A Benchmark Open-Source Implementation of COSMO-SAC

Article *in* Journal of Chemical Theory and Computation · February 2020

DOI: 10.1021/acs.jctc.9b01016

7 authors, including:

Ian H. Bell
National Institute of Standards and Technology
95 PUBLICATIONS **2,229** CITATIONS
SEE PROFILE

Erik Mickoleit
Technische Universität Dresden
9 PUBLICATIONS **52** CITATIONS
SEE PROFILE

Chieh-Ming Hsieh
National Central University
52 PUBLICATIONS **793** CITATIONS
SEE PROFILE

Shiang-Tai Lin
National Taiwan University
165 PUBLICATIONS **4,918** CITATIONS
SEE PROFILE

A Benchmark Open-Source Implementation of COSMO-SAC

Ian H. Bell,[*,†] Erik Mickoleit,[‡] Chieh-Ming Hsieh,[¶] Shiang-Tai Lin,[§] Jadran Vrabec,[‖] Cornelia Breitkopf,[‡] and Andreas Jäger[‡]

†*Applied Chemicals and Materials Division, National Institute of Standards and Technology, Boulder, CO 80305*

‡*Institute of Power Engineering, Faculty of Mechanical Science and Engineering, Technische Universität Dresden, Helmholtzstraße 14, 01069 Dresden, Germany*

¶*Department of Chemical & Materials Engineering, National Central University, Taoyuan 32001, Taiwan*

§*Department of Chemical Engineering, National Taiwan University, 10617 Taipei City, Taiwan*

‖*Thermodynamics and Process Engineering, Technische Universität Berlin, Ernst-Reuter-Platz 1, 10587 Berlin, Germany*

E-mail: ian.bell@nist.gov

Keywords: sigma profile; vapor-liquid-equilibria; COSMO-SAC; open-source

Abstract

The COSMO-SAC modeling approach has found wide application in science as well as in a range of industries due to its good predictive capabilities. While other models for liquid phases, as for example UNIFAC, are in general more accurate than COSMO-SAC, these models typically contain many adjustable parameters and can be limited in their applicability. In contrast, the COSMO-SAC model only contains a few universal parameters and subdivides the molecular surface area into charged segments that interact with each other. In recent years, additional improvements to the construction of the sigma profiles and evaluation of activity coefficients have been made. In this work, we present a comprehensive description how to postprocess the results of a COSMO calculation through to the evaluation of thermodynamic properties. We also assembled a large database of COSMO files, consisting of 2261 compounds, freely available to academic and noncommercial users.

We especially focus on the documentation of the implementation and provide the optimized source code in C++, wrappers in Python, sample sigma profiles calculated from each approach, as well as tests and validation results. The misunderstandings in the literature relating to COSMO-SAC are described and corrected. The computational efficiency of the implementation is demonstrated.

1 Introduction

The calculation of thermodynamic properties of multi-component mixtures is of great importance for the chemical industry. The reason for this is that experimental measurements of mixtures are very time-consuming, costly or involve a high risk when measuring in extreme conditions or considering toxic fluids. In the past, many successful models have been developed. The accuracy of results of predictive

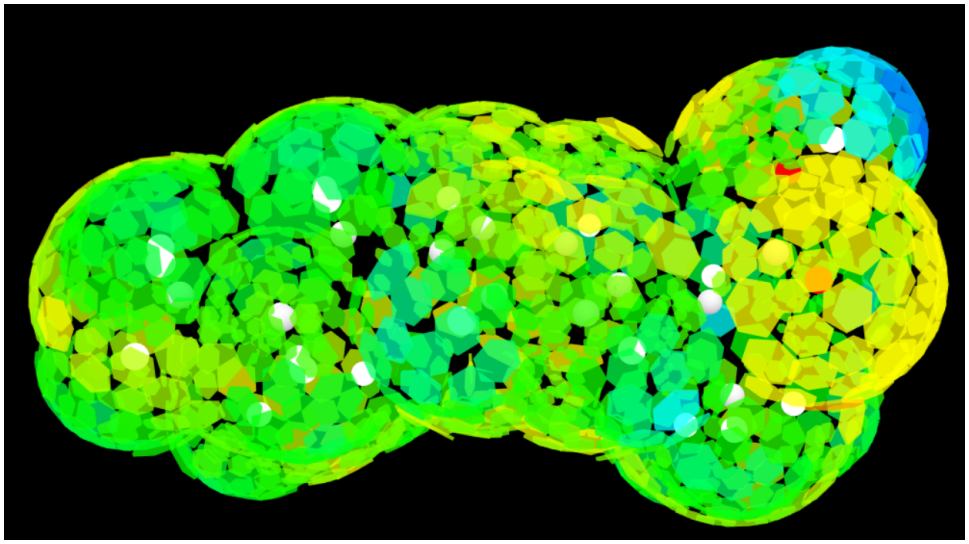

Figure 1: A three-dimensional view of ibuprofen and its segment charge densities (not averaged charge densities) with the visualization tool developed in this work. An interactive HTML file is available in the supplemental material.

models such as the Group Contribution Method (GCM) is based on numerous adjustable parameters which have to be fitted to experimental data. If no adjusted parameters for specific group interactions exist, the model cannot be used. A more predictive alternative to GCM's such as UNIFAC,[1–6] are models based on quantum mechanical conductor-like screening model (COSMO) calculations, originally proposed by Klamt et al. (conductor-like screening model for real solvents, COSMO-RS).[7–9] Based on the COSMO-RS model, Lin and Sandler[10] developed the COSMO segment activity coefficient model (COSMO-SAC). In these models, the interactions of molecules in a mixture are not modeled as pairwise molecular group interactions, but rather as pairwise interactions of charged surface segments of the molecule that can be obtained from quantum mechanical calculations when the molecule is placed in a perfect conductor. COSMO-based models are models for liquid mixtures which typically depend on temperature and composition only, i.e., the pressure-dependency is usually neglected. However, note that pressure-dependency can be introduced to the model by either combining COSMO-based models with equations of state (see, e.g., refs 11–20) or by modifying the model itself (see, e.g., refs 21,22). While COSMO-based models are useful tools for predicting properties of mixtures, their correct implementation can be tricky and time-consuming. The purpose of this work is therefore to provide a reference implementation of three COSMO-SAC models, which are: the original COSMO-SAC model by Lin and Sandler[10] and the modifications of this model by Hsieh et al.[23,24]. These models will be discussed in more detail in section 3.1 (the original model COSMO-SAC-2002[10]), section 3.2 (COSMO-SAC-2010[23]), and section 3.3 (COSMO-SAC-dsp[24]). The COSMO-SAC model can in principle be applied to all types of liquid mixtures. Fingerhut et al. examined thoroughly its performance for over ten thousand binary mixtures based on 2295 compounds, including water.[25] The method can also be used for polymers[26] and, when combined with the Pitzer-Debye-Hückel model, for electrolyte[27,28] and ionic liquids.[29,30] The article is organized thematically by addressing the following aspects of the COSMO-SAC model:

