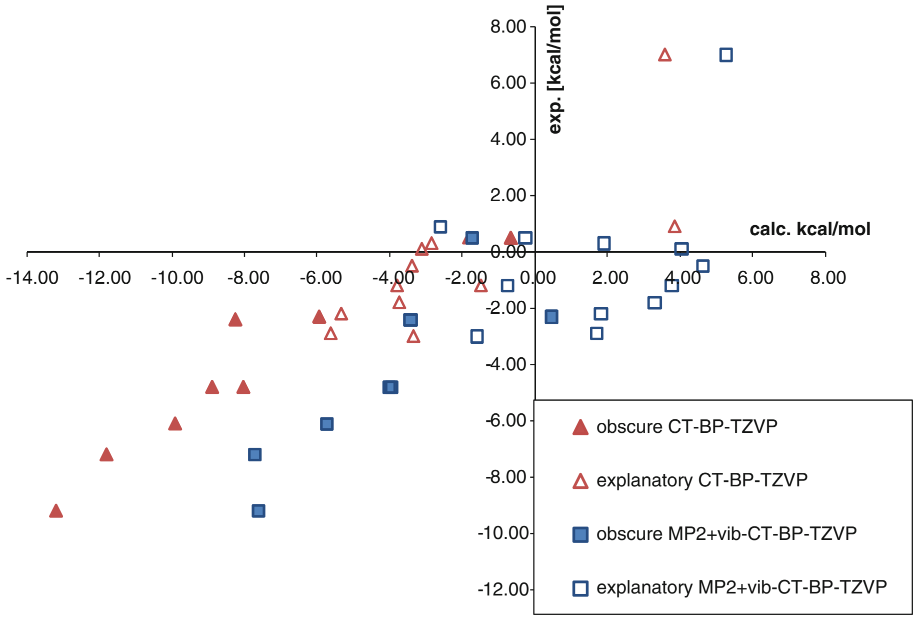

Fig. 1 Comparison of Experiment tautomerization free energies with calculated results for the CT-BP-TZVP and MP2+vib-CT-BP-TZVP, respectively, for the obscure and explanatory subsets

cases are given as enol to keto (ek) equilibrium in the TAUT09 data set, while three are given the other way round, i.e. as keto-enol equilibrium (ke). The other four cases consist of three equilibria between different keto forms and one equilibrium between an enol form and a zwitterionic form. If we focus on the 16 keto-enol type cases and convert the cases given as three ke to ek for consistency, we get a rather different picture as shown in Figs. 2 and 3.

The CT-BP-TZVP data show an excellent correlation ($r^2 = 0.97$) with the experimental tautomerization free energies, if we leave out diketone, which is a strong outlier. Indeed, Taylor [1] considers the experimental value for the diketone tautomerization as a crude estimate, just indicating that the equilibrium should be far on the keto side. Hence all further discussion is based on the remaining 15 cases. These still cover a large diversity of chemistry, ranging from the very common case of 1,2-keto-iminol tautomerism in heterocyclic 6-rings, over the same in heterocyclic 5-rings, 2 cases of of 1,3-keto-iminol-tautomerization in 5-rings, i.e. a hydrogen transfer to the next-nearest neighbor (nnn), one case of true keto-enol equilibrium in a heterocyclic 5-ring system, and the keto-enol tautomerization of cyclohexadione. The RMSD of this regression is as good as 0.61 kcal/mol. The correlation coefficient r^2 reduces to 0.68, if we just use gas phase BP-TZVP energy differences. Hence, the solvation contribution of COSMOtherm is crucial for the good correlation.

The surprising result is the much worse correlation ($r^2 = 0.65$) which we find on the MP2+vib-CT-BP-TZVP method, since this level was expected to be more accurate for reaction energy differences than the CT-BP-TZVP method. In order to analyze the origin of the much larger scatter, we tested a combination of the vib-corrections with CT-BP-TZVP. This slightly decreased the correlation to $r^2 = 0.95$, but the large decrease of correlation, i.e. the large scatter, appears to result from the MP2 correction.

