

An open source COSMO-RS implementation and parameterization supporting the efficient implementation of multiple segment descriptors

Thomas Gerlach^{a,b,*}, Simon Müller^a, Andrés González de Castilla^a, Irina Smirnova^a

^a*Institute of Thermal Separation Processes, Hamburg University of Technology, Hamburg*

^b*Bayer AG, Leverkusen*

Abstract

The model COSMO-RS (Conductor-like screening model for realistic solvation) is based on the approximation of molecular interactions by the interactions of surface segments from the molecular cavities. These segment interactions are functions of the segment properties i.e. descriptors, with the most important descriptor being the screening charge density. Whereas the inclusion of further descriptors can lead to improved accuracy for the models predictions, it also increases the computational demand due to the growth of the central nonlinear equation system of COSMO-RS. In this work, we present an algorithm that efficiently allows the introduction of further segment descriptors at moderate size increase of the central equation system. It is included into our implementation of COSMO-RS, which is provided as open source code and will be named openCOSMO-RS in the future. Different parameterizations of the model are performed, implementing an RDKit/ORCA based workflow for the required quantum chemical calculations. In this way, the implementation is open for evaluation and refinement for a large range of academic users.

Keywords: COSMO-RS, open-source, parameterization, ORCA

1. Introduction

COSMO-RS is a model with numerous applications within different industries as well as in academia. Its predictive power for fluid phase thermodynamics allows its use in a wide range of applications without requiring experimental data. The model was first devised and implemented by Klamt et al. [1, 2, 3]. It is based on the combination of information from quantum chemical calculations with a fast model for molecular interactions. The molecular interactions are accounted for by introducing a simplified picture of interacting surface segments, with the interaction strengths being described as general functions of the

*Corresponding author: thomas.gerlach1@bayer.com

properties of these segments. The most important property, or descriptor, was proposed to be the screening charge density σ at the molecular surface, which is derived from a quantum chemical calculation of the molecule being embedded in an ideal conductor. As proposed by Klamt et al. [2], other properties, like the chemical element of the atom to which a surface segment belongs, can additionally be used to refine the model and its parameterization.

The full model equations as well as a basic parameterization were already published in 1998, allowing a full reimplementations of the model [2]. A parameterization based on DFT quantum chemical calculations with the BP86 functional and a TZVP basis set was published by Eckert and Klamt [3]. In the following years, Klamt et al. pursued all further refinements of their implementation within the proprietary software *COSMOtherm*, which is now widely applied within academic as well as industrial groups [4]. Several different reimplementations of the model were developed by other groups and are in wide use today.

The reimplementations COSMO-SAC by Lin and Sandler [5] is based on the fundamental COSMO-RS theory albeit introducing several model modifications. Their first implementation was based on the single segment descriptor of the screening charge density. Several refinements of the model were introduced after the initial publication. Hsieh et al. [6] introduced a separation of σ -profiles for hydroxy groups, other hydrogen bonding groups, and non-hydrogen bonding groups. Later, a new contribution was introduced, accounting for dispersive interactions between molecules as an additional term apart from the combinatorial and residual contribution to the activity coefficient [7]. Recently, a reference implementation of the COSMO-SAC model variants was published as an open source code by Bell et al. [8].

Other reimplementations of COSMO-RS include COSMO-RS(OL) by Grensemann and Gmehling [9], the ADF implementation by Pye et al. [10], an implementation by Banerjee et al. [11], JCOSMO as published by Gerber et al. [12], and an implementation developed by Franke and Hannebauer [13]. Of these, only the source code of JCOSMO was published as an open source project.

In all of the COSMO-RS implementations, the main segment descriptor defining the interactions between segments is the screening charge density σ . Most reimplementations of the model start from a precalculated σ -profile that is based on the creation of a histogram of the molecular cavity surface area over the screening charge density using predefined bins. A notable extension includes the separation into three distinct σ -profiles in newer variants of COSMO-SAC [6]. The implementations by Franke and Hannebauer as well as the ADF implementation support the additional segment descriptor $\sigma^\perp$, which was proposed by Klamt et al. [2] to improve the description of electrostatic interactions.

In general, the addition of further segment descriptors can be expected to significantly increase the size of the non-linear equation system that constitutes COSMOSPACE [14], the central part of the residual term of COSMO-RS. An efficient algorithm for the inclusion of multiple segment descriptors can be beneficial by allowing further refinements of the model, while keeping the equation

system at a reasonable size.

To the best of our knowledge, no open-source implementation of COSMO-RS includes such a flexible extension to multiple segment descriptors yet. We developed a solution that we integrated into our implementation of the model named COSMO-RS(HH) in the past [15, 16]. The focus was the efficient implementation of a full COSMO-RS model including additional segment descriptors like $\sigma^\perp$, the chemical element of each segment, or the total charge of the molecule as the basis for the development of a model for electrolyte systems. Nevertheless, the basic algorithm and its implementation may be beneficial in other domains. Therefore, we now make the software available as an open source code and renamed it to 'openCOSMO-RS'.

To present our implementation, we first outline the fundamentals of the COSMO-RS model. Subsequently, we introduce the method for the efficient introduction of multiple segment descriptors into the model implementation, focusing on a specific additional descriptor, $\sigma^\perp$. The analysis also demonstrates the capability of openCOSMO-RS to determine additional properties from contact statistics, including the partial molar energy contributions. In the following, we present different parameterizations of the model based on data from infinite dilution activity coefficients. The parameterizations are performed using the quantum chemical program package ORCA [17, 18], which is free to use for non-commercial users and is available for Linux and Windows.

2. The model COSMO-RS

In the following the central equations of COSMO-RS [2, 3, 19, 14] are summarized and their implementation in the software is discussed. COSMO-RS is based on the assumption that the interactions between molecules in liquid phases can be represented by an ensemble of interacting molecular surface segments. The most important properties of these surface segments are derived from quantum chemical calculations that are performed in the conductor state, i.e. a state, in which the molecule is embedded in its cavity in a perfect conductor. The idea is that the interaction free energies between segments can be described as functions of a set of segment descriptors with the main descriptor being the screening charge density σ (cf. figure 1). Typically, the value of the screening charge density from the COSMO calculations is not used directly, but an averaging is performed first to reduce numerical noise [2]:

$$\sigma_\alpha = \frac{\sum_\beta \sigma_\beta^* \frac{r_\beta^2 r_{av}^2}{r_\beta^2 + r_{av}^2} \exp(\frac{d_{\alpha\beta}^2}{r_\beta^2 + r_{av}^2})}{\sum_\beta \frac{r_\beta^2 r_{av}^2}{r_\beta^2 + r_{av}^2} \exp(\frac{d_{\alpha\beta}^2}{r_\beta^2 + r_{av}^2})} \quad (1)$$

The sums include all surface segments β from the quantum chemical cavity surface, σ_β^* and r_β are the non-averaged screening charge density and the radius of surface segment β , and $d_{\alpha\beta}$ is the distance between surface segments α and β .