Preprocessing:

1. Generate sigma profile(s) from the results of the COSMO quantum mechanical calculation

2. Split the profile into hydrogen bonding parts if desired

3. Calculate dispersive contributions

Use:

1. Calculate activity coefficients and the excess Gibbs energy
2. Calculate phase equilibria by combining COSMO-SAC with the ideal gas law

2 Part I: COSMO file processing

The COSMO file obtained as the output of a quantum mechanical density-functional-theory calculation is a text file of non-standardized format containing the results of the calculation and information about the molecule. The information needed for creating the sigma profile (see section 2.2) in order to conduct a COSMO-SAC calculation is essentially:

- Volume and surface area of the molecule
- Positions of all nuclei
- Location of each segment patch of the molecule, together with its area and charge

The units of the parameters in the COSMO files are frequently not specified. This unfortunate historical design decision has led to many mistakes in publications and implementations (see for instance Section 2.2). Therefore, users must be exceptionally careful to ensure that a consistent set of units is used. The most frequent source of confusion is the length unit, which is sometimes given in Bohr radius (atomic units), and sometimes in Ångstroms (Å). The conversion factor from Bohr radius to Å is not large in magnitude (1 Bohr radius a_0 $\approx$ 0.52918 Å[31]), further muddying the waters.

2.1 Atoms, Bonds, and Dispersion

The structural information of the molecules is not required for the original model COSMO-SAC-2002 (see section 3.1). However, the structure of the molecule is important for the more advanced models COSMO-SAC-2010 (see section 3.2) and COSMO-SAC-dsp (see section 3.3).

Position data of all nuclei are read from the COSMO file and converted to the Ångstrom scale by multiplying with the conversion factor 0.52917721067 Å/(Bohr radius) as given by CODATA.[31] These nuclei position data are used to determine: a) which atoms are bonded to each other, and b) what type of hybridization of the electron orbitals is present in the atom, used in the analysis of dispersion.

The pairwise distance between each pair of nuclei m and n, in Å, is calculated from

$$d_{mn} = \sqrt{(x_m - x_n)^2 + (y_m - y_n)^2 + (z_m - z_n)^2}. \quad (1)$$

Whether atoms m and n are covalently bonded is determined by comparing their distance and the sums of the covalent radii of the atoms of the pair. If the distance between atoms is less than the sum of the covalent radii, the atoms are assumed to be bonded together. Covalent radii were obtained from Ref.,[32] and these values are also used in the OpenBabel (v2.3.1) cheminformatics library. The covalent radius for carbon was taken to be 0.76 Å (equal to that of the sp^3 hybridization).

2.1.1 Hydrogen bonding

COSMO-SAC-2010 and COSMO-SAC-dsp take different types of hydrogen bonding into account. The molecular surface is separated into segments that are non-hydrogen-bonding (nhb) and segments that form hydrogen bonds. The hydrogen-bonding segments are further divided into hydrogen bonds of an hydroxyl group (OH) and other hydrogen bonds (OT). Other hydrogen bonds (OT) consider surface segments of the atoms nitrogen (N), oxygen (O), and fluorine (F) as well as hydrogen (H) atoms bonded to N or F. Therefore, information about bonding needs to be obtained from the COSMO file. The source code for determining nhb, OT, and OH segments is organized as follows: Once it has been determined which atoms are bonded to each other, the hydrogen bonding class of each atom is determined. If the atom is not in the

set of (O, H, N, F), the atom is not considered to be a candidate for hydrogen bonding in the COSMO-SAC framework, and is given the hydrogen bonding flag of "NHB" (non-hydrogen-bonding). If the atom is an N or an F, the atom is considered to hydrogen bond, but is not in the OH family, and therefore, it is given the "OT" designation (hydrogen bonding, but not OH). If the atom is O or H, the hydrogen bonding class of the atom is:

1. OH: if the atom is O and is bonded to an H, or *vice versa*
2. OT: if the atom is O and is bonded to an atom other than H, or if the atom is H and is bonded to N or F
3. NHB: otherwise

2.1.2 Dispersion

The models COSMO-SAC-dsp[24] and COSMO-SAC 2013[33] take the dispersion contribution to the activity coefficient into account. The dispersive interactions have been considered by Hsieh et al.[24] by assuming equally sized atoms with a size parameter $\sigma = 3$ Å of the Lennard-Jones potential and assigned a dispersion parameter ϵ_{Atom} to each atom forming a molecule. They proposed to compute the dispersion parameter of the molecule $\epsilon_{\text{Molecule}}$ from the relation

$$\frac{\epsilon_{\text{Molecule}}}{k_{\text{B}}} = \frac{1}{N_{\text{Atom}}} \sum_{i=1}^{n} \frac{\epsilon_{\text{Atom},i}}{k_{\text{B}}}, \qquad (2)$$

where $\epsilon_{\text{Atom},i}/k_{\text{B}}$ is the dispersion parameter of atom i, n is the number of atoms in molecule i, and N_{Atom} is the total number of atoms for which $|\epsilon_{\text{Atom},i}/k_{\text{B}}| > 0$. The molecular dispersion parameter for the molecule $\epsilon_{\text{Molecule}}/k_{\text{B}}$ depends on the atomic structure of the molecule and on the dispersion parameters of each atom $\epsilon_{\text{Atom},i}/k_{\text{B}}$. A dispersion parameter is attached to each atom, depending on its orbital hybridization. The orbital hybridization of an atom in a molecule is determined by the number of atoms that are bonded to it. In the case of carbon, sp^3 hybridization corresponds to four, sp^2 to three, and sp to two bonded neighbor atoms. In the case of nitrogen, sp^3 and sp^2 hybridization corresponds to three and two bonded neighbor atoms, respectively, and sp to one bonded neighbor atom.

In addition, the w parameter for the COSMO-SAC-dsp model contains additional molecule specific information (see section 3.3). To calculate w, the dispersive nature of the molecules are classified into categories:

- `DSP_WATER` indicates water
- `DSP_COOH` indicates a molecule with a carboxyl group
- `DSP_HB_ONLY_ACCEPTOR` indicates that the molecule is only a hydrogen bonding acceptor
- `DSP_HB_DONOR_ACCEPTOR` indicates that the molecule is a hydrogen bonding acceptor and donor
- `DSP_NHB` indicates that the molecule is non-hydrogen-bonding

If the molecule is a water molecule or if the molecule contains a COOH-group, the molecule is tagged as `DSP_WATER` or `DSP_COOH`, respectively. Following the implementation of COSMO-SAC-dsp,[24] a molecule is treated as `DSP_HB_ONLY_ACCEPTOR` if the molecule contains any of the atoms O, N, or F but no H-atoms bonded to any of these O, N, or F. Molecules with NH, OH, or FH (but not OH of COOH or water) functional groups are treated as `DSP_HB_DONOR_ACCEPTOR`. If the molecule meets neither of the hydrogen-bonding criteria and is not water and does not contain a COOH group, it is handled as a non-hydrogen bonding molecule and tagged as `DSP_NHB`. For the COSMO-SAC-dsp model, if an atom other than C, H, O, N, F, Cl is included, the associated value of $\epsilon_{\text{Atom},i}/k_{\text{B}}$ is set to an undefined value and the calculation is aborted.