Summary and conclusions

In order to benchmark our standard COSMOtherm solvation level CT-BP-TZVP we have participated in SAMPL09 solvation and tautomerization bind tests. While for the solvation part (XFER09) CT-BP-TZVP yielded the most accurate predictions, for the aqueous tautomerization free energies, which essentially are a combination of reaction energies in gas phase and solvation, the bare CT-BP-TZVP level yielded large errors of ~4 kcal/mol. This outcome was expected due to the known inaccuracies of DFT with respect to reaction energies. In order to crosscheck the DFT results with an ab initio method, we supplemented the BP/TZVP energies with MP2/QZVP//BP-TZVP reaction energy corrections and vibrational corrections on the BP/TZVP level. On a first glance this appeared to improve the

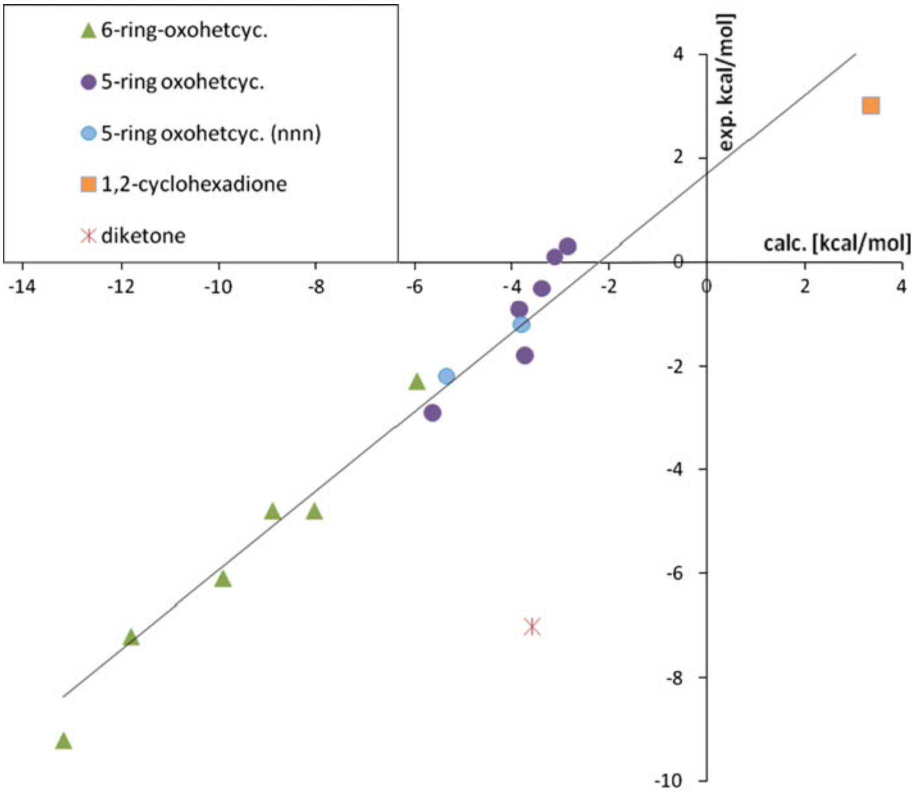

Fig. 2 Experimental tautomerization free energies of enol-keto type tautomeric equilibria versus calculated values from the CT-BP-TZVP method. Diketone is considered as outlier and is not included in the regression

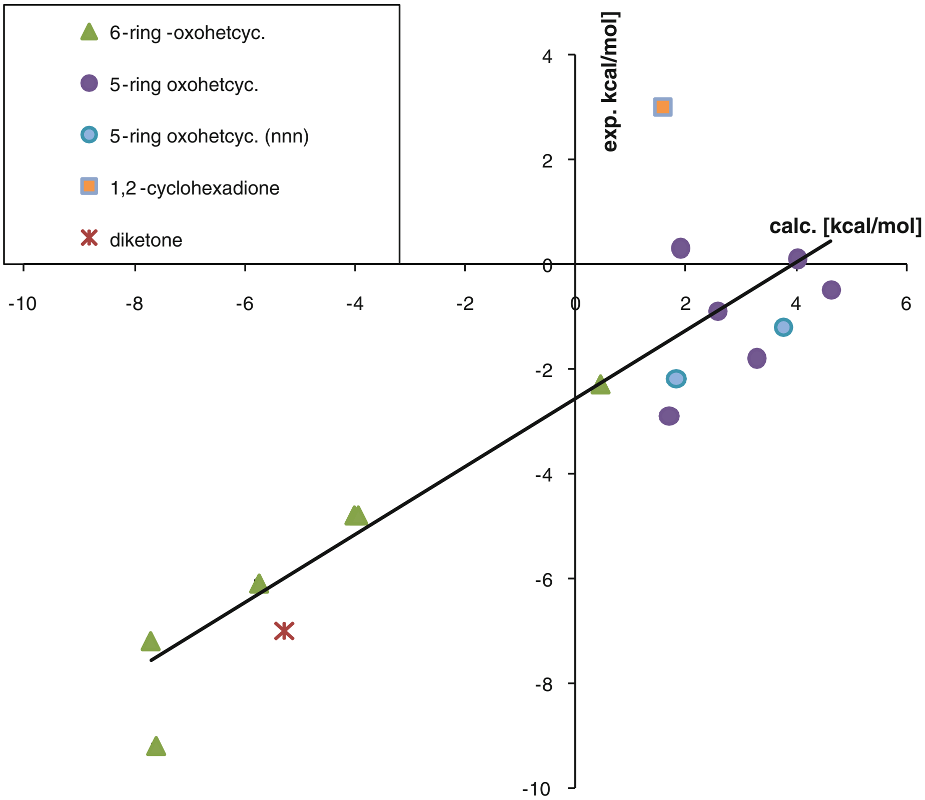

Fig. 3 Experimental tautomerization free energies of enol-keto type tautomeric equilibria versus calculated values from the MP2+vib-CT-BP-TZVP method. Diketone is considered as outlier and is not included in the regression

accuracy of the tautomerization free energy predictions considerably on the small TAUT09 blind test set, but a closer analysis focussing on those 16 of the 20 examples that can be considered as generalized keto-enol tautomer equilibria, showed that the CT-BP-TZVP results yield a very good correlation with the experimental data, while the MP2+vib-CT-BP-TZVP level does not show such a systematic deviation. Hence, finally the simpler CT-BP-TZVP level combined with one empirical correction taken from a regression analysis appears to yield very good results for aqueous tautomerization equilibria of generalized keto-enol tautomeric systems. Due to the transferability of the COSMO-RS method with respect to solvents, the combination of CT-BP-TZVP with this empirical correction should also allow for the quantitative prediction of the tautmeric equilibria in other solvents.

References

1. Geballe MT, Skillman AG, Nicolls A, Guthrie JP, Taylor PJ (2010) J Comp Aided Mol Design 24. doi:10.1007/s10822-010-9350-8
2. Eckert F, Klamt A (2008) COSMOtherm, Version C2.1-Revision 01.08; COSMOlogic GmbH&CoKG. Leverkusen, Germany, see also. http://www.cosmologic.de
3. Klamt A (1995) J Phys Chem 99:2224
4. Klamt A, Jonas V, Bürger T, Lohrenz JCW (1998) J Phys Chem 102:5074
5. Klamt A (2005) COSMO-RS from quantum chemistry to fluid phase thermodynamics and drug design. Elsevier, Amsterdam
6. Klamt A, Diedenhofen M (2010) J Comp Aid Mol Des. doi: 10.1007/s10822-010-9354-4
7. Becke AD (1988) Phys Rev A 38:3098
8. Perdew JP (1986) Phys Rev B 33:8822
9. Schäfer A, Huber C, Ahlrichs R (1994) J Chem Phys 100:5829
10. Eichkorn K, Weigend F, Treutler O, Ahlrichs R (1997) Theor Chem Acc 97:119
11. Klamt A, Schüürmann G (1993) J Chem Soc Perkins Trans 2:799
12. Klamt A, Eckert F, Diedenhofen MJ (2009) Phys Chem B 113:4508–4510
13. Weigend F, Furche F, Ahlrichs R (2003) J Chem Phys 119:12753
14. Haettig C (2005) Phys Chem Chem Phys 7:59
15. TURBOMOLE (1989–2007) A development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH. TURBOMOLE GmbH, since 2007, see also http://www.turbomole.com