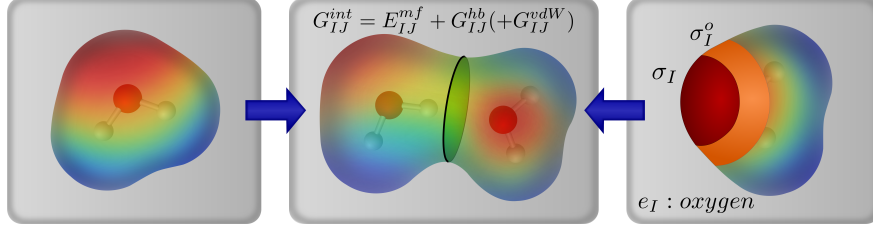


Figure 1: Schematic representation of the interaction energy concept within COSMO-RS. Molecular interactions are replaced by segment contacts. The interaction free energies are functions of a set of segment descriptors.

The surface segments are subsequently distributed onto a predefined grid of screening charge density values to create the σ -profiles. These clustered segment types later enter the COSMO-RS calculations. Assuming a fixed area of a segment contact, the fraction of a segment type I in a segment mixture can be expressed as:

$$X_I = \frac{\sum_i x_i A_I^i}{\sum_i \sum_j x_i A_j^i} \quad (2)$$

Here, A_I^i is the surface area of segment type I on a molecule of type i , and x_i is the mole fraction of molecule i .

The central part of COSMO-RS is the non-linear equation system COSMO-SPACE. It relates the interaction free energies of the segments to their activity coefficients Γ_I in the mixture of contacting segments [14]:

$$\Gamma_I = \frac{1}{\sum_J X_J \Gamma_J \tau_{IJ}} \quad (3)$$

If the COSMO reference state is retained for the segments, the interaction parameter τ_{IJ} can be defined as:

$$\tau_{IJ} = \exp\left(-\frac{G_{IJ}^{int}}{RT}\right) \quad (4)$$

The interaction free energy G_{IJ}^{int} of a segment contact is often reported as an interaction energy, but must be considered as an interaction free energy in practice, unless the individual contributions to this term are temperature independent. The interaction free energy is specified as a function of the set of segment type descriptors.

Apart from the descriptor of the screening charge density itself, a descriptor describing the polarity of the segment surrounding ($\sigma^\perp$) is generated. For this purpose, an additional averaging using equation 1 with twice the averaging radius r_{av} is performed, resulting in a charge density σ^o . In many cases a strong correlation is present between both averaged screening charge densities. During the derivation of the COSMO-RS model for electrostatic interactions

between segments it is commonly assumed that neighbouring segment contacts are of negligible influence for the interaction energy of a segment contact. If this assumption is not true, it can be expected that the electrostatic potential from the molecular surroundings at the segment contact site does not vanish and an effect on the interaction energy of the contact site should be accounted for. Klamt et al. [2] integrated such a correction for the screening charge correlation by first creating an orthogonal descriptor from the analysis of the σ and σ^o distributions of a set of molecules and using this descriptor as additional information in the electrostatic interaction energy relation. The new descriptor is calculated from:

$$\sigma_{\alpha}^{\perp} = \sigma_{\alpha} - 0.816\sigma_{\alpha}^o \quad (5)$$

The contribution of electrostatic misfit interactions to the interaction free energy can then be calculated as:

$$E_{IJ}^{mf} = 0.5 a_{eff} \alpha [(\sigma_I + \sigma_J)^2 + f_{corr}(\sigma_I + \sigma_J)(\sigma_I^{\perp} + \sigma_J^{\perp})] \quad (6)$$

where α is the misfit prefactor, a_{eff} is the area of a segment contact, and f_{corr} is a parameter that was adjusted based on the dielectric screening energies of a set of diverse molecules [2].

For the calculation of the hydrogen bonding free energy contribution, additional segment descriptors are applied. Whereas in newer variants of COSMO-SAC the σ -profile is separated into a hydroxy, a hydrogen bonding non-hydroxy and a non-hydrogen bonding part, here we introduce the additional descriptor of the segment class [6]. With this descriptor, we distinguish segments from hydrogen atoms with negative screening charges, segments from potentially hydrogen bond accepting atoms (at present C, N, O, F, P, S, Cl, Br, I) with positive screening charges, and other segments. Additionally, we use the segment descriptor of the chemical element e_I to open the opportunity to perform an element specific scaling for the hydrogen bond strength [20, 21]. For hydrogen atoms, this descriptor additionally encodes information about the chemical element of the bound atom. Nevertheless, in the parameterizations presented in this work, the hydrogen bond scaling parameters are set to unity, i.e. the original expression by Klamt et al. [2] is applied. The hydrogen bond interaction energy can then be calculated from:

$$G_{IJ}^{hb} = \begin{cases} c_{hb}(T)a_{eff}[\min(0; \sigma_I + \sigma_{hb}) \cdot \max(0; \sigma_J - \sigma_{hb})] & \text{if } \sigma_I < \sigma_J \\ c_{hb}(T)a_{eff}[\min(0; \sigma_J + \sigma_{hb}) \cdot \max(0; \sigma_I - \sigma_{hb})] & \text{if } \sigma_I \geq \sigma_J \end{cases} \quad (7)$$

σ_{hb} is the hydrogen bond threshold parameter. Although the temperature dependence of hydrogen bonds is not of relevance in this work, the following equation for the temperature dependence of the hydrogen bond interaction free energy is implemented in the software [19]:

$$c_{hb}(T) = c_{hb} \max[0, 1 - c_{hb}^T + c_{hb}^T \cdot 298.15/T] \quad (8)$$

Here, c_{hb} is the hydrogen bond strength parameter, and c_{hb}^T is a parameter for the temperature dependence of the hydrogen bond.

An additional contribution for the introduction of dispersion interactions can be introduced into the model. In COSMO-RS, this is implemented by a set of element specific energy parameters. In its simplest form, the interaction energies are not dependent on the contacts but additive for the elements ($E_{IJ}^{disp} = f(e_I) + f(e_J)$) [19]. They do not change the relative interaction energies of different segment contacts when the reference state is also in the liquid phase and therefore do not influence activity coefficients. As we focus on the liquid phase predictions by COSMO-RS and do not consider gas phase predictions in this work, these factors are neglected. If instead, the parameters are chosen to be dependent on the different element pairs ($E_{IJ}^{disp} = f(e_I, e_J)$) as done in newer versions of COSMOtherm [22], they become relevant also for activity coefficients. Nevertheless, a regression of a full set of such parameters was not in the scope of this work. With regard to the required segment descriptors, only the chemical element e_I of each segment type would be required, which is readily available in the software. Restricting the interaction free energies to the hydrogen bonding and misfit contribution, the complete interaction free energy is calculated from:

$$G_{IJ}^{int} = E_{IJ}^{mf} + G_{IJ}^{hb} \quad (9)$$

When the COSMOSPACE equation system has been solved, the activity coefficient at composition $\mathbf{x}$ with respect to the COSMO reference state can be calculated from:

$$\ln(\gamma_i^\#(\mathbf{x})) = \sum_I \left(\frac{A_I^i}{a_{eff}} \ln(\Gamma_I) \right) \quad (10)$$

A conversion to a different molecular reference state, e.g. the pure component, leads to the residual contribution to the molecular activity coefficient:

$$\ln(\gamma_i^{res}(\mathbf{x})) = \ln(\gamma_i^\#(\mathbf{x})) - \ln(\gamma_i^\#(\mathbf{x}^{ref})) \quad (11)$$

For the combinatorial term, different variants are implemented in openCOSMO-RS, with the Staverman-Guggenheim term being used in this work [23]:

$$\frac{\phi}{x_i} = \frac{V_i}{\sum_j x_j V_j} \quad (12)$$

$$\frac{\theta}{x_i} = \frac{A_i}{\sum_j x_j A_j} \quad (13)$$

$$\ln(\gamma^{comb}(\mathbf{x})) = \ln\left(\frac{\phi}{x_i}\right) + 1 - \frac{\phi}{x_i} - 0.5z \frac{A_i}{A_{std}} \left(\ln\left(\frac{\phi}{\theta}\right) + 1 - \left(\frac{\phi}{\theta}\right) \right) \quad (14)$$

Here, z is the coordination number and is typically chosen to be equal to 10. The total activity coefficient is then calculated from:

$$\ln(\gamma_i(\mathbf{x})) = \ln(\gamma_i^{res}(\mathbf{x})) + \ln(\gamma_i^{comb}(\mathbf{x})) \quad (15)$$

In addition to the calculation of the activity coefficient, COSMO-RS allows the calculation of the contact probabilities between segments [14]. The expected value of the number of contacts between segments of type $(I - J)$ divided by the number of segments N in a mixture is calculated from [24]:

$$P_{IJ} = \frac{E[N_{IJ}]}{N} = X_I X_J \Gamma_I \Gamma_J \tau_{IJ} \quad (16)$$

This can be used to derive several useful quantities that can aid in the analysis of the model results. In openCOSMO-RS, at present the calculation of the contributions of the different interaction energy terms to the partial molar enthalpy, as well as the calculation of the average interaction energies of molecules in mixtures are supported. The derivation of the equations was reported by Gerlach et al. [15]. The average interaction energies of the molecules in a mixture is calculated from:

$$\epsilon_i^\# = 0.5 \sum_I \sum_J N_I^i \frac{P_{IJ}}{X_I} E_{IJ} \quad (17)$$

For the contribution of the different interaction energy terms to the partial molar enthalpy, the following relation can be used:

$$\bar{h}_i^\# = 0.5 \sum_I \sum_J \left(N^i P_{IJ} + \left(\sum_k x_k N^k \right) \left[n \left(\frac{\partial P_{IJ}}{\partial n_i} \right)_{T, n_j} \right] \right) E_{IJ} \quad (18)$$

Here, N^i being the number of segments on a molecule i , which is calculated from the ratio of the cavity surface area and the contact area a_{eff} . Equation 17 and 18 can either be applied to the total interaction energy or its individual contributions due to misfit interactions or hydrogen bonding. For a correct calculation of the partial molar enthalpy from eq. 18, the interaction energy contribution of interest E_{IJ} has to be calculated first from the respective interaction free energy, e.g. by applying the Gibbs-Helmholtz relation. The partial derivative of the contact probability multiplied with the total number of molecules n can be calculated analytically, requiring the solution of a linear equation system. The required equations were reported previously [15].

3. Computational methods

The choice of conformational isomers to represent each molecule can influence the predictions by COSMO-RS. Although the required modifications of the model to represent a molecule by a set of conformers have been described by Klamt [1] with the first description of COSMO-RS, in this work we limit all considerations to the representation of molecules by a single conformer. For the conformer search and quantum chemical calculations, two approaches were

applied within this work. On the one hand, a conformer search was applied using *COSMOconf* 3.0, integrating TURBOMOLE 6.6 based quantum chemical calculations [25, 26]. Geometry optimizations were performed in the conductor state using DFT calculations with the BP86 functional and a def-TZVP basis set. For further applications in *COSMOtherm*, additionally single point calculations were performed with the def2-TZVPD basis set and the FINE cavity construction algorithm [27].

Additionally, an alternative workflow was applied using software available for free to non-commercial users, based on the following steps:

- Molecular mechanics based conformer generation using RDKit [28, 29]
- ALPB [30] geometry optimizations in water using GFN2-xTB [31, 32] calculations from within ORCA [17, 18, 33]
- Filter conformers by an energy window of 6 kcal/mol
- Cluster conformers by an RMSD window of 1 Å, taking the conformer with the lowest energy
- COSMO geometry optimizations at DFT/BP86/def2-TZVP(-f) level in ORCA
- Selection of conformer with lowest energy
- COSMO geometry optimizations at DFT/BP86/def2-TZVP level in ORCA
- COSMO single point calculation at DFT/BP86/def2-TZVPD level in ORCA

Whereas RDKit is available as open source software, ORCA is available free of charge for academic and private use. The python script that was developed to control the calculations from an initial SMILES code is also made available with the source code of openCOSMO-RS.

The subsequent COSMO-RS calculations were performed with the source code that is made available in our git repository. Two variants of the openCOSMO-RS code are made available. The openCOSMO-RS(py) implementation is created for conceptual simplicity and rapid development, while relying purely on the python library NumPy [34] for speed improvements. The openCOSMO-RS(c++) implementation is a faster, fully C++ based implementation making use of the Eigen library [35] for linear algebra and developed with computational speed in mind. Still, it supports full access on python via the pybind11 library [36] and on MATLAB [37] as a MEX function. Examples are provided within the repository to facilitate the usage. For comparative purposes, for some calculation the proprietary COSMO-RS implementation *COSMOtherm* [4] was employed.

For all optimizations, the following objective function was minimized:

$$OF = \frac{1}{N_p} \sum_i \left(\ln(\gamma_i^{\infty,calc}) - \ln(\gamma_i^{\infty,exp}) \right)^2 \quad (19)$$

Here, $\gamma_i^{\infty,calc}$ and $\gamma_i^{\infty,exp}$ are the calculated and experimental infinite dilution activity coefficient of data point i , and N_p is the number of data points. For the optimization the Nelder-Mead algorithm was chosen. Satisfactory experience with regard to the identification of suitable optima for this type of problem is expected and the calculation of numerical derivatives is avoided using this algorithm.

The root mean square deviation was calculated from:

$$RMSD_{IDAC} = \sqrt{\frac{1}{N_p} \sum_i \left(\ln(\gamma_i^{\infty,calc}) - \ln(\gamma_i^{\infty,exp}) \right)^2} \quad (20)$$

For the partition ratios, the reported experimental partition coefficients were first converted to a mole fraction basis, accounting for the phase equilibrium composition of the solvents. The root mean square deviation was then calculated from:

$$RMSD_{\ln(K)} = \sqrt{\frac{1}{N_p} \sum_i \left(\ln \left(\frac{\gamma_i^{\infty,calc,aq}}{\gamma_i^{\infty,calc,org}} \right) - \ln(K_i^{\infty,exp}) \right)^2} \quad (21)$$

Here, $\ln(K_i^{\infty,exp})$ is the experimental mole fraction based partition ratio, $\gamma_i^{\infty,calc,aq}$ is the calculated infinite dilution activity coefficient in the aqueous phase and $\gamma_i^{\infty,calc,org}$ the respective property in the organic phase.

4. Results and discussion

4.1. Extended σ -profile generation

In the following, the algorithm for the creation of the required extended σ -profiles is outlined. "Extended" refers here to the fact, that multiple segment descriptors are included in addition to the screening charge density. As in other implementations of COSMO-RS, properties from the space of real numbers, like σ and $\sigma^\perp$, are first distributed onto a predefined grid [1, 5]. In this work, a bin size of $0.001 \text{ e}/\text{\AA}^2$ was used for both descriptors [1].

In figure 2 the resulting extended σ -profiles of two active pharmaceutical ingredients and two solvents are shown. In particular the molecules indometacin and chloramphenicol consist of several different chemical elements while covering a large range of values of the descriptors σ and $\sigma^\perp$. Although the segment descriptor of the chemical element is not used in the COSMO-RS parameterizations in this work, it is nevertheless of promising utility for advanced parameterizations, e.g. by introducing element specific hydrogen bonding strengths, or binary van der Waals interaction parameters [20, 21, 22].

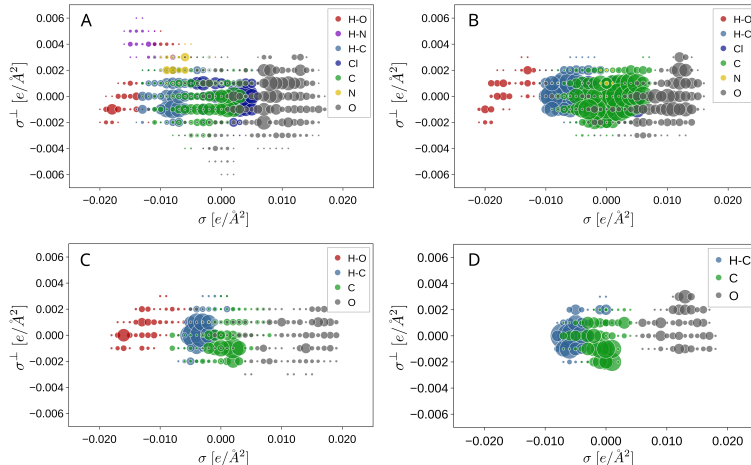


Figure 2: Extended σ -profiles of chloramphenicol (A), indometacin (B), ethanol (C), and acetone (D). The segment descriptors of σ_I , $\sigma_I^\perp$, and the chemical element e_I are shown. Circle sizes indicate the area of the respective segment type on the molecular surface.

Table 1: Number of segment types of pure components and their mixture in openCOSMO-RS.

Name	Number of segment types
Chloramphenicol	553
Indometacin	358
Acetone	146
Ethanol	272
Mixture	639

If all possible combinations of the relevant descriptors covering the respective ranges were included in the calculations, a very large number of segment types would result, leading to the need to solve a very large non-linear COSMOSPACE equation system. Nevertheless, as expected, it can be observed that a correlation between these segment descriptors exists, leading e.g. to higher screening charges for oxygen atoms in comparison to carbon atoms. Additionally, it has to be noted that for the COSMO-RS calculations only those segment types are of relevance, which are associated with molecules being considered in the mixture. This encompasses molecules at zero and non-zero concentration. Based on this fact, a more efficient alternative to the use of predefined σ -profiles becomes obvious.

Instead of covering the complete space of possible descriptors, a set of only the occupied segment types is created and used for the subsequent creation of interaction energy arrays and the solution of the COSMOSPACE equation set. By still relying on fixed values for the descriptors the extended σ -profiles can be independently stored and combined during the mixture

creation if higher computational speeds are of relevance, as it is implemented in openCOSMO-RS(c++). In figure 3 an illustration of the segment type creation procedure is shown. Whereas a related algorithm is applied in newer implementations of COSMO-SAC to reduce the number of entries during the COSMOSPACE calculation step below the fixed number of $3*51$ from the three separate σ -profiles [8], the method that is described here goes further by reducing the number of segment types directly during the processing of the COSMO information, and extending the approach flexibly to multiple segment descriptors. The same principle of using a set of only the occupied segment types is applicable to the σ -profiles of mixtures. In table 1 the number of segment types of two complex, polyatomic molecules, two common solvents, as well as their mixture is shown for the case of three descriptors, applying this approach.

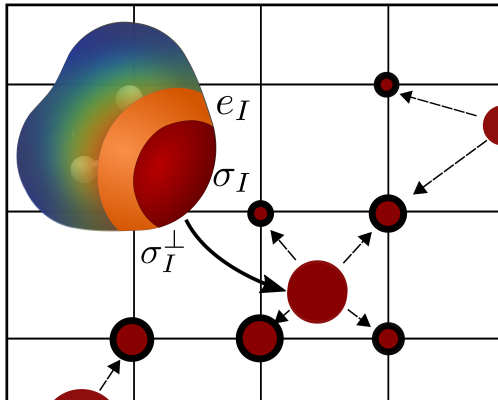


Figure 3: Procedure of the segment type generation. Segments from the cavity surface are distributed onto neighbouring grid points on a predefined grid in the space spanned by the segment descriptors. Only occupied grid points are stored and are subsequently used for the creation of the interaction energy arrays and the COSMOSPACE calculations.

4.2. Parameterizations

Using the new implementation of COSMO-RS, parameterizations of the model were performed. For this purpose, the collection of infinite dilution activity coefficients by Voutsas and Tassios [38] and He and Zhong [39] were used. All data points with either water or organic acids as a solute were removed from the data set before starting the regressions. It has already been reported that COSMO-RS can show significant deviations for the solute water and for organic acids as solutes, the latter due to dimerization [2, 40]. To correct the outliers for water as a solute, in different implementations of COSMO-RS corrections were introduced [41, 9]. As the introduction of corrections like σ -profile scalings

should be accompanied by a parameterization at finite concentrations, we did not attempt such a correction in this work.

Parameterizations were performed using RDKit for the initial conformer generation and ORCA for the quantum chemical calculations. Due to its much higher computational efficiency, the openCOSMO-RS(c++) implementation was used for the parameterizations but calculations were validated with openCOSMO-RS(py). Main differences included in the c++ variant are the precalculation of extended σ -profiles, the efficient vectorization of floating point number calculations, the multi-threaded calculation of systems and the caching of intermediate values to accelerate the COSMOSPACE convergence.

Table 2: Parameterizations of COSMO-RS

	TURBOMOLE	ORCA-A	ORCA-B
Method	TZVP	TZVPD	TZVPD
$\sigma^\perp$ used	yes	no	yes
r_{av} [$\text{\AA}$]	0.5 ^a	0.5 ^a	0.5 ^a
a_{eff} [$\text{\AA}^2$]	6.25 ^b	5.943	6.115
α_{mf} [$kJ \text{\AA}^2/(mol e^2)$]	5.950e6 ^b	5.953e6	7.584e6
f_{corr} [-]	2.4 ^a		2.4 ^a
c_{hb} [$kJ \text{\AA}^2/(mol e^2)$]	3.670e7 ^b	3.671e7	3.093e7
σ_{hb} [$e/\text{\AA}^2$]	8.5e-3 ^b	7.469e-3	7.876e-3
c_{hb}^T [-]	1.5 ^c	1.5 ^c	1.5 ^c
A_{std} [$\text{\AA}^2$]	79.53 ^d	51.28	41.89

^a Klamt et al., 1998 [2]; ^b Eckert and Klamt, 2002 [3]; ^c Klamt and Eckert, 2000 [19]. ^d Lin and Sandler, 2002 [5];

Table 3: Root mean square deviations for the training and the test set.

	TURBOMOLE	ORCA-A	ORCA-B
Training set ^a	0.76	0.78	0.65
Test set ^b	0.96	1.27	1.06

^a Infinite dilution activity coefficient collection by He and Zhong [39] and Voutsas and Tassios [38]; ^b Partition ratio collection by Klamt et al. [2].

The model parameterizations were performed with (ORCA-B) and without (ORCA-A) the screening charge correlation correction shown in eq. 6. The evaluation of this parameter may serve as an example for the application of an interaction energy correction that relies on an additional, continuously distributed segment descriptor, in this case $\sigma^\perp$. In each of these parameterizations, five parameters were simultaneously adjusted to the data set. Other parameters that were determined independently are listed together with the regressed values in table 2. Additionally, a full set of parameters from literature references is listed for the use with σ -profiles calculated from TURBOMOLE.

In table 3, the root mean square deviations for the training set of logarithmic infinite dilution activity coefficients is shown. Additionally, a set of parti-

tion ratios in the octanol+water, benzene+water, hexane+water, and diethyl ether+water systems as collected by Klamt et al. [2] was used for the validation of the predictions. Only data points with N_2 as the solute were removed from the data set. In figures 4 and 5, results for the training set and the test set are shown for selected parameterizations.

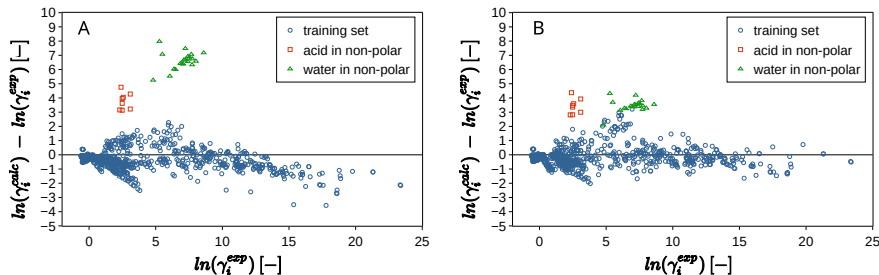


Figure 4: Deviations between the training set of infinite dilution activity coefficients and the results of the TURBOMOLE based calculations implemented in openCOSMO-RS (A) as well as the parameterization ORCA-B (B). The results for the excluded data points of acids and water as solutes in non-polar solvents are also shown.

As can be seen, a higher accuracy is achieved for the regression as well as the predictions for the test set if the screening charge correlation correction is applied. This is the case although the same number of parameters was adjusted to the training set of experimental data. The parameters of the screening charge correlation correction were in contrast only regressed to results from quantum chemical calculations, and were taken from the literature in this work. The application of the RDKit/ORCA based workflow led to lower deviations in comparison to the use of the published parameters for TURBOMOLE based calculations for the training set but higher deviations for the test set. Differences between the approaches include the applied conformer search algorithm as well as differences in the algorithms used for the quantum chemical calculations, including e.g. the cavity construction algorithm as well as the treatment of outlying charges. Nevertheless, the integration of this simple RDKit/ORCA based workflow widens the applicability of openCOSMO-RS by making it accessible to a broader range of academic users. In figure 6 a comparison of the σ -profiles of ethanol using different methods for the COSMO calculations is shown. A high similarity between the different approaches is observed.

Additionally, a comparison with results from the proprietary software package COSMOtherm was performed for the test set. A higher accuracy is observed, with particularly lower deviations being reached by the BP_TZVPD_FINE_19 parameterization of COSMO-RS. The calculated RMSD is $RMSD = 0.72$ using this level, and $RMSD = 0.87$ using the BP_TZVP_19 level. Whereas even the BP_TZVP_19 version has an advanced parameterization in comparison to openCOSMO-RS, it also contains several model differences including conformation equilibria and binary van der Waals interaction parameters. In particular

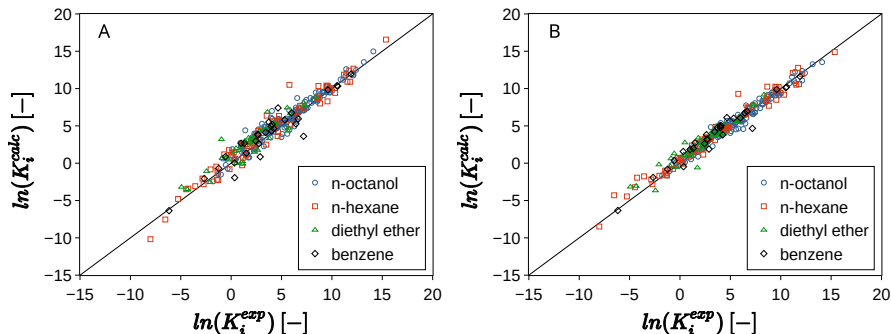


Figure 5: Parity plot for the prediction results of the parameterization ORCA-B (A) as well as COSMOtherm with the BP_TZVPD_FINE_19 parameterization (B) based on the test set containing n-octanol+water, n-hexane+water, diethyl ether+water, and benzene+water partition coefficients.

for the BP_TZVPD_FINE_19 level of the model, refinements of key parts of the model itself were introduced, that have not been described in the literature yet [41]. A comparison of openCOSMO-RS with predictions from a recent version of COSMOtherm was reported for the training set in our previous work [15].

4.3. Property analysis

openCOSMO-RS includes a functionality for the evaluation of contact probabilities and derived properties, which can aid in the understanding of model results [42, 43, 44, 15]. In the following, this functionality is applied to the system of acetone+water, to demonstrate potential applications for a selected system. For this purpose, the effects of the screening charge correlation correction are investigated.

In figure 7 the logarithmic activity coefficient as well as the contribution to the partial molar energies in the binary acetone+water system are shown for the component acetone. Applying parameterization ORCA-B, the results for the activated as well as deactivated screening charge correlation correction are compared. The infinite dilution activity coefficient of acetone in water decreases with the inclusion of this term at otherwise fixed parameters. This is in particular connected to a decrease of the contribution of the misfit term to the partial molar enthalpies. As expected, for the hydrogen bonding term only a small change occurs, which is due to a change in the contact probabilities between segments within the model.

In figure 8 the average interaction energies of the acetone molecule in the same system are shown. Here, the chosen reference state is the COSMO reference state. Only minor differences are also observed for the average hydrogen bonding interaction energies. In contrast to this, in pure acetone a higher average misfit interaction energy is observed if the screening charge correlation correction is applied. This can be explained by the strong correlation of the

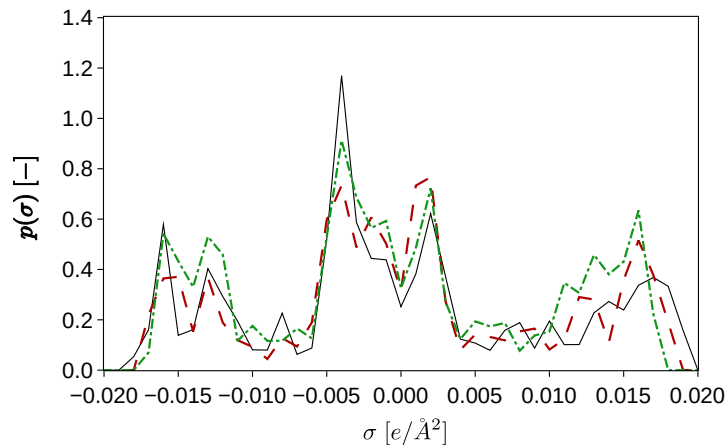


Figure 6: σ -profile of ethanol using TURBOMOLE with the BP functional and a def-TZVP basis set (solid line), TURBOMOLE with a def2-TZVPD basis set and the FINE cavity construction algorithm (dashed line), and ORCA with a def2-TZVPD basis set (dash-dotted line).

screening charges on the surface of the oxygen atom of the acetone molecule. In pure acetone, no adequate screening partners of opposite charge are found for these surface segments (cf. figure 2), leading to high misfit charges with the interaction energies being increased by the correlation. In contrast to this, at infinite dilution of acetone in water, the average interaction energies are reduced. In this environment, more adequate screening partners are present for the oxygen surface segments and the hydrogen bonding contribution strongly influences the segment pairing.

It becomes clear that the analysis of the properties that are based on the contact probabilities is a powerful tool to help in the understanding of the prediction results even for complex systems and allows deeper insights that may lead to further model improvements.

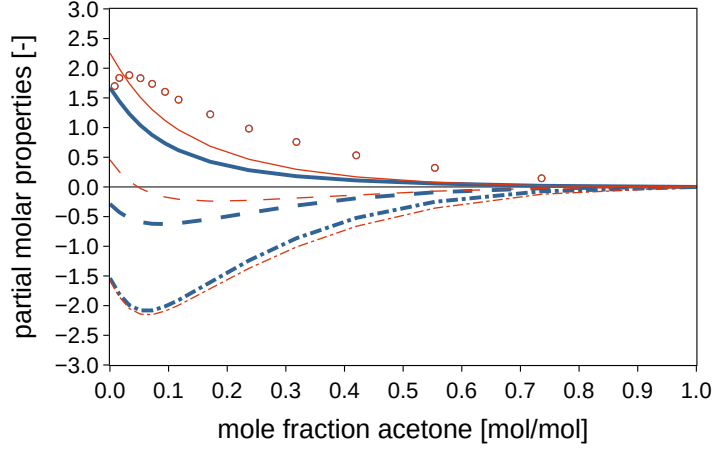


Figure 7: Partial molar properties of acetone in the binary acetone+water system at vapor-liquid equilibrium conditions and 1 atm. Logarithmic activity coefficients $\ln(\gamma)$ (solid lines: calculated, markers: experimental data [45]), partial molar energy contribution $\bar{h}_i^\#/(RT)$ due to the misfit energy term (dashed line) and the hydrogen bonding energy term (dashed-dotted line). Thick lines indicate results of parameterization ORCA-B with activated screening charge correlation correction, thin lines correspond to results of parameterization ORCA-B with the deactivated term.

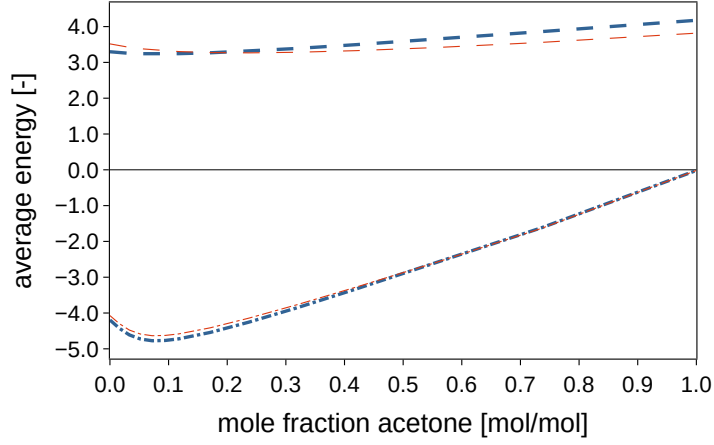


Figure 8: Average interaction energy of acetone in the binary acetone+water system at vapor-liquid equilibrium conditions and 1 atm. For the assignment of the lines to the different interaction energy terms, cf. fig 7.

5. Conclusions

In this work, an approach for the efficient introduction of multiple segment descriptors into COSMO-RS was presented. The improvements brought about by the introduction of the screening charge correlation correction proposed by Klamt et al. [2] was used as an example of how multiple descriptors might improve the model. Nevertheless, applications can reach beyond this term including other descriptors to further refine the model. Demonstrated applications for such additional descriptors include the parameterization of binary element specific dispersion parameters [22], the element specific parameterization of hydrogen bonds [21, 46], the parameterization for ionic molecules [15], or the fine-tuning of the model for specific chemical groups [16]. The implementation of COSMO-RS that is presented in this work integrates this approach and has now been made available as open source software. The implementation was renamed openCOSMO-RS in the process. Additionally, we presented different parameterizations of the model COSMO-RS. For this purpose we implemented a workflow that integrates RDKit and ORCA. Although the resulting accuracy was lower than for the COSMO*therm* based calculations, the workflow and software licenses facilitate the evaluation, further development and refinement of COSMO-RS for a broad range of academic users.

Acknowledgements

We would like to thank Andreas Klamt, Frank Eckert, and Jens Reinisch at Dassault Systemes for the valuable discussions. We would also like to thank Miquel Garcia-Ratés and Christoph Riplinger at FAccTs GmbH for their valuable insights with regard to the application of ORCA for these calculations.

Supporting information

The relevant source code of openCOSMO-RS(py), openCOSMO-RS(c++), as well as the script for the creation of COSMO files with the RDKit/ORCA based workflow can be found on GitHub at <https://github.com/TUHH-TVT>.

References

- [1] A. Klamt, Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena, *The Journal of Physical Chemistry* 99 (7) (1995) 2224–2235.
- [2] A. Klamt, V. Jonas, T. Bürger, J. C. W. Lohrenz, Refinement and Parametrization of COSMO-RS, *J. Phys. Chem. A* 102 (26) (1998) 5074–5085. doi:10.1021/jp980017s.
- [3] F. Eckert, A. Klamt, Fast solvent screening via quantum chemistry: COSMO-RS approach, *AIChE Journal* 48 (2) (2002) 369–385.
- [4] COSMOtherm, Release 19; COSMOlogic GmbH & Co. KG, a Dassault Systèmes company.
- [5] S.-T. Lin, S. I. Sandler, A Priori Phase Equilibrium Prediction from a Segment Contribution Solvation Model, *Industrial & Engineering Chemistry Research* 41 (5) (2002) 899–913. doi:10.1021/ie001047w.
- [6] C.-M. Hsieh, S. I. Sandler, S.-T. Lin, Improvements of COSMO-SAC for vapor–liquid and liquid–liquid equilibrium predictions, *Fluid Phase Equilibria* 297 (1) (2010) 90–97. doi:10.1016/j.fluid.2010.06.011.
- [7] C.-M. Hsieh, S.-T. Lin, J. Vrabec, Considering the dispersive interactions in the COSMO-SAC model for more accurate predictions of fluid phase behavior, *Fluid Phase Equilibria* 367 (2014) 109–116. doi:10.1016/j.fluid.2014.01.032.
- [8] I. H. Bell, E. Mickoleit, C.-M. Hsieh, S.-T. Lin, J. Vrabec, C. Breitkopf, A. Jäger, A Benchmark Open-Source Implementation of COSMO-SAC, *J. Chem. Theory Comput.* 16 (4) (2020) 2635–2646. doi:10.1021/acs.jctc.9b01016.
- [9] H. Grensemann, J. Gmehling, Performance of a Conductor-Like Screening Model for Real Solvents Model in Comparison to Classical Group Contribution Methods, *Ind. Eng. Chem. Res.* 44 (5) (2005) 1610–1624. doi:10.1021/ie049139z.
- [10] C. C. Pye, T. Ziegler, E. van Lenthe, J. N. Louwen, An implementation of the conductor-like screening model of solvation within the Amsterdam density functional package — Part II. COSMO for real solvents, *Canadian Journal of Chemistry* 87 (7) (2009) 790–797. doi:10.1139/V09-008.
- [11] T. Banerjee, M. K. Singh, A. Khanna, Prediction of Binary VLE for Imidazolium Based Ionic Liquid Systems Using COSMO-RS, *Industrial & Engineering Chemistry Research* 45 (9) (2006) 3207–3219. doi:10.1021/ie051116c.

- [12] R. P. Gerber, R. d. P. Soares, Prediction of Infinite-Dilution Activity Coefficients Using UNIFAC and COSMO-SAC Variants, *Ind. Eng. Chem. Res.* 49 (16) (2010) 7488–7496. doi:10.1021/ie901947m.
- [13] R. Franke, B. Hannebauer, On the influence of basis sets and quantum chemical methods on the prediction accuracy of COSMO-RS, *Physical Chemistry Chemical Physics* 13 (48) (2011). doi:10.1039/c1cp22317h.
- [14] A. Klamt, G. J. P. Krooshof, R. Taylor, COSMOSPACE: Alternative to conventional activity-coefficient models, *AIChE Journal* 48 (10) (2002) 2332–2349. doi:10.1002/aic.690481023.
- [15] T. Gerlach, S. Müller, I. Smirnova, Development of a COSMO-RS based model for the calculation of phase equilibria in electrolyte systems, *AIChE J.* 64 (1) (2018) 272–285. doi:10.1002/aic.15875.
- [16] T. Gerlach, T. Ingram, G. Sieder, I. Smirnova, Modeling the solubility of CO₂ in aqueous methyl diethanolamine solutions with an electrolyte model based on COSMO-RS, *Fluid Phase Equilibria* 461 (2018) 39–50. doi:10.1016/j.fluid.2018.01.007.
- [17] F. Neese, The ORCA program system, *WIREs Comput Mol Sci* 2 (1) (2012) 73–78. doi:10.1002/wcms.81.
- [18] F. Neese, Software update: the ORCA program system, version 4.0, *WIREs Comput Mol Sci* 8 (1) (Jan. 2018). doi:10.1002/wcms.1327.
- [19] A. Klamt, F. Eckert, COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids, *Fluid Phase Equilibria* 172 (1) (2000) 43–72.
- [20] A. Klamt, J. Reinisch, F. Eckert, A. Hellweg, M. Diedenhofen, Polarization charge densities provide a predictive quantification of hydrogen bond energies, *Phys. Chem. Chem. Phys.* 14 (2) (2012) 955–963. doi:10.1039/C1CP22640A.
- [21] A. Klamt, J. Reinisch, F. Eckert, J. Graton, J.-Y. Le Questel, Interpretation of experimental hydrogen-bond enthalpies and entropies from COSMO polarisation charge densities, *Physical Chemistry Chemical Physics* 15 (19) (2013) 7147. doi:10.1039/c3cp44611e.
- [22] A. Klamt, F. Eckert, Prediction, fine tuning, and temperature extrapolation of a vapor liquid equilibrium using COSMOtherm, *Fluid Phase Equilibria* 260 (2) (2007) 183–189. doi:10.1016/j.fluid.2007.07.055.
- [23] A. J. Staverman, The entropy of high polymer solutions. Generalization of formulae, *Recl. Trav. Chim. Pays-Bas* 69 (2) (2010) 163–174. doi:10.1002/recl.19500690203.

- [24] A. Klamt, COSMO-RS: from quantum chemistry to fluid phase thermodynamics and drug design, 1st Edition, Elsevier, Amsterdam, 2005.
- [25] COSMOconf, Version 3.0; COSMOlogic GmbH & Co. KG, a Dassault Systèmes company.
- [26] TURBOMOLE V6.6 2010; A Development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007; TURBOMOLE GmbH, since 2007.
- [27] A. Klamt, M. Diedenhofen, A refined cavity construction algorithm for the conductor-like screening model, *Journal of Computational Chemistry* 39 (21) (2018) 1648–1655. doi:10.1002/jcc.25342.
- [28] J.-P. Ebejer, G. M. Morris, C. M. Deane, Freely Available Conformer Generation Methods: How Good Are They?, *J. Chem. Inf. Model.* 52 (5) (2012) 1146–1158. doi:10.1021/ci2004658.
- [29] RDKit, version 2021.09.4.
URL <https://www.rdkit.org/>
- [30] S. Ehlert, M. Stahn, S. Spicher, S. Grimme, Robust and efficient implicit solvation model for fast semiempirical methods 17 (7) 4250–4261. doi:10.1021/acs.jctc.1c00471.
URL <https://doi.org/10.1021/acs.jctc.1c00471>
- [31] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, S. Grimme, Extended tight-binding quantum chemistry methods 11 (2) e1493. doi:10.1002/wcms.1493.
URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1493>
- [32] xTB, version 6.4.1.
URL <https://github.com/grimme-lab/xtb/>
- [33] ORCA, version 5.0.3.
URL <https://www.faccts.de/>
- [34] NumPy, version 1.22.0.
URL <https://numpy.org/>
- [35] Eigen, version 3.4.0.
URL <https://eigen.tuxfamily.org/>
- [36] pybind11, version 2.9.1.
URL <https://github.com/pybind/pybind11>
- [37] Matlab, version 2019a.
URL <https://mathworks.com/>

- [38] E. C. Voutsas, D. P. Tassios, Prediction of Infinite-Dilution Activity Coefficients in Binary Mixtures with UNIFAC. A Critical Evaluation, *Ind. Eng. Chem. Res.* 35 (4) (1996) 1438–1445. doi:10.1021/ie9503555.
- [39] J. He, C. Zhong, A QSPR study of infinite dilution activity coefficients of organic compounds in aqueous solutions, *Fluid Phase Equilibria* 205 (2) (2003) 303–316. doi:10.1016/S0378-3812(02)00296-0.
- [40] T. Sachsenhauser, S. Rehfeldt, A. Klamt, F. Eckert, H. Klein, Consideration of dimerization for property prediction with COSMO-RS-DARE, *Fluid Phase Equilibria* 382 (2014) 89–99. doi:10.1016/j.fluid.2014.08.030.
- [41] A. Klamt, COSMO-RS for aqueous solvation and interfaces, *Fluid Phase Equilibria* 407 (2016) 152–158. doi:10.1016/j.fluid.2015.05.027.
- [42] F. Eckert, M. Diedenhofen, A. Klamt, Towards a first principles prediction of pK_a : COSMO-RS and the cluster-continuum approach, *Molecular Physics* 108 (3-4) (2010) 229–241. doi:10.1080/00268970903313667.
- [43] R. E. Isele-Holder, B. D. Rabideau, A. E. Ismail, Definition and Computation of Intermolecular Contact in Liquids Using Additively Weighted Voronoi Tessellation, *The Journal of Physical Chemistry A* 116 (18) (2012) 4657–4666. doi:10.1021/jp3021886.
- [44] S. Tshepelevitsh, M. Oss, A. Pung, I. Leito, Evaluating the cosmo-rs method for modeling hydrogen bonding in solution, *ChemPhysChem* 14 (9) (2013) 1909–1919. doi:10.1002/cphc.201300186.
- [45] W. Reinders, C. H. de Minjer, Vapour-liquid equilibria in ternary systems. VI. The system water-acetone-chloroform, *Recl. Trav. Chim. Pays-Bas* 66 (9) (2010) 573–604. doi:10.1002/recl.19470660906.
- [46] T. Ingram, T. Gerlach, T. Mehling, I. Smirnova, Extension of COSMO-RS for monoatomic electrolytes: Modeling of liquid–liquid equilibria in presence of salts, *Fluid Phase Equilibria* 314 (2012) 29–37. doi:10.1016/j.fluid.2011.09.021.