2.1.3 Example

The block of the COSMO file with atom locations looks something like the example in Fig. 2 taken from the database of Mullins et al.[34] In

case of a .cosmo file from DMol^3, the location of the atoms (x, y, z) are given in Å. For instance, the x, y, z position of the first hydrogen nucleus (H1) is (0.888162953 Å, -1.326789759 Å, -0.880602803 Å).

```
Molecular car file :
3.car

!BIOSYM archive 3
PBC=OFF

!DATE     Dec 22 14:07:04 2003
C1       0.000000000   -1.278805452   -0.229129879 XXXX 1      xx      C   0.000
C2       0.000000000    0.000000000    0.616062014 XXXX 1      xx      C   0.000
C3       0.000000000    1.278805452   -0.229129879 XXXX 1      xx      C   0.000
H1       0.888162953   -1.326789759   -0.880602803 XXXX 1      xx      H   0.000
H2      -0.888162953   -1.326789759   -0.880602803 XXXX 1      xx      H   0.000
H3       0.000000000   -2.182013939    0.401203849 XXXX 1      xx      H   0.000
H4       0.880298599    0.000000000    1.281100629 XXXX 1      xx      H   0.000
H5      -0.880298599    0.000000000    1.281100629 XXXX 1      xx      H   0.000
H6       0.000000000    2.182013939    0.401203849 XXXX 1      xx      H   0.000
H7      -0.888162953    1.326789759   -0.880602803 XXXX 1      xx      H   0.000
H8       0.888162953    1.326789759   -0.880602803 XXXX 1      xx      H   0.000
end
end
```

Figure 2: The atom locations (in Å) from the .cosmo file for propane (C_3H_8) from DMol^3.

2.2 Sigma Profile Construction

When doing a quantum mechanical COSMO calculation, the output file supplies a charge density σ_m^* on each surface element with area a_n as well as its charge. These numerical values for the surface charge density are truncated to a few significant digits in the COSMO file. When using the charge density values as given in the COSMO file instead of recalculating the charge density by dividing the charge by the surface area of the segment, the differences can be on the order of a few percent. Therefore, for full replicability of numerical values with the values in this work, the charge density of each segment must be calculated from the charge given in the COSMO file divided by the area given in the COSMO file.

The following averaging equation was originally used by Klamt et al.[8] for COSMO-RS

$$\sigma_m = \frac{\sum_n \sigma_n^* \frac{r_n^2 r_{\mathrm{av}}^2}{r_n^2 + r_{\mathrm{av}}^2} \exp\left(-\frac{d_{mn}^2}{r_n^2 + r_{\mathrm{av}}^2}\right)}{\sum_n \frac{r_n^2 r_{\mathrm{av}}^2}{r_n^2 + r_{\mathrm{av}}^2} \exp\left(-\frac{d_{mn}^2}{r_n^2 + r_{\mathrm{av}}^2}\right)}, \quad (3)$$

where σ_n^* is the original, non-averaged, surface charge of the n-th segment given in elementary charge e coming directly from the COSMO file, $r_n = (a_n/\pi)^{0.5}$, $r_{\mathrm{av}} = 0.5$ Å, and d_{mn} is the distance (in Å) between the centers of the surface segments n and m in Å.

Lin and Sandler[10] used an effective radius for averaging of $r_{\mathrm{eff}} = (a_{\mathrm{eff}}/\pi)^{0.5}$ with $a_{\mathrm{eff}} = 7.5$ Å^2, but otherwise applied a similar methodology as Klamt. Due to a confusion of units in the COSMO file generated by DMol^3 (Lin and Sandler[10] thought the coordinates of the centers of the segments used to calculate the distances were in Å but they were in Bohr radii), in the erratum[35] to their original article Lin and Sandler[10] had to provide a unit conversion parameter f_{decay} to correct the distance d_{mn} from Å to Bohr radius (1 Bohr radius $a_0 \approx 0.52918$ Å,[31] thus $f_{\mathrm{decay}} = 0.52918^{-2} \approx 3.57$). The corrected equation is given as

$$\sigma_m = \frac{\sum_n \sigma_n^* \frac{r_n^2 r_{\mathrm{eff}}^2}{r_n^2 + r_{\mathrm{eff}}^2} \exp\left(-f_{\mathrm{decay}} \frac{d_{mn}^2}{r_n^2 + r_{\mathrm{eff}}^2}\right)}{\sum_n \frac{r_n^2 r_{\mathrm{eff}}^2}{r_n^2 + r_{\mathrm{eff}}^2} \exp\left(-f_{\mathrm{decay}} \frac{d_{mn}^2}{r_n^2 + r_{\mathrm{eff}}^2}\right)}, \quad (4)$$

where d_{mn} is the distance (in Å) between the centers of the surface segments n and m, $r_{\mathrm{eff}} = (a_{\mathrm{eff}}/\pi)^{0.5}$ in Å. Nonetheless, this equation remains dimensionally inconsistent (the argument of the exponential function has units of bohr^2/Å^2); the parameter f_{decay} should therefore be thought of as a dimensionless scaling quantity (only). Hence, the practical interpretation of f_{decay} is that, first, the coordinates of the segments given in bohr are converted to Å and, second, f_{decay} is used to scale the values given in Å back to the numerical values of bohr (keeping Å as unit).

The model of Lin and Sandler[10,35] was made available as a Fortran source code together with a comprehensive sigma-profile database in the very useful work of Mullins et al.[34] They computed the sigma profiles with density-functional-theory calculations using the software DMol^3. Note that the parametrization and the results of COSMO models in general depend on the underlying method and software with which the sigma profiles are calculated.[36] Hence, it is very important for the comparability and evaluation of these models to use exactly the same set of sigma profiles.

It is furthermore important to note that in the Fortran code for the computation of the sigma profiles, Mullins et al.[34] used the same averaging equation as Klamt (Eq. (3)), but with $r_{\mathrm{av}} = 0.81764$ Å. The use of an averaging radius of $r_{\mathrm{av}} = 0.81764$ Å and $f_{\mathrm{decay}} = 1$ is equivalent (when assuming $r_n^2 << r_{\mathrm{av}}^2$) to the assumption of dividing numerator and denominator by f_{decay}, such that the f_{decay} correction is applied to the averaging radius.

Once the averaged value of σ_m has been obtained for each segment m, the $p(\sigma)A_i$ values (probability $p(\sigma)$ of finding a given segment with specified value of σ multiplied with the entire surface area A_i of molecule i, which gives the surface areas $A_i(\sigma_m)$ of molecule i with

charge densities σ_m) need to be obtained on gridded values. The values of σ for which the $p(\sigma)A$ values are to be obtained is generally -0.025 e/Å^2 to 0.025 e/Å^2 in increments of 0.001 e/Å^2, forming a set of 51 points.

Subsequently for each value of σ (in e/Å^2), the (0-based) index corresponding to the left of the value is obtained from

$$i_{\text{left}} = \left\lfloor \frac{\sigma - (-0.025 \text{ e/Å}^2)}{0.001 \text{ e/Å}^2} \right\rfloor, \quad (5)$$

the fractional distance of the value between the left and right edges of the cell are given by

$$w = \frac{\sigma[i_{\text{left}} + 1] - \sigma}{0.001 \text{ e/Å}^2}, \quad (6)$$

which is by definition between 0 and 1. The area of the segment is then distributed between the gridded sigma values above and below the value, according to the weighting parameter w:

$$p(\sigma)A_i[i_{\text{left}}] += wA_n \quad (7)$$
$$p(\sigma)A_i[i_{\text{left}} + 1] += (1 - w)A_n. \quad (8)$$

Figure 3 illustrates the construction of the sigma profile for ethanol. For each patch, the value of sigma is obtained, and from that, the value of sigma is then distributed amongst the gridded values. As can be seen in this plot, the histogram is constructed from a relatively small number of patches.

2.2.1 Example

Figure 4 gives an example of a COSMO file from DMol^3. It is important to note that the units are not specified for the areas or the charge. It is especially problematic that the length units differ within the same line of the file (the area in units of Å^2, and segment positions are in Bohr radius (from the atomic units (a.u.) system of measurement)), a source of confusion for many authors, ourselves included. Nonetheless, this is a standard file format.

2.3 Splitting of Profiles

Lin et al.[37] proposed to split the sigma profile into hydrogen-bonding (hb) and non-hydrogen-bonding (nhb) segments with $p_i(\sigma) = p_i^{\text{nhb}}(\sigma) + p_i^{\text{hb}}(\sigma)$. Hydrogen-bonding atoms were defined to be oxygen, nitrogen, and fluorine atoms as well as hydrogen atoms bound to one of oxygen, nitrogen, or fluorine. Hence, all surfaces belonging to the aforementioned atoms contribute to the sigma profile $p_i^{\text{hb}}(\sigma)$ and the other atoms forming molecule i contribute to the sigma profile $p_i^{\text{nhb}}(\sigma)$. Hsieh et al.[23] suggested to further split the hydrogen-bonding sigma profile into interactions of surfaces belonging to groups of oxygen and hydrogen (OH) and surfaces belonging to other groups (OT).

Each atom of the molecule is assigned to be in one of the hydrogen-bonding classes:

- NHB: the atom is not a candidate to hydrogen bond
- OH: the atom is either the oxygen or the hydrogen in a OH hydrogen-bonding pair
- OT: the atom is N, F, or an oxygen that is not part of an OH bonding group

Though these classes are consistent with the work of Hsieh et al.,[23] they do not consider the fact that the H of a COOH group (likewise for other similar groups) is delocalized between the two oxygens of the group.

$$p(\sigma) = p^{\text{nhb}}(\sigma) + p^{\text{OH}}(\sigma) + p^{\text{OT}}(\sigma). \quad (9)$$

A currently undocumented feature of the profile splitting is that the contribution for a given segment is deposited into the NHB, OH, or OT sigma profiles depending on its *averaged* charge density in the following manner:

- If the segment belongs to an O atom, and the hydrogen-bonding class of the atom is OH, and the averaged charge density value of the segment is greater than zero, the segment goes into the OH profile.
- If the segment belongs to an H atom, and the hydrogen-bonding class of the atom

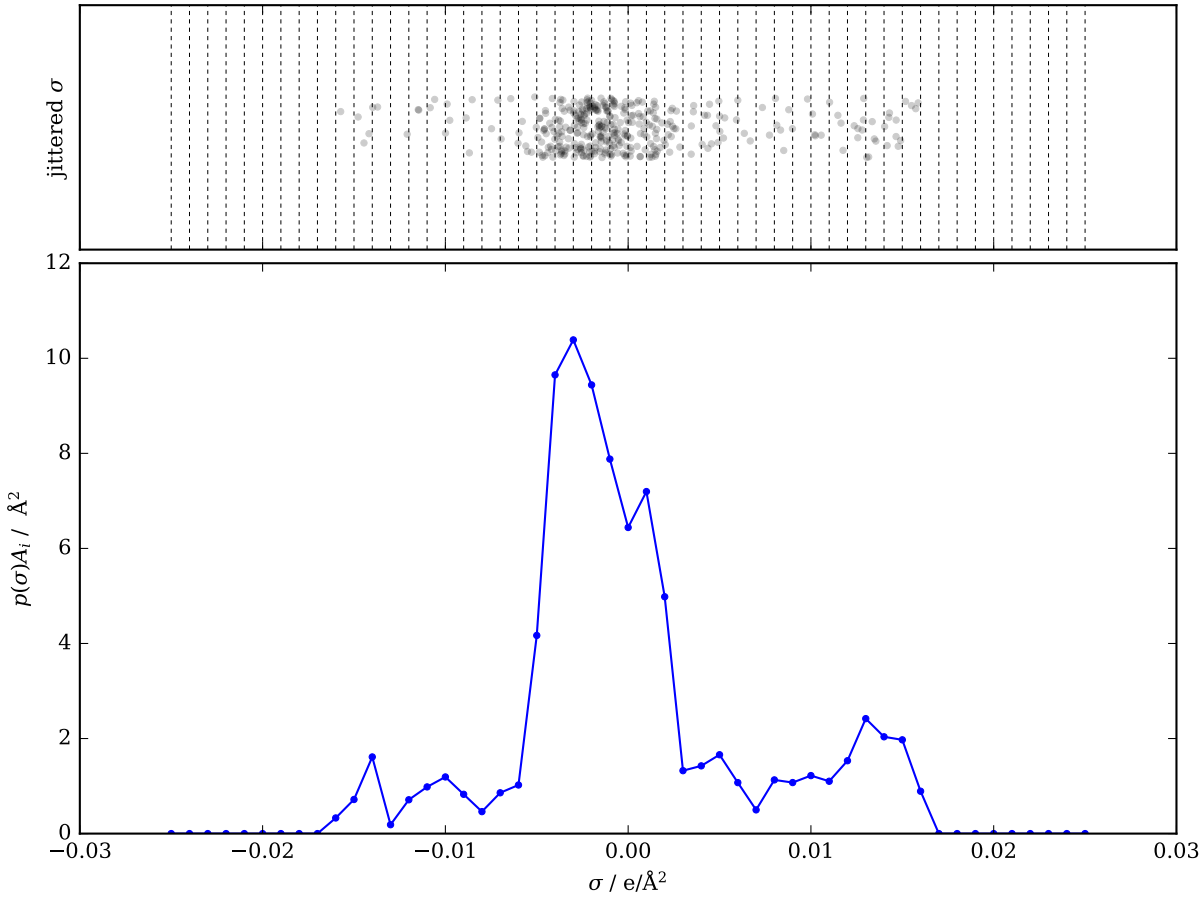

Figure 3: Sigma values for ethanol (single profile, with averaging scheme of Mullins) Top: Values of averaged charge densities on segments (randomly jittered in the vertical direction). Vertical lines correspond to the nodes at which the sigma profile will be generated. Bottom: Sigma profile generated from these σ values.

is OH, and the averaged charge density value of the segment is less than zero, the segment goes into the OH profile.

- If the segment belongs to an O, N, or F atom, and the hydrogen-bonding class of the atom is OT, and the averaged charge density value of the segment is greater than zero, the segment goes into the OT profile.
- If the segment belongs to an H atom, and the hydrogen-bonding class of the atom is OT, and the averaged charge density value of the segment is less than zero, the segment goes into the OT profile.
- Otherwise, the segment goes into the NHB profile.

By definition, the superpositioned profiles (NHB + OH + OT) *must* equal the original profile (see Eq. (9)). Wang et al.[38] proposed the use of a Gaussian-type function for the probability P of a hydrogen-bonding segment to indeed form a hydrogen bond

$$P^{\mathrm{hb}}(\sigma) = 1 - \exp\left(-\frac{\sigma^2}{2\sigma_0^2}\right) \qquad (10)$$

with $\sigma_0 = 0.007$ e/Å^2. This probability function considers that not all surfaces belonging to potentially hydrogen-bond forming atoms in fact form hydrogen bonds. With the probability function for hydrogen bonding according to Eq. (10), it follows

```
Segment information:

total number of segments:    438
n             - segment number
atom          - atom associated with segment n
position      - segment coordinates
charge        - segment charge
area          - segment area
potential     - solute potential on a segment


n   atom        position (X, Y, Z) [au]        charge      area      charge/area  potential


1    1       2.59824  -5.10444   0.12280     0.00132     0.23228     0.00570    -0.01662
2    1      -0.06797  -5.22869  -2.95721     0.00148     0.37165     0.00398    -0.02163
3    1      -0.37610  -1.61747  -4.10780     0.00176     0.32519     0.00541    -0.01857
4    1      -2.88843  -3.73710   1.61574     0.00190     0.41810     0.00453    -0.01385
5    1      -2.80596  -4.61858   0.81676     0.00116     0.23228     0.00500    -0.01516
...
```

Figure 4: The first five segments of a COSMO file from DMol^3. The charge is in units of e, the area in units of Å^2, and segment positions are in Bohr radius (from the atomic units (a.u.) system of measurement)

$$p^{\mathrm{nhb}}(\sigma) = \frac{A_i^{\mathrm{nhb}}(\sigma)}{A_i} + \frac{A_i^{\mathrm{hb}}(\sigma)}{A_i}\left[1 - P^{\mathrm{hb}}(\sigma)\right], \tag{11}$$

$$p^{\mathrm{OH}}(\sigma) = \frac{A_i^{\mathrm{OH}}(\sigma)}{A_i} P^{\mathrm{hb}}(\sigma), \tag{12}$$

$$p^{\mathrm{OT}}(\sigma) = \frac{A_i^{\mathrm{OT}}(\sigma)}{A_i} P^{\mathrm{hb}}(\sigma), \tag{13}$$

where $A_i^{\mathrm{nhb}}(\sigma)$ is the surface area of non-hydrogen-bonding atoms of molecule i, $A_i^{\mathrm{hb}}(\sigma) = A_i^{\mathrm{OH}}(\sigma) + A_i^{\mathrm{OT}}(\sigma)$ is the surface area of hydrogen-bonding segments of molecule i, $A_i^{\mathrm{OH}}(\sigma)$ is the surface area of hydrogen-bonding segments of OH-groups of molecule i, and $A_i^{\mathrm{OT}}(\sigma)$ is the surface area of hydrogen-bonding segments other than OH.

2.4 Implementation, Validation and Verification

A Python script `profiles/to_sigma.py` was written that fully automates the process of reading in a COSMO file from a DMol^3 calculation and generates the split profiles as well as calculates the dispersion parameters. The output file with sigma profiles is a space-delimited text file with additional metadata stored in JSON (JavaScript Object Notation) format in the header of the sigma profile. In case of discrepancies between the description above and the Python code, the latter should be used as the reference. This Python script makes heavy use of vectorized matrix operations of the numpy matrix library, especially in case of the sigma averaging, the computationally most expensive part of sigma profile generation. Furthermore, the sigma profile generation from COSMO files is automated with the Python script `profiles/generate_all_profiles.py`, which generates sigma profiles in parallel. All of these scripts are available in the provided code.

The sigma profiles, dispersion flags, and dispersion parameters are available in the supplemental material for 2261 molecular species. Parameters to carry out the sigma profile averaging (r_{av}, f_{decay}) are documented in the header

of each sigma profile for reproducibility. All parameters are specified along with their units (where appropriate).

Before carrying out any COSMO-SAC calculations, users should first verify that their implementation yields exactly the same sigma profiles compiled in the supplemental material when processing the provided COSMO files. Values of $p(\sigma)A$ should agree to within at least 10^{-15}.

Finally, a segment charge visualization tool was written with the three.js javascript library. This tool, driven by a Python-based script, reads the COSMO file and generates an HTML file with the data of the locations and orientations of the segments (and the atoms). The visualization scene is constructed with three.js and behind the scenes, WebGL powers the 3D visualization, allowing for a seamless visualization in three dimensions, even for rather large molecules. This approach is cross-platform, fully open-source, and while intended to be rudimentary, could easily be extended by users for their own application. An example of the visualization tool is provided in Fig. 1, and other examples are available.

3 Part II: Activity Coefficient Calculation

3.1 Model of Lin and Sandler – COSMO-SAC 2002

To generate the sigma profile, in accordance with the work of Mullins et al.,[39] we have used the equation from Klamt (Eq. (3)) with the value of effective radius defined by $r_{\mathrm{av}} = 0.81764$ Å, and all radii in Å. The model parameters are summarized in Table 1.

According to Lin and Sandler,[10] the activity coefficient $\gamma_{i,\mathrm{S}}$ of component i in the liquid mixture S can be obtained from the equation

$$\ln\left(\gamma_{i,\mathrm{S}}\right) = \ln\left(\gamma^{\mathrm{c}}_{i,\mathrm{S}}\right) + \ln\left(\gamma^{\mathrm{r}}_{i,\mathrm{S}}\right), \qquad (14)$$

where the combinatorial part $\ln\left(\gamma^{\mathrm{c}}_{i,\mathrm{S}}\right)$ accounts for the size and shape differences of the molecules. This quantity is usually described by the Staverman-Guggenheim combinatorial term

$$\ln\left(\gamma^{\mathrm{c}}_{i,\mathrm{S}}\right) = \ln\left(\frac{\phi_i}{x_i}\right) + \frac{z}{2} q_i \ln\left(\frac{\theta_i}{\phi_i}\right) + l_i - \frac{\phi_i}{x_i}\sum_j x_j l_j, \qquad (15)$$

with

$$\theta_i = \frac{x_i q_i}{\sum_j x_j q_j}, \qquad (16)$$

$$\phi_i = \frac{x_i r_i}{\sum_j x_j r_j}, \qquad (17)$$

$$l_i = \frac{z}{2}\left(r_i - q_i\right) - \left(r_i - 1\right), \qquad (18)$$

and

$$r_i = V_i / r_0, \qquad (19)$$

Here $r_0 = 66.69$ Å^3 denotes the normalized volume parameter and

$$q_i = A_i / q_0, \qquad (20)$$

where $q_0 = 79.53$ Å^2 denotes the normalized surface area parameter and z is the coordination number, which was chosen to be 10. Note that r_0 is not needed to calculate the combinatorial term as it cancels out internally in Eq. (15), see, e.g., Ref. 33. V_i is the molecular volume of component i and A_i is the molecular surface area of component i coming from the COSMO calculations. Note that in the article by Lin and Sandler,[10] there are misplaced parentheses in the equation for l_i, which was corrected in Eq. (18) (compare, e.g., Ref.[1]).

In the case of infinite dilution, when one of the mole fractions goes to zero, Eqs. (16) and (17) are ill-defined because division by zero occurs. The terms in Eq. (15) leading to division by zero can be rewritten as

$$\frac{\theta_i}{x_i} = \frac{q_i}{\sum_j x_j q_j}, \tag{21}$$

$$\frac{\phi_i}{x_i} = \frac{r_i}{\sum_j x_j r_j}, \tag{22}$$

$$\frac{\theta_i}{\phi_i} = \frac{\frac{\theta_i}{x_i}}{\frac{\phi_i}{x_i}}, \tag{23}$$

and were used throughout.

The residual part $\ln\left(\gamma^{\mathrm{r}}_{i,\mathrm{S}}\right)$, which is also called the restoring free energy part, mainly accounts for electrostatic interactions between the molecules in the mixture. According to a statistical mechanical derivation by Lin and Sandler,[10] the residual part of the activity coefficient can be obtained as follows

$$\ln\left(\gamma^{\mathrm{r}}_{i,\mathrm{S}}\right) = n_i \sum_{\sigma_m} p_i\left(\sigma_m\right)\left[\ln\left(\Gamma_{\mathrm{S}}\left(\sigma_m\right)\right) - \ln\left(\Gamma_i\left(\sigma_m\right)\right)\right], \tag{24}$$

where n_i denotes the number of surface segments of molecule i with a standard segment surface area a_{eff} and can be calculated according to:

$$n_i = \frac{A_i}{a_{\mathrm{eff}}} \tag{25}$$

where the sum on the right hand side goes over all charge densities σ_n in the mixture. Note that Eq. (28) needs to be solved numerically

σ_m is the screening charge density of segment m, which is the average screening charge of the surface segment divided by a_{eff}, $p_i\left(\sigma_m\right)$ denotes the probability of finding a segment with screening charge density σ_m on the surface of component i, $\Gamma_{\mathrm{S}}\left(\sigma_m\right)$ is the activity coefficient of segment m in the mixture, and $\Gamma_i\left(\sigma_m\right)$ is the activity coefficient of segment m in the mixture of segments of only pure component i. The quantity $p_i\left(\sigma_m\right)$ is called the sigma profile of pure component i and is defined as

$$p_i\left(\sigma_m\right) = \frac{A_i\left(\sigma_m\right)}{A_i}, \tag{26}$$

where $A_i(\sigma_m)$ is the surface area with screening charge density σ_m of a molecule of species i and A_i is again the entire surface area of the molecule of species i. The sigma profile of the mixture S can then be obtained by

$$p_{\mathrm{S}}\left(\sigma_m\right) = \frac{\sum_{i=1}^{N} x_i A_i p_i\left(\sigma_m\right)}{\sum_{i=1}^{N} x_i A_i}, \tag{27}$$

where x_i is the mole fraction of component i. The segment activity coefficient in the mixture can be calculated from

$$\ln\left(\Gamma_{\mathrm{S}}\left(\sigma_m\right)\right) = -\ln\left\{\sum_{\sigma_n} p_{\mathrm{S}}\left(\sigma_n\right)\Gamma_{\mathrm{S}}\left(\sigma_n\right)\exp\left[-\frac{\Delta W\left(\sigma_m,\sigma_n\right)}{RT}\right]\right\} \tag{28}$$

for the segment activity coefficients $\Gamma_S\left(\sigma_m\right)$. The quantity $\Delta W\left(\sigma_m,\sigma_n\right)$ is called the exchange energy and can be calculated from

$$\Delta W(\sigma_m,\sigma_n) = \left(\frac{\alpha'}{2}\right)\left(\sigma_m+\sigma_n\right)^2 + c_{\mathrm{hb}}\max\left[0,\sigma_{\mathrm{acc}}-\sigma_{\mathrm{hb}}\right]\min\left[0,\sigma_{\mathrm{don}}+\sigma_{\mathrm{hb}}\right] \tag{29}$$

where the first term on the right hand side is the misfit energy, accounting for the electrostatic interactions, and the second term on the right hand side accounts for hydrogen-bonding interactions. The values of the generalized parameters are: $\alpha' = 16466.72$ kcal Å^4 mol^{-1} e^{-2}, $c_{\mathrm{hb}} = 85580$ kcal Å^4 mol^{-1} e^{-2}, and $\sigma_{\mathrm{hb}} = 0.0084$ e Å^{-2} (with 1 kcal = 4184 J). Note that – in accordance with the FORTRAN code supplied by Mullins et al.[34] – the value for α' given in the article of Lin and Sandler[10] was

used rather than the different value provided in the article by Mullins et al.[34] Furthermore, $\sigma_{\mathrm{acc}} = \max(\sigma_m, \sigma_n)$ and $\sigma_{\mathrm{don}} = \min(\sigma_m, \sigma_n)$) holds. $\Gamma_i(\sigma_m)$ and $\Gamma_{\mathrm{S}}(\sigma_m)$ have a similar form

$$\ln\left(\Gamma_i(\sigma_m)\right) = -\ln\left\{\sum_{\sigma_n} p_i(\sigma_n)\,\Gamma_i(\sigma_n)\exp\left[-\frac{\Delta W(\sigma_m,\sigma_n)}{RT}\right]\right\}. \tag{30}$$

3.2 Model of Hsieh et al. – COSMO-SAC 2010

In 2010, Hsieh et al.[23] suggested an improvement of COSMO-SAC for phase equilibrium calculations based on the modifications published by Lin et al.[37] and Wang et al.[38] Hsieh et al. proposed two modifications for Eq. (29), which in their model reads

$$\Delta W\left(\sigma_m^t,\sigma_n^s\right) = c_{\mathrm{ES}}(T)\cdot\left(\sigma_m^t+\sigma_n^s\right)^2 - c_{\mathrm{hb}}\left(\sigma_m^t,\sigma_n^s\right)\left(\sigma_m^t-\sigma_n^s\right)^2, \tag{31}$$

where the superscripts t and s denote different types of sigma profiles. The first modification concerns the electrostatic interaction parameter c_{ES}, which was made temperature dependent

$$c_{\mathrm{ES}} = A_{\mathrm{ES}} + \frac{B_{ES}}{T^2}, \tag{32}$$

with $A_{\mathrm{ES}} = 6525.69$ kcal Å^4 mol^{-1} e^{-2} and $B_{\mathrm{ES}} = 1.4859\times10^8$ kcal Å^4 K^2 mol^{-1} e^{-2}. The second modification concerns the hydrogen-bonding term given in Eq. (31). With this distinction, the parameter c_{hb} is defined as follows

$$c_{\mathrm{hb}}\left(\sigma_m^t,\sigma_n^s\right) = \begin{cases} c_{\mathrm{OH-OH}} & \text{if } s=t=\mathrm{OH} \text{ and } \sigma_m^t\cdot\sigma_n^s<0 \\ c_{\mathrm{OT-OT}} & \text{if } s=t=\mathrm{OT} \text{ and } \sigma_m^t\cdot\sigma_n^s<0 \\ c_{\mathrm{OH-OT}} & \text{if } s=\mathrm{OH},\ t=\mathrm{OT}, \text{ and } \sigma_m^t\cdot\sigma_n^s<0 \\ 0 & \text{otherwise} \end{cases}, \tag{33}$$

where $c_{\mathrm{OH-OH}} = 4013.78$ kcal Å^4 mol^{-1} e^{-2}, $c_{\mathrm{OT-OT}} = 932.31$ kcal Å^4 mol^{-1} e^{-2}, and $c_{\mathrm{OH-OT}} = 3016.43$ kcal Å^4 mol^{-1} e^{-2}. Due to the separation of the sigma profiles into nhb, OT, and OH contributions, an additional sum over the different sigma-profiles needs to be introduced in Eqs. (24), (28), and (30). Equation (24) becomes

$$\ln\left(\gamma_{i,\mathrm{S}}^{\mathrm{r}}\right) = n_i \sum_{t}^{\mathrm{nhb,OH,OT}} \sum_{\sigma_m} p_i^t\left(\sigma_m^t\right)\left[\ln\left(\Gamma_{\mathrm{S}}^t\left(\sigma_m^t\right)\right) - \ln\left(\Gamma_i^t\left(\sigma_m^t\right)\right)\right], \tag{34}$$

and Eq. (29) becomes

$$\ln\left(\Gamma_{\mathrm{S}}^t\left(\sigma_m^t\right)\right) = -\ln\left\{\sum_{s}^{\mathrm{nhb,OH,OT}} \sum_{\sigma_n} p_{\mathrm{S}}^s(\sigma_n^s)\,\Gamma_{\mathrm{S}}^s(\sigma_n^s)\exp\left[-\frac{\Delta W(\sigma_m^t,\sigma_n^s)}{RT}\right]\right\}. \tag{35}$$

Table 1: Parameters of COSMO-SAC 2002 as implemented by Mullins et al.[34] and their SI equivalents. The value for the charge of the electron is $e = 1.602176634 \times 10^{-19}$ C.

Parameter	Value	Value (SI)
q_0	79.53 Å^2	79.53×10^{-20} m^2
r_0	66.69 Å^3	66.69×10^{-30} m^3
z	10	10
r_{av}	0.81764 Å	8.1764×10^{-11} m
a_{eff}	7.5 Å^2	7.5×10^{-20} m^2
c_{hb}	85580 kcal Å^4 mol^{-1} e^{-2}	$1.3949003091892562\times10^{6}$ J m^{-4} mol^{-1} C^{-2}
σ_{hb}	0.0084 e Å^{-2}	0.134582837256 C m^{-2}
α' †	16466.72 kcal Å^4 mol^{-1} e^{-2}	$2.6839720518033312\times10^{5}$ J m^{-4} mol^{-1} C^{-2}
R	0.001987 kcal mol^{-1} K^{-1}	8.313608 J mol^{-1} K^{-1}

†: Mullins et al. used an erroneous value in their Table 1. The correct value for the misfit energy parameter[10] is obtained from $(0.3 f_{\mathrm{pol}} \cdot a_{\mathrm{eff}}^{3/2})/\varepsilon_0$, with $\varepsilon_0 = 2.395 \times 10^{-4}$ (e^2 mol)/(kcal Å), $\epsilon = 3.667$, $f_{\mathrm{pol}} = (\epsilon - 1)/(\epsilon + 0.5)$, and $a_{\mathrm{eff}} = 7.5$ Å^2, according to the FORTRAN code of Mullins et al.

Equation (30) can again be obtained by changing the index S to i in Eq. (35). For COSMO-SAC 2010, Hsieh et al.[23] used the sigma profile database of Mullins et al.[34]
Furthermore the value for a_{eff} was changed to 7.25 Å^2 and the sigma profile averaging equation according to Eq. (4) was used. A summary of the model parameters is given in Table 2.

3.3 Model of Hsieh at el. – COSMO-SAC-dsp

On the basis of their model modification summarized in section 2.2, Hsieh et al.[24] proposed to also take dispersive interactions between the molecules into account, which were entirely neglected before. The activity coefficient of component i in the mixture S then becomes

$$\ln\left(\gamma_{i,\mathrm{S}}\right) = \ln\left(\gamma_{i,\mathrm{S}}^{\mathrm{c}}\right) + \ln\left(\gamma_{i,\mathrm{S}}^{\mathrm{r}}\right) + \ln\left(\gamma_{i,\mathrm{S}}^{\mathrm{dsp}}\right), \quad (36)$$

with $\gamma_{i,\mathrm{S}}^{\mathrm{dsp}}$ being the contribution to the activity coefficient due to dispersion. The combinatorial part $\ln\left(\gamma_{i,\mathrm{S}}^{\mathrm{c}}\right)$ and the residual part $\ln\left(\gamma_{i,\mathrm{S}}^{\mathrm{r}}\right)$ are calculated in the same way and with the same parameters as given in sections 2.1 and 2.2, respectively. Hsieh et al.[24] suggest the use of the one-constant Margules equation for the calculation of the dispersive interaction and give the following equations for a binary mixture of components 1 and 2

$$\ln\left(\gamma_{1,\mathrm{S}}^{\mathrm{dsp}}\right) = A x_2^2 \text{ and } \ln\left(\gamma_{2,\mathrm{S}}^{\mathrm{dsp}}\right) = A x_1^2. \quad (37)$$

As given in the article by Hsieh et al.,[24] the parameter A can be calculated according to

$$A = w\left[0.5\left(\epsilon_1 + \epsilon_2\right) - \sqrt{\epsilon_1 \epsilon_2}\right], \quad (38)$$

with the definition of w given in a corrigendum by Hsieh et al.[40]

$$w = \begin{cases} -0.27027 & \text{if water + hb-only-acceptor} \\ -0.27027 & \text{if COOH + (nhb or hb-only-acceptor)} \\ -0.27027 & \text{if water + COOH} \\ 0.27027 & \text{otherwise} \end{cases}. \quad (39)$$

Table 2: Parameters of COSMO-SAC 2010 and their SI equivalents. The value for the charge of the electron is $e = 1.602176634 \times 10^{-19}$ C

Parameter	Value	Value (SI)
q_0	79.53 Å^2	79.53 $\times 10^{-20}$ m^2
r_0	66.69 Å^3	66.69 $\times 10^{-30}$ m^3
z	10	10
a_{eff}	7.25 Å^2	7.25 $\times 10^{-20}$ m^2
r_{eff}	$(a_{\text{eff}}/\pi)^{0.5}$	
f_{decay}	3.57	3.57
$c_{\text{OH-OH}}$	4013.78 kcal Å^4 mol^{-1} e^{-2}	6.542209585204081$\times 10^4$ J m^4 mol^{-1} C^{-2}
$c_{\text{OT-OT}}$	932.31 kcal Å^4 mol^{-1} e^{-2}	1.5196068091379239$\times 10^4$ J m^4 mol^{-1} C^{-2}
$c_{\text{OH-OT}}$	3016.43 kcal Å^4 mol^{-1} e^{-2}	4.916591656517583$\times 10^4$ J m^4 mol^{-1} C^{-2}
σ_0	0.007 e Å^{-2}	0.11215236437999998 C m^2
A_{ES}	6525.69 kcal Å^4 mol^{-1} e^{-2}	1.06364652940795$\times 10^5$ J m^4 mol^{-1} C^{-2}
B_{ES}	1.4859×10^8 kcal Å^4 K^2 mol^{-1} e^{-2}	2.421923778247623$\times 10^9$ J m^4 K^2 mol^{-1} C^{-2}
N_{A}	$6.022140758 \times 10^{23}$ mol^{-1}	$6.022140758 \times 10^{23}$ mol^{-1}
k_{B}	$1.38064903 \times 10^{-23}$ J K^{-1}	$1.38064903 \times 10^{-23}$ J K^{-1}
R	$k_{\text{B}}N_{\text{A}}/4184$ kcal mol^{-1} K^{-1}	$k_{\text{B}}N_{\text{A}}$ J mol^{-1} K^{-1}

Table 3: Dispersion parameters for atom types in COSMO-SAC-dsp as implemented by Hsieh et al.[24]

Atom type i	$(\epsilon_{\text{Atom},i}/k_{\text{B}})$ / K	Note
C (sp^3)	115.7023	bonded to four others
C (sp^2)	117.4650	bonded to three others
C (sp)	66.0691	bonded to two others
N (sp^3)	15.4901	bonded to three others
N (sp^2)	84.6268	bonded to two others
N (sp)	109.6621	bonded to one other
–O–	95.6184	bonded to two others
=O	−11.0549	double-bonded O
F	52.9318	bonded to one other
Cl	104.2534	bonded to one other
H (water)	58.3301	H in water
H (OH)	19.3477	H-O bond (not water)
H (NH)	141.1709	H bonded to N
H (other)	0	
other	invalid	

Hsieh et al.[24] define substances as hb-only-acceptor, if they are able to form a hydrogen-bond by accepting a proton from its neighbor and hb-donor-acceptors as substances that are able to form hydrogen-bonds by either providing or accepting a proton from its neighbors. All substances containing a carboxyl group are denoted with COOH. Note that there is a typographical error in the corrigendum, where the value $w = -0.27027$ is proposed for the combination of "COOH + nhb or hb-only-acceptor", whereas in the article, this value for w is given for the combination of "COOH + nhb or hb-donor-acceptor". Furthermore, note that there also is a confusion of units in the original article, as the constant A should be dimensionless and the dispersion parameters are usually divided by Boltzmann's constant $k_\mathrm{B} = 1.380649 \times 10^{-23}$ J K^{-1} (value taken from[41]). Therefore, we implemented Eq. (38) and Eq. (39) as follows

$$A = w \left[0.5 \left(\frac{\epsilon_1}{k_\mathrm{B}} + \frac{\epsilon_2}{k_\mathrm{B}} \right) - \sqrt{\frac{\epsilon_1}{k_\mathrm{B}} \frac{\epsilon_2}{k_\mathrm{B}}} \right], \quad (40)$$

and

$$w = \begin{cases} -0.27027 \text{ K}^{-1} & \text{if water + hb-only-acceptor} \\ -0.27027 \text{ K}^{-1} & \text{if COOH + (nhb or hb-donor-acceptor)} \\ -0.27027 \text{ K}^{-1} & \text{if water + COOH} \\ 0.27027 \text{ K}^{-1} & \text{otherwise} \end{cases} . \quad (41)$$

The dispersion parameters of the atoms have been fitted to experimental data by Hsieh et al.[24] and are listed in Table 3 (already considering the corrigendum[40]). The other model parameters stayed the same as already given in Table 2.

The QM/COSMO calculation results from the University of Delaware database[33] were used for the development of COSMO-SAC-dsp model. This database can be considered as a revised and extended version of the VT-database[34,39] and was developed with the cooperation of Stanley Sandler's research group at the University of Delaware and Dr. S. Lustig, formerly at DuPont. The basis set used in Dmol^3 was the GGA/VWN-BP/DNP functional with double numerical basis with polarization functions (DNP). The detailed procedure for obtaining the equilibrium geometry and the screening charges can be found in Ref. 34. The effects from using different combination of DFT methods and basis set were studied by Chen et al.[42] More details are available in Xiong et al.[33] A fully-open-source set of sigma profiles generated from an open-source quantum chemical tool is available[43] at `https://github.com/lvpp/sigma` for the release 18.07,[44] and their use in a COSMO approach has been investigated previously.[45,46]

4 Implementation Details for the COSMO-SAC Models

The numerical method used to solve the nonlinear system of Eqs. (28) and (35) is the successive substitution method. This method is the solver used in the code of Mullins et al.[34] forming the basis of this implementation. