



# COSMO-RS blind prediction of distribution coefficients and aqueous pKa values from the SAMPL8 challenge

Michael Diedenhofen<sup>1</sup> · Frank Eckert<sup>1</sup> · Selman Terzi<sup>1</sup>

Received: 9 January 2023 / Accepted: 18 June 2023 / Published online: 27 June 2023  
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2023

## Abstract

The SAMPL8 blind prediction challenge, which addresses the acid/base dissociation constants (pKa) and the distribution coefficients (logD), was addressed by the Conductor like Screening Model for Realistic Solvation (COSMO-RS). Using the COSMOtherm implementation of COSMO-RS together with a rigorous conformational sampling, yielded logD predictions with a root mean square deviation (RMSD) of 1.36 log units over all 11 compounds and seven bi-phasic systems of the data set, which was the most accurate of all contest submissions (logD).

For the SAMPL8 pKa competition, participants were asked to report the standard state free energies of all microstates, which were then used to calculate the macroscopic pKa. We have used COSMO-RS based linear free energy fit models to calculate the requested energies. The assignment of the calculated and experimental pKa values was made on the basis of the popular transitions, i.e. the transition that was predicted by the majority of the submissions. With this assignment and a model that covers both, pKa and base pKa, we achieved an RMSD of 3.44 log units (18 pKa values of 14 molecules), which is the second place of the six ranked submissions. By changing to an assignment that is based on the experimental transition curves, the RMSD reduces to 1.65. In addition to the ranked contribution, we submitted two more data sets, one for the standard pKa model and one for the standard base pKa model of COSMOtherm. Using the experiment based assignment with the predictions of the two sets we received a RMSD of 1.42 log units (25 pKa values of 20 molecules). The deviation mainly arises from a single outlier compound, the omission of which leads to an RMSD of 0.89 log units.

**Keywords** COSMO-RS · COSMOtherm · logD prediction · pKa prediction · Partitioning

## Abbreviations

COSMO-RS	Conductor like Screening Model for Realistic Solvation
COSMO	Conductor like Screening Model
DMF	Dimethylformamide
LFER	Linear free energy relationship
MAD	Mean absolute deviation between predicted and experimental data
MEK	Methyl ethyl ketone
RMSD	Root mean square deviation between predicted and experimental data
TBME	Tert butyl methyl ether

## Introduction

The Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) is a series of blind challenges aimed to improve the accuracy of computational models to predict physical properties relevant for modern rational drug design. Relevant data assessed during the course of the SAMPL challenge series are free energies of binding, hydration free energies, protonation equilibria, as well as partition and distribution equilibria. Among the latter, the distribution/partition coefficient of the active pharmaceutical ingredients (APIs) are of particular interest, as it provides guidance on the rational solvent selection for purification [1, 2] and can also be used to estimate the API's distribution between the compartments of the human body [3]. Other areas where the distribution coefficients are of interest are the modeling of environmental properties like bioaccumulation [4], plastics recycling [5–7], and other types of technical extraction processes [8–10]. The assessment of prediction methods for

✉ Michael Diedenhofen  
Michael.diedenhofen@3ds.com

<sup>1</sup> Dassault Systèmes Deutschland GmbH, Am Kabellager 11-13, 51063 Cologne, Germany

practical application provides important information for the user who want to use prediction tools, but it also makes an important contribution to the development and evaluation of physical property prediction methods. Blind prediction challenges like SAMPL provide the rare opportunity to benchmark a model under real conditions and ensure a fair assessment of the different methods.

Since the COSMO-RS method was originally developed for the prediction of partition coefficients and Henry's law constants [11], there is a long history of COSMO-RS contributions to the SAMPL blind prediction challenge series. The results underline the good quality of COSMO-RS derived molecular free energies in solution [12–15].

The recent SAMPL8 challenge contained two categories: the prediction of host–guest binding affinities and a physical property challenge, focusing on the prediction of distribution coefficients (logD) and acid dissociation constants (pKa) of a series of drug molecules. The experimental data collection of the latter part was provided by GlaxoSmithKline [1] and contains pKa data for twenty-three drug molecules, as well as logD values for eleven of those compounds. The distribution data was determined for seven bi-phasic systems, ranging from un-polar/polar combinations like heptane/water to more hydrophilic organic phases like methyl ethyl ketone or octanol, where the water solubility in the organic phase has to be taken into account. The SAMPL8 compound set is a diverse set (see Fig. 1) with only two groups of molecules that share the same scaffold with varying substituents. These are three 2-aminobenzimidazole derivatives (SM8-7, SM8-9, SM8-17) and two 2-chloroquinazolin-4-amines (SM8-15, SM8-18).

Here we report the results of the COSMO-RS method for the blind prediction of the SAMPL8 acid dissociation constants and distribution coefficients.

## Computational methods

The pKa and logD predictions were done with the Conductor like Screening Model for Realistic Solvation (COSMO-RS) [16]. COSMO-RS is a statistical thermodynamics model that describes intermolecular interactions from an ensemble of pair-wise interacting surface segments. The surface interaction terms are based on the screening charge density of the Conductor like Screening Model (COSMO) [17], which is a continuum solvation model commonly used in computational quantum chemistry. Thus, we have a two-step procedure that starts with the COSMO calculation, followed by the COSMO-RS prediction. The time-consuming quantum chemical COSMO calculations need to be done only once per compound. Technically, the results of the COSMO calculations are saved in the so-called COSMO files, which

can be stored in a database. For a detailed description of the COSMO-RS method, please refer to [18] and [19].

We started our study with the generation of the (de)protonated states and tautomeric states of the SAMPL8 molecules and their conformers as described in reference 15. The only difference to the process described in reference 15 is that in addition to the tautomers, the (de)ported states were also generated. In this procedure, we used version 21 of COSMOconf [20] and COSMOquick [21, 22] software together with version 7.5 of the quantum chemistry software TURBOMOLE [23]. As a result, we obtained a set of COSMO files for the conformers of each microstate, which can be used for the following COSMO-RS calculations. For the logD predictions, the conformers of the neutral microstates were merged to form the conformer set of the compound. The number of microstates and conformers is given in Table S4 of the Supplementary Material. The COSMO files of the solvents used in this study were taken from the COSMObase 21 database [24]. All COSMO-RS calculations were done with the COSMOtherm program version 21 using the BP\_TZVPD\_FINE\_21 parametrization [25].

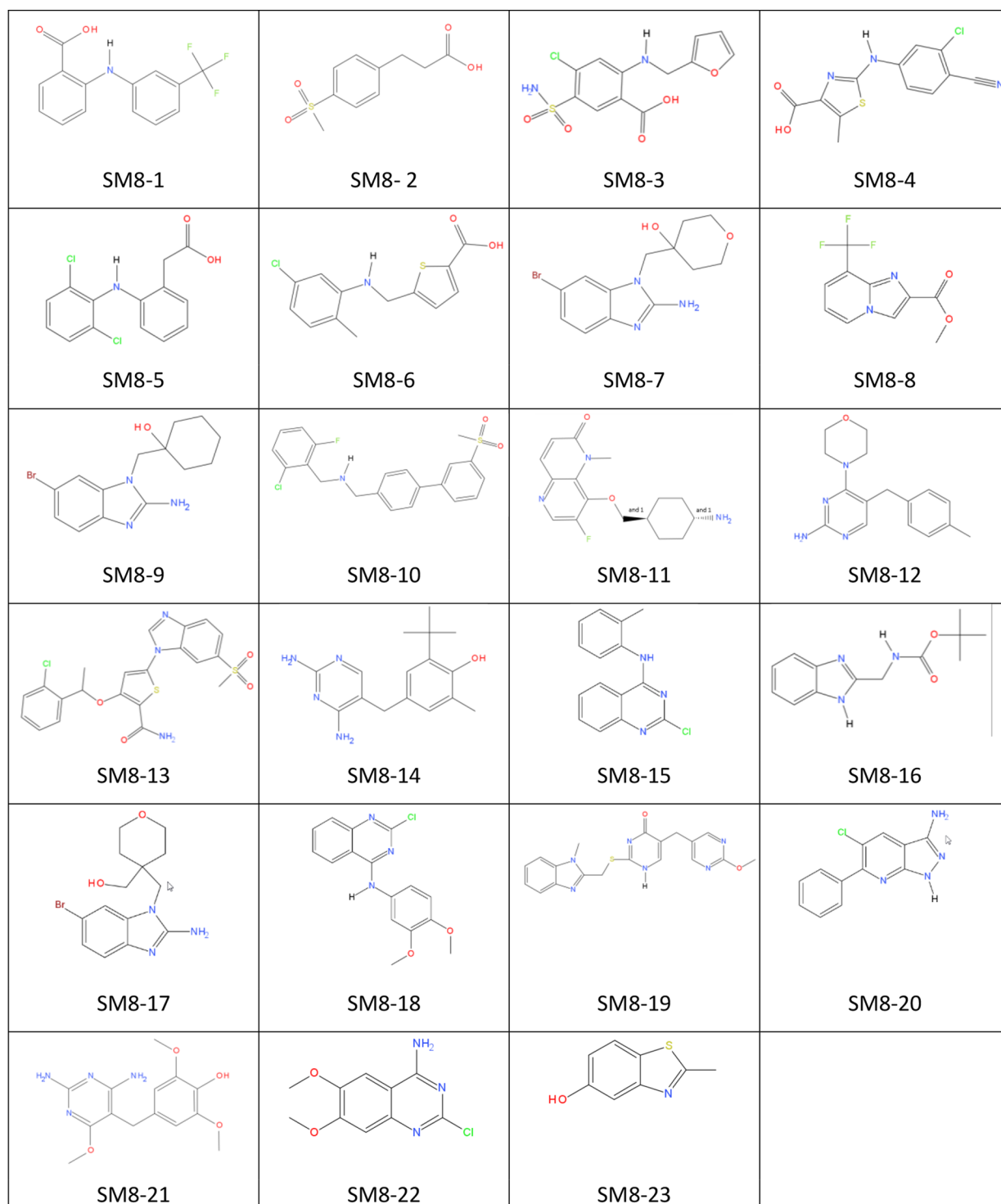
## pKa

The aqueous dissociation constant of an acid (pKa), or the conjugated acid of a base (base-pKa) can be estimated from a linear free energy relationship (LFER) that connects the COSMO-RS predicted free energy  $\Delta G_i$  to the dissociation constant pKa by two adjustable parameters, slope  $c_1$  and shift  $c_0$  [26, 27].

$$pK_a^i = c_1 \Delta G_i + c_0 \quad (1)$$

For the pKa of an acid,  $\Delta G_i$  is the free energy difference between the anion and the neutral state. In case of the base pKa, the free energy difference between the neutral state and the cation is used. The LFER parameters are specific to the solvent and the reacting system, i.e. an optimal correlation (and thus best prediction quality) is reached for two independent sets of LFER parameters, one for the (acid) pKa and one for the base pKa.

As in the SAMPL7 [28] challenge, the standard state free energies approach of Gunner et al. [29] should be used to describe the (de)protonation equilibria of the compounds under consideration. These relative free energies can be directly calculated from the LFER given in Eq. 1. The participants were asked to submit the relative free energies of the all microstates (neutral and ionic) with respect to a given reference state. In the evaluation, this information was used to calculate titration curves and the macroscopic pKa values as described in reference [29]. The assignment to the experimental data was done with the help of the popular



**Fig. 1** Molecules of the SAMPL8 logD and pKa challenge

transition, i.e. the protonation transition, which was found by the majority of the submissions [30].

In order to determine a consistent set of standard state free energies for the entire system, from the protonated cations to the anions, we used a new combined LFER model fit that

covers both (acid) pKa and base pKa data. This model was called "one pKa fit" in the submission, and will be referred to as the "unified" pKa model in the following<sup>1</sup>. Besides

<sup>1</sup> In cases where the full protonation range could be described by the COSMOtherm pKa (SAMPL8-2) or the base pKa model (SAMPL8-10,

this, we prepared two additional submissions, one with the (acid) pKa LFER fit [26], and one with the base pKa LFER fit [27], both of which are available in COSMOtherm. In addition to the relative free energies, we have also submitted the optional macroscopic pKa data.

## logD

The distribution coefficient, commonly used in the logarithmic form logD, describes the distribution of a compound between the two liquid phases of a bi-phasic system. It is defined as the ratio of the sum of the concentrations of all forms of the compound in the two phases. In contrast to the partition coefficient logP, which considers only the neutral solute, the logD considers the sum of dissociated and non-dissociated species.

$$\log D^{(2,1)} = \log_{10} \left( \frac{[\text{neutral} + \text{ionic in phase 2}]}{[\text{neutral} + \text{ionic in phase 1}]} \right) \quad (2)$$

In this study, we examined 6 bi-phasic system that consist of an organic phase and a water phase and one system with two organic phases (cyclohexane/DMF).

Using the assumption that the dissociated species will not migrate into the organic solvent phase, we can calculate the fraction of the dissociated solute in the aqueous phase from the dissociation constant pKa and the pH. As a result, we obtain the logD from the logP and a dissociation correction for monoprotic acids or bases.

SAMPL8-11, SAMPL8-15, SAMPL8-8), these models were used instead of the unified model.

**Table 1** pKa and pH values used for the logD dissociation correction

Compound	pKa <sup>b</sup>	Type	pH	Dissociation correction [log units] <sup>a,b</sup>
SM8-1	5.01(3.99)	acid	8	-2.99 (-4.01)
SM8-3	4	acid	8	-4.00
SM8-5	3.92 (4.23)	acid	8	-4.08 (-3.77)
SM8-6	4.17	acid	8	-3.83
SM8-7	6.63	base	3	-3.63
SM8-9	6.08	base	3	-3.08
SM8-10	7.71	base	3	-4.71
SM8-12	6.98	base	3	-3.98
SM8-14	7.27	base	3	-4.27
SM8-16	5.1	base	3	-2.10
SM8-17	6.58	base	3	-3.58

a)  $-\log_{10} \left( 1 + 10^{\Delta_{acid/base}} \right)$  of Eq. 3.

b) pKa measured by Gretz, Czodrowski, Tielker and Kast [30] and corresponding correction are given in parenthesis.

$$\log D^{(org.,water)} = \log P^{(org.,water)} - \log_{10} \left( 1 + 10^{\Delta_{acid/base}} \right) \quad (3)$$

$$\Delta_{acid} = pH - pK_a (acid)$$

$$\Delta_{base} = pK_a (base) - pH$$

In case of the cyclohexane/DMF system we did not consider dissociation and the logD values correspond to the logP values of the neutral solutes. For the dissociation corrections, the experimental values provided by the SAMPL8 organizers were used (see Table 1). The pH=8 was used for SM8-1,3,5,6 and pH=3 for the rest.

The neglect of the dissociation in the organic phases with significant water content should be a valid assumption for common acids and bases [33]. Scott and Clymer estimate a logD error of ~0.3 log units for a pH - pKa difference of 3 units and a ratio of the partitioning of the neutral and the ionic form of 0.001 [33]. The pH - pKa difference of some cases of this study is higher than the 3 units. However, we do not have reliable data for the partitioning ratio to estimate the error.

Since the goal of the SAMPL 8 challenge was to predict the partition coefficient, which is defined for the salt-free phases, we did not consider the influence of counter ions discussed in the literature [31, 32]. The background ions introduced by the aqueous buffers used in the experimental work influence the distribution coefficients but the effect depend on the ion type [31] and thus contradicts a general definition of the distribution coefficients. Nevertheless, this discrepancy is a source of deviations between calculation and experiment.

The partition coefficient used in Eq. 3 was calculated from the difference of the solute chemical potentials at infinite dilution in the two phases  $\mu_i^{(water/org.)}$ .

$$\log_{10} \left( P_i^{(org.,water)} \right) = \left( \mu_i^{(water)} - \mu_i^{(org.)} \right) / RT \ln(10) + \log_{10} (VQ) \quad (4)$$

In Eq. 4 the quotient of the molar volumes of the solvent phases  $VQ = V^{(water)} / V^{(org.)}$  is used to convert the partition coefficient from the mole fraction framework to a molar concentration-based definition. If available, experimental densities were utilized to obtain the molar solvent phase volume quotient  $VQ$ . Otherwise, COSMOtherm estimates were used. For heptane/water and cyclohexane/water systems we assumed pure phases. For the remaining systems the mutual solubility of the solvents was taken into account. The phase compositions used are listed in Table 2. All logP values were calculated at 25 °C.

## Results and discussion

### pKa predictions

Among the six ranked SAMPL8 entries, COSMO-RS (using the “unified” pKa LFER) and the Deep Gaussian Process submission provide the best predictions [30]. COSMO-RS showed the better mean absolute deviation (MAD) of 2.49 compared to 2.62 for the Deep Gaussian Process method. Considering the root mean square deviations (RMSD) the Deep Gaussian Process submission showed a lower deviation of 3.17 compared to COSMO-RS with 3.44. The SAMPL8 evaluation was done on a data set with 16 pKa values for 14 molecules and used the popular transition for the assignment of the experimental values. This assignment method was used because it provides an objective criterion for the mapping of calculation on experiment and should avoid predictions that are good for the wrong reasons, i.e. when a value is predicted correctly, but is assigned to the wrong transition (private communication D. L. Mobley 2022). It is assumed that the transition, neutral to anion or cation to neutral, found by the majority of the submissions is the dominant one for the experiment, which works well whenever the correct transition is popular, and badly when it is not. A summary of the SAMPL8 COSMO-RS results is given in Fig. 2.

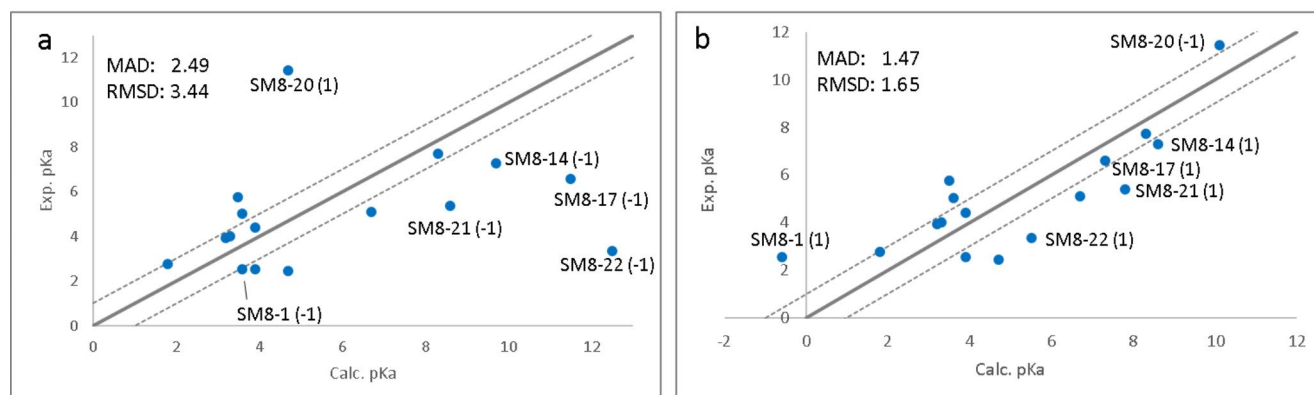
The plot of the original assignment of transitions done by SAMPL8 depicted in Fig. 2a shows a number of compounds where the calculations deviate several pKa units from the experiment. Two of them, SM8-1 and SM8-20, have two experimental pKas, which were compared with the same calculated value, i.e. the same transition. For SM8-20 this results in a large deviation, whereas the deviations for SM8-1 are moderate. Since the deviations for SM8-14,17,21,22 were larger than expected, we checked the experimental titration curves (pH dependent solubilities)

**Table 2** Compositions of the bi-phasic systems used for the logP calculations

Bi-phasic System	Water (DMF) content of the phases [mol %]	Reference
octanol/water	27.0 / 100.0	Dallos et al. 1995 [34]
cyclohexane/water	0.0 / 100.0	
ethylacetate /water	20.6 / 98.0	COSMOtherm prediction
heptane/water	0.0 / 100.0	
MEK/water	35.1 / 91.8	Lladosa et al. 2011 [35]
TBME/water	7.4 / 99.2	Ashour 2005 [36]
cyclohexane /DMF	5.3 / 83.1	COSMOtherm prediction

provided by the organizers [30]. To determine the transitions, we assumed that the neutral organic substances have a lower water solubility than the ionic species and exist in the pH range considered. Thus, they represent the minimum of the solubility curves. Starting from the solubility minimum, we now consider pKas that are at a lower pH value as base pKa (cation to neutral transition) and pKas that are at a higher pH as acid pKas (neutral to anion transition). With this assignment, we obtain the results shown in Fig. 2b. The large deviations have been reduced and the MAD has decreased to 1.47 (RMSD 1.65). Interestingly, the deviation for the base pKa of SM8-1 has increased.

At this point, we were interested in how the individual aqueous (acid) pKa and base pKa models of COSMOtherm perform on the SAMPL8 dataset. These two models were submitted outside the official evaluation as “COSMO-RS base pKa fit” for bases and “COSMO-RS pKa fit” for acids. Depending on the experiment-based transition assignment, the prediction of the (acid) pKa or base pKa model was compared with the experiment. The resulting dataset is bigger than the set used for the original SAMPL8 evaluation described above. The results for the 25 pKa values of 20



**Fig. 2** COSMO-RS (unified model) results of the SAMPL8 pKa challenge evaluation of 16 pKa values for 14 molecules. Dashed lines mark the corridor of 1 pKa unit. The charge of the ionic state is given in

parentheses. **a:** Original SAMPL8 assignment of experiment and calculation. **b:** The assignment of experimental and calculated values was derived from the experimental pH dependent solubilities

molecules are shown in Fig. 3 and Table S1 in the Supplementary Material.

Apart from the SM8-1 cation to neutral transition, the predictions are in good agreement with the experiment. Nevertheless, due to the SM8-1 outlier the RMSD of 1.42 (MAD 0.95) is higher than expected. The RMSDs of the pKa and base pKa fit sets are 0.49 pKa units for the aqueous acid pKa LFER fit [26], and 0.56 pKa units for the base pKa LFER fit [27], respectively.

A careful check of the SM8-1 COSMO files did not reveal any significant anomalies. Two energetically unfavorable isomers that remained in the set do not change the results (see Table S3 of the Supplementary Material). We tried to optimize a zwitterionic tautomer, but could not detect a stable structure. Besides COSMO-RS there are two other SAMPL8 submissions (“EC\_RISM” and “RFE-uESE-extra”) that predict two pKa values for the amphoteric SM8-1 compound. Similar to the COSMO-RS prediction, these submission also show negative pKa values (EC\_RISM: -2.5; RFE-uESE-extra: -11.1) for the cation-neutral transition and thus a large deviations from the experimental value of 2.54. The re-measurement of the acid constants of SM8-1, 2, 5, 22 by Gretz, Czodrowski, Tielker and Kast from the TU Dortmund University [30] did not change the overall picture, but for SM8-1 they report a pKa for the neutral-anion transition of 3.99 and no base pKa above 2 pKa-units.

If we omit the SM8-1 base pKa outlier, we obtain a good agreement with the experiment with a MAD of 0.75 and an RMSD of 0.89. This result is within the expected range for molecules that are not present in the model’s fit set.

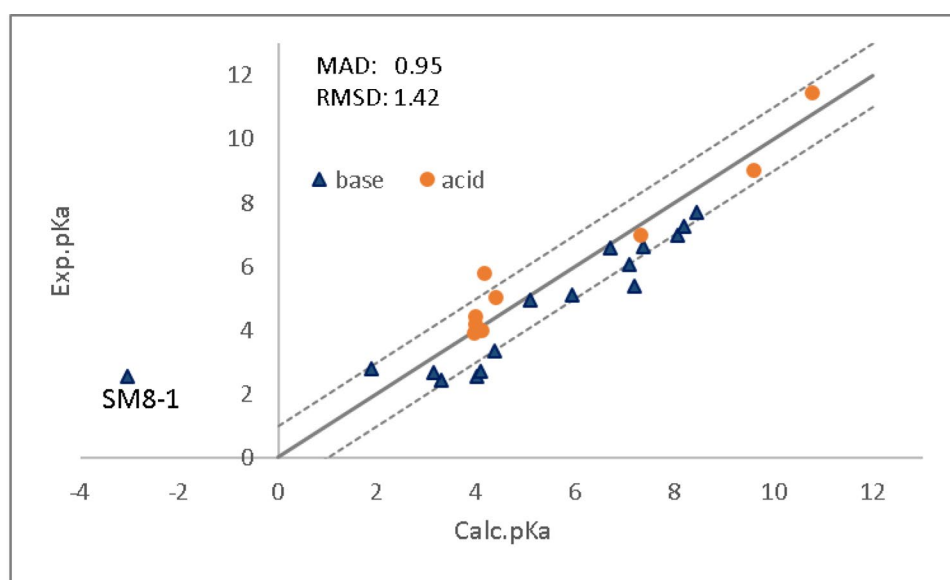
## logD prediction

The SAMPL8 blind challenge counted 17 logD submissions [30]. Five of them were submitted to be included in the ranking. This subset consist of four methods classified as physical approaches that include quantum chemical calculations and one that uses a machine learning model built on molecular dynamics simulations. As can be seen from the overall deviations in the Supplementary Material Fig. S1, the COSMO-RS method provided the most accurate predictions in this challenge. The mean absolute deviation (MAD) over all bi-phasic systems and molecules is 1.07 log units and the root mean square deviation (RMSD) for the same data set is 1.36 log units. Looking at the MADs of all submissions and bi-phasic systems COSMO-RS can be found on rank one for the MEK/water and ethylacetate/water systems (see Supplementary Material Fig. S2). For cyclohexane/DMF, heptane/water, octanol/water and cyclohexane/water the COSMO-RS submission is on rank two, sharing this second place for cyclohexane/water with the “EC\_RISM\_logDexp” submission. The COSMO-RS predictions for the TBME/water system are better than the ones for cyclohexane/DMF, but the deviation of three other submission are smaller which puts the method in fourth place.

The MAD and RMSD deviations for the COSMO-RS logD predictions for all bi-phasic systems are given in Table 3. Figure 4 shows the comparison against the experiment. The calculated and experimental values are listed in the Supplementary Material Table S2.

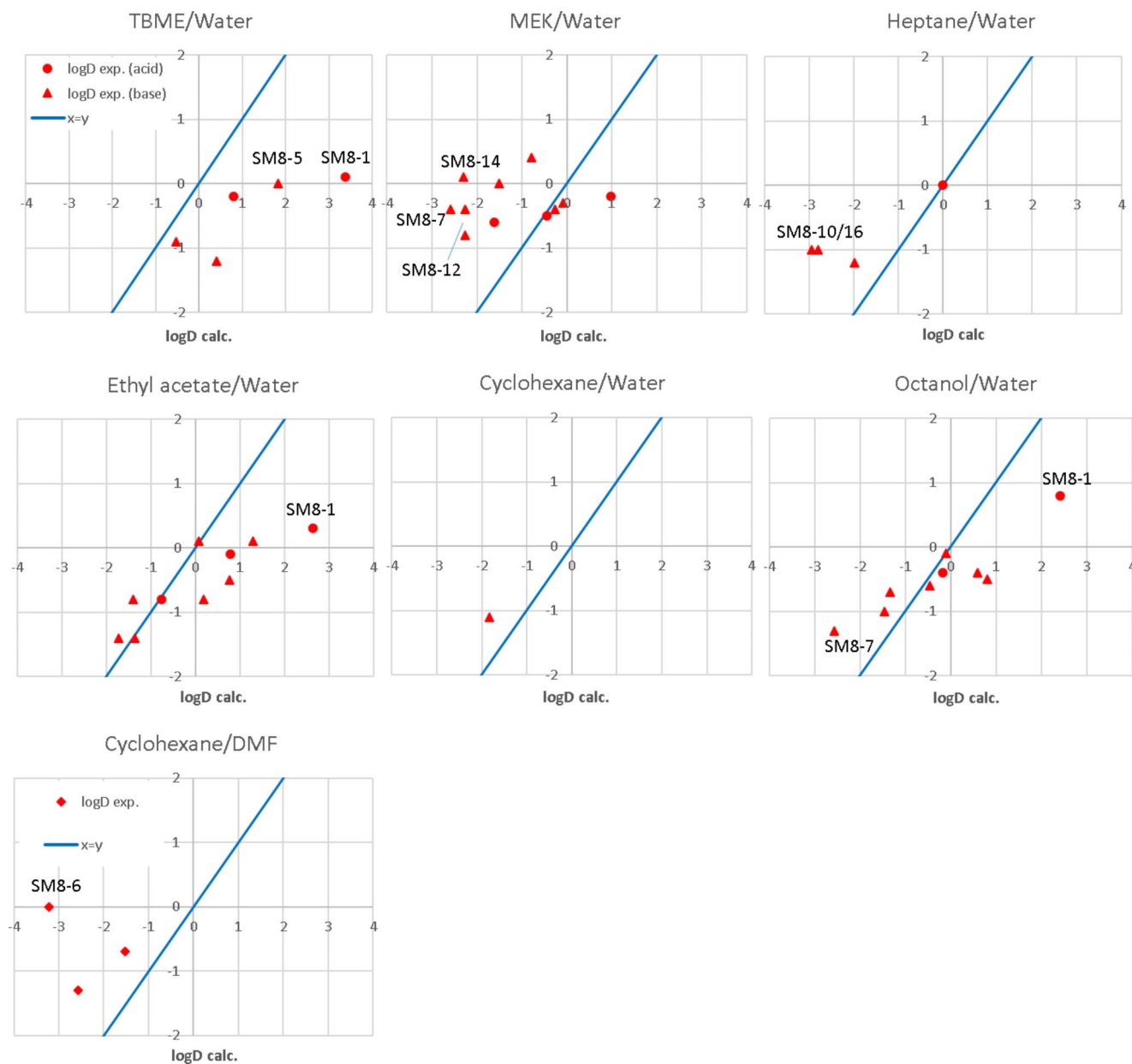
Due to the “first principles based” nature of the COSMO-RS method it is not possible to provide error bars for individual data points. However, the error can be expected to be in the order of 0.5 log units (in the sense of a root mean square deviation) for partition coefficients [14].

**Fig. 3** COSMOtherm pKa predictions of 25 pKa values for 20 molecules. The transitions were derived from the experimental titration curves. The LFER fit for acids was used for the neutral to anion transition (circles) and the base pKa fit for the cation to neutral transition (triangles). The dashed lines mark the corridor of 1 pKa unit deviation



**Table 3** Mean absolute deviation (MAD) and root mean square deviation (RMSD) between COSMO-RS predictions and experiment

	cyclohexane/ DMF	TBME/ water	MEK/ water	heptane/ water	ethylacetate/ water	cyclohexane/ water	octanol/ water
MAD	1.77	1.62	1.20	1.13	0.77	0.73	0.74
RMSD	2.05	1.89	1.42	1.38	1.03	0.73	0.92
Data points	3	5	11	4	10	1	9

**Fig. 4** COSMO-RS logD predictions at 25 °C. Acids are marked by circles and bases by triangles. For the nonaqueous solvent system cyclohexane/DMF the pKa correction was not applied

The solvent system cyclohexane/DMF is the only nonaqueous system in SAMPL8, and therefore does not require a dissociation correction. The structures of the solutes SM8-1,6,16 are chemically diverse and belong to different compound classes. The mutual solubilities of cyclohexane

and DMF were predicted by COSMOtherm (see Table 2). It was found that the logP of the solutes is not very sensitive with respect to the cyclohexane content of the DMF phase. Varying the cyclohexane content of the DMF phase (0–10 mol%), using a pure cyclohexane phase, does not

change the results significantly, whereas the DMF content of the nonpolar cyclohexane phase has a large influence (see Fig. 5). This trend can be explained roughly from the phase and solute polarities. The solutes are all dipolar and protic, with the capability of hydrogen bond acceptance as well as donation. Thus, the solutes strongly prefer the polar (DMF rich) phase, which contains hydrogen bond acceptors that can interact with the solutes. The nonpolar (hexane rich) phase prefers to not contact with the polar solutes simply because of electrostatic repulsion. Thus, all solutes show negative partition coefficients in the pure DMF and hexane solvents. Adding hexane to the DMF phase does not change the picture qualitatively, as the nonpolar hexane does not offer any additional hydrogen bonding contacts for the solutes to interact with. This is different for the hexane phase, where adding DMF also adds a significant amount of hydrogen bonding sites that the solutes can interact with, making the hexane rich phase much more attractive for the solutes, and thus shifting the apparent partition coefficients towards positive values (see Fig. 5). For the actual partition equilibrium, the self-interaction of the solvent also play a role, but even from this simplified picture it becomes clear that the logP prediction significantly depends on the composition of the two phases. Using the COSMOtherm prediction to obtain the composition of the two phases thus provides a reasonable initial guess, but shows a similar error quantitatively, as the logP prediction itself, which may cause a substantial bias on the prediction, explaining the systematic shift seen for this solvent system. Hence, it is recommended to use experimental phase compositions, if available, and COSMOtherm predictions as a fallback if no experimental data is available.

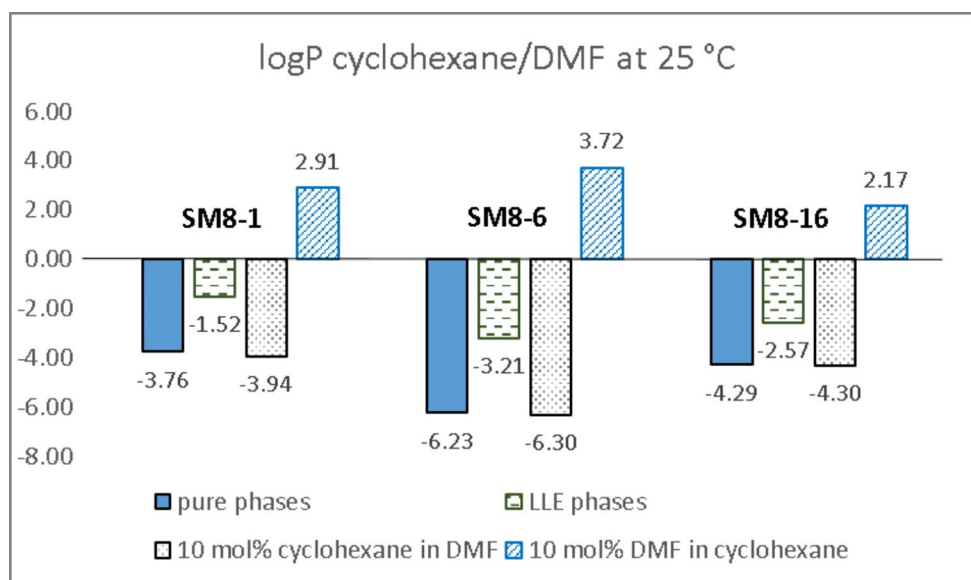
In this study we assumed that protonation and dissociation only take place in the water phase, not in the organic

phase. Hence, the dissociation correction of Eq. 3 was applied to the water phase only. The actual pH values of the measurements, which were not known at the time of the challenge, differ slightly from the values used for the predictions [1]. Nevertheless, this small differences do not change the predictions significantly.

Since the neutral forms of the substances under consideration have low water solubilities, their distribution coefficient logD were determined at a pH value where the molecules were significantly ionized, to ensure that the solubility is large enough to be measurable [1]. The resulting dissociation corrections are listed in Table 1. All molecules have relatively large corrections, which are in the order of magnitude of the calculated logP values. The absolute ratio of the corrections and the logP values is between 0.5 and 3.5. Therefore, the errors of the pKa values play an important role when considering the deviation from the experimental logD values. The dissociation correction of SM8-1, which shows large deviations for four bi-phasic systems, changes from  $-2.99$  to  $-4.01$ , if we use the pKa value of 3.99 measured by Gretz, Czodrowski, Tielker and Kast [30]. Using this correction, the large deviations of SM8-1 for the organic/water systems reduces substantially, except for heptane/water, which is spot on using the SAMPL8 provided pKa values (see Fig. 6).

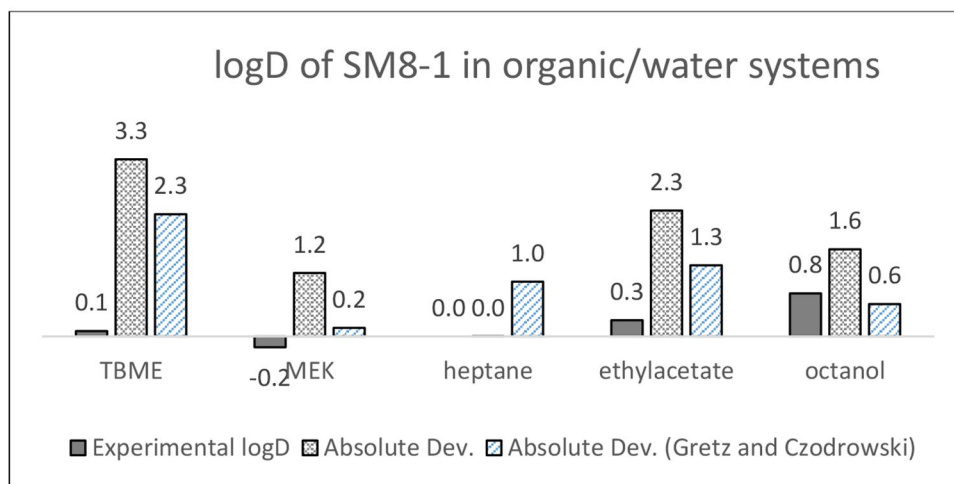
Typically, the solubility of the organic solvent in the water phase is negligibly small, but the solubility of water in the organic phase can be significant. For the octanol/water system the equilibrium solubility of water was 27 mol%, for the ethylacetate/water system it was 20.6 mol%, and for MEK/water 35.1%. For the systems heptane/water, and cyclohexane/water pure solvent phases were assumed, as the mutual solubility of the nonpolar solvents and water is very small.

**Fig. 5** Partition coefficient logP for the cyclohexane/DMF system using different solvent phase compositions. LLE denotes the phases given in Table 2





**Fig. 6** Experimental logD values and absolute deviations to experiment of SM8-1 in the different organic/water systems in log units. The logP corrections were calculated from the SAMPL8 pKa, or the pKa measured by Gretz, Czodrowski, Tielker and Kast [30]



A further source of the unusual high deviation from prediction to the measured data in some solutes could be grounded in the possibility that metastable extraction equilibria could have been measured. As stated in reference 1, due to the high-throughput nature of the experiments not all solvent combinations could be pre-saturated.

## Conclusion

We submitted three COSMO-RS based contributions for the SAMPL8 pKa challenge. Since the task was to predict the standard state free energies for the whole protonation equilibrium, our ranked submission was based on a new combined LFER model that covers the acid and the base pKa and can be used for the prediction of the required values (unified model). The macroscopic pKa values were calculated from the standard state free energies and compared with the experimental data. The assignment of the experimental and calculated values was made using the popular transition method. This COSMO-RS approach yields an RMSD deviation of 3.44 log units and is in the second place of the ranked submissions. A different assignment of experimental and calculated values, which depends on the experimental pH dependent solubilities reduces the RMSD to 1.65. We consider this method to be more consistent, as it is not based on the results of the submitted predictions. Besides the unified model we submitted the results of the COSMOtherm pKa and base pKa models. Together with the experiment-based assignment, these models yield an RMSD of 1.42 log units. This deviation is mainly due to the base pKa of SM8-1. The experimental value of this cation-neutral transition could not be confirmed by the re-measurement of Gretz, Czodrowski, Tielker and Kast and the omission of this outlier leads to an RMSD of 0.89. The results show that the separate LFER fits for pKa and base pKa of COSMOtherm

are clearly advantageous and recommended for the calculations of the macroscopic pKa. However, if the entire (de)protonation equilibrium needs to be described consistently, a unified model is needed.

With an RMSD of 1.36 log units the COSMO-RS logD calculations were the most accurate predictions of the SAMPLE 8 challenge. Nevertheless, some predictions show unexpectedly large deviations from the experiment. The discussion of these deviations is difficult, as they depend on several factors. An important factor is the composition of the phases of the bi-phasic systems. In particular, the fraction of polar substance in the un-polar phase, e.g. the DMF fraction of the cyclohexane phase, has a major influence on the distribution coefficients. To avoid the additional error of predicting the phase composition, it is recommended to use the experimental phase composition whenever possible.

Another important point is the dissociation correction for the aqueous phases, which in some cases becomes dominant at the pH values used in SAMPL8. For a more accurate analysis of the deviations of the logD predictions, the experimental error of the pKa values used would have to be known. The same applies to the achievement of phase equilibrium during the high throughput measurements, which may be another source of error.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10822-023-00514-4>.

**Acknowledgements** We thank the organizers for setting up the SAMPL8 challenge and the National Institutes of Health for its support of the SAMPL project via R01GM124270 to David L. Mobley (UC Irvine). We appreciate Juliana Gretz, Paul Czodrowski, Nicolas Tielker, and Stefan M. Kast of the TU Dortmund for sharing their acidity constants. MD thanks David L. Mobley and Aakanshit Nandkeolyar for fruitful discussions.

**Author contributions** MD and FE wrote the manuscript. All authors reviewed the manuscript and approved the version to be published.

FE and MD analyzed the data and interpreted the results. ST fitted the parameters of the unified pKa LFER model. MD ran the tautomer/conformer generation and performed the COSMO-RS predictions.

**Funding** This research was funded solely by Dassault Systèmes.

**Data Availability** All data generated or analyzed during this study are available in this published article, its supplementary information files, or reference [30].

## Declarations

**Competing interests** The authors declare the following competing financial interests: MD, FE, and ST are employees of Dassault Systèmes, BIOVIA. Dassault Systèmes commercially distributes the software COSMOtherm, COSMOconf, COSMOquick, COSMObase, and TURBOMOLE, which were used in the present study.

**Ethics approval and consent to participate** Not applicable.

**Consent for publication** Not applicable.

## References

- Bahr MN, Nandkeolyar A, Kenna JK et al (2021) Automated high throughput pKa and distribution coefficient measurements of pharmaceutical compounds for the SAMPL8 blind prediction challenge. *J Comput Aided Mol Des* 35:1141–1155. <https://doi.org/10.1007/s10822-021-00427-0>
- Abramov YA (2018) Rational solvent selection for Pharmaceutical Impurity Purge. *Cryst Growth Des* 18:1208–1214. <https://doi.org/10.1021/acs.cgd.7b01748>
- Moss GP, Cronin MTD (2002) Quantitative structure–permeability relationships for percutaneous absorption: re-analysis of steroid data. *Int J Pharm* 238:105–109. [https://doi.org/10.1016/S0378-5173\(02\)00057-1](https://doi.org/10.1016/S0378-5173(02)00057-1)
- Mackay D, Celsie AKD, Powell DE, Parnis JM (2018) Bio-concentration, bioaccumulation, biomagnification and trophic magnification: a modelling perspective. *Environ Sci: Processes Impacts* 20:72–85. <https://doi.org/10.1039/c7em00485k>
- Walker TW, Frelka N, Shen Z, Chew AK, Huber GW (2020) Recycling of multilayer plastic packaging materials by solvent-targeted recovery and precipitation. *Sci Adv* 6:eaba7599. <https://doi.org/10.1126/sciadv.aba7599>
- Sánchez-Rivera KL, Zhou P, Kim MS, González Chávez LD, Grey S, Nelson K, Wang S-C, Hermans I, Zavala VM, Van Lehn RC, Huber GW (2021) Reducing Antisolvent Use in the STRAP process by enabling a temperature-controlled polymer dissolution and precipitation for the recycling of Multilayer Plastic Films. *Chem Sus Chem* 14:4317–4329. <https://doi.org/10.1002/cssc.202101128>
- Mohan M, Keasling JD, Simmons BA, Singh S (2022) In silico COSMO-RS predictive screening of ionic liquids for the dissolution of plastic. *Green Chem* 24:4140–4152. <https://doi.org/10.1039/d1gc03464b>
- Gutiérrez JP, Meindersma GW, de Haan AB (2012) COSMO-RS-Based ionic-liquid selection for extractive distillation processes. *Ind Eng Chem Res* 51:11518–11529. <https://doi.org/10.1021/ie301506n>
- Janoschek L, Grozdev L, Berensmeier S (2018) Membrane-assisted extraction of monoterpenes: from in silico solvent screening towards biotechnological process application. *R Soc Open Sci* 5:172004–172018. <https://doi.org/10.1098/rsos.172004>
- Yara-Varón E, Li Y, Balcells M, Canela-Garayoa R, Fabiano-Tixier AS, Chemat F (2017) Vegetable oils as alternative solvents for green Oleo-Extraction, purification and formulation of Food and Natural Products. *Molecules* 22:1474. <https://europepmc.org/article/med/28872605>
- Klamt A (2016) COSMO-RS for aqueous solvation and interfaces. *Fluid Phase Equilibria* 40:152–158. <https://doi.org/10.1016/j.fluid.2015.05.027>
- Klamt A, Diedenhofen M (2010) Blind prediction test of free energies of hydration with COSMO-RS. *J Comput Aided Mol Des* 24:357–360. <https://doi.org/10.1007/s10822-010-9354-4>
- Klamt A, Eckert F, Reinisch J, Wichmann K (2016) Prediction of cyclohexane-water distribution coefficients with COSMO-RS on the SAMPL5 data set. *J Comput Aided Mol Des* 30:959–967. <https://doi.org/10.1007/s10822-016-9927-y>
- Loschen C, Reinisch J, Klamt A (2020) COSMORS based predictions for the SAMPL6 logP challenge. *J Comput Aided Mol Des* 34:385–392. <https://doi.org/10.1007/s10822-019-00259-z>
- Warnau J, Wichmann K, Reinisch J (2021) COSMO-RS predictions of LogP in the SAMPL7 blind challenge. *J Comput Aided Mol Des* 35:813–818. <https://doi.org/10.1007/s10822-021-00395-5>
- Klamt A (1995) Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. *J Phys Chem* 99:2224–2235. <https://doi.org/10.1021/j100007a062>
- Klamt A, Schüürmann G (1993) COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J Chem Soc Perkin Trans 2* 1993:799–805. <https://doi.org/10.1039/P29930000799>
- Klamt A (2018) The COSMO and COSMO-RS solvation models: COSMO and COSMO-RS. *Wiley Interdiscip Rev Comput Mol Sci* 8:e1338. <https://doi.org/10.1002/wcms.1338>
- Eckert F, Klamt A (2002) Fast solvent screening via quantum chemistry: COSMO-RS approach. *AIChE J* 48:369–385. <https://doi.org/10.1002/aic.690480220>
- BIOVIA COSMOconf 21. Dassault Systèmes, <https://www.3ds.com>, Cologne, Germany
- BIOVIA COSMOquick 21. Dassault Systèmes, <https://www.3ds.com>, Cologne, Germany
- Loschen C, Klamt A (2012) COSMOquick: a novel interface for fast  $\sigma$ -profile composition and its application to COSMO-RS solvent screening using multiple reference solvents. *Ind Eng Chem Res* 51:14303–14308. <https://doi.org/10.1021/ie3023675>
- TURBOMOLE V7.5. University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, TURBOMOLE GmbH, since 2007; available from <http://www.turbomole.com>, Karlsruhe, Germany
- BIOVIA COSMObase 21. Dassault Systèmes. <https://www.3ds.com>, Cologne, Germany
- BIOVIA COSMOtherm 21. Dassault Systèmes. <https://www.3ds.com>, Cologne, Germany
- Klamt A, Eckert F, Diedenhofen M, Beck ME (2003) First Principles calculations of aqueous pKa values for Organic and Inorganic acids using COSMO-RS reveal an inconsistency in the slope of the pKa scale. *J Phys Chem A* 107(44):9380–9386. <https://doi.org/10.1021/jp034688o>
- Eckert F, Klamt A (2006) Accurate prediction of basicity in aqueous solution with COSMO-RS. *J Comput Chem* 27(1):11–19. <https://doi.org/10.1002/jcc.20309>
- Bergazin TD, Tielker N, Zhang Y et al (2021) Evaluation of log P, pKa, and log D predictions from the SAMPL7 blind challenge. *J Comput Aided Mol Des* 35:771–802. <https://doi.org/10.1007/s10822-021-00397-3>

29. Gunner MR, Murakami T, Rustenburg AS, Işık M, Chodera JD (2020) Standard state free energies, not pK<sub>a</sub>s, are ideal for describing small molecule protonation and tautomeric states. *J Comput Aided Mol Des* 34:561–573. <https://doi.org/10.1007/s10822-020-00280-7>
30. Mobley DL, Amezcuca M, Nandkeolyar A, Bergazin TD, Tielker N, Ray D (2023) samplchallenges/SAMPL8: 1.0.0 (1.0.0). Zenodo. <https://doi.org/10.5281/zenodo.7535037>
31. Ingram T, Richter U, Mehling T, Smirnova I (2011) Modelling of pH dependent n-octanol/water partition coefficients of ionizable pharmaceuticals. *Fluid Phase Equilibria* 305:197–203. <https://doi.org/10.1016/j.fluid.2011.04.006>
32. Chen C-S, Lin S-T (2016) Prediction of pH Effect on the octanol – water partition coefficient of Ionizable Pharmaceutical. *Ind Eng Chem Res* 55:9284–9294. <https://doi.org/10.1021/acs.iecr.6b02040>
33. Scott DC, Clymer JW (2002) Estimation of distribution coefficients from the partition coefficient and pK<sub>a</sub>. *Pharm Technol* 26:30–39
34. Dallos A, Liszi JJ (1995) (liquid + liquid) equilibria of (octan-1-ol + water) at temperatures from 288.15 K to 323.15 K. *Chem Thermodyn* 27:447–448. <https://doi.org/10.1006/jcht.1995.0046>
35. Lladosa E, Montón JB, de la Torre J, Martínez NF (2011) Liquid – liquid and vapor – liquid – liquid equilibrium of the 2-Butanone + 2-Butanol + water system. *J Chem Eng Data* 56:1755–1761. <https://doi.org/10.1021/je1004643>
36. Ashour I (2005) Liquid – liquid equilibrium of MTBE + ethanol + water and MTBE + 1-Hexanol + water over the temperature range of 288.15 to 308.15 K. *J Chem Eng Data* 50:113–118. <https://doi.org/10.1021/je049799a>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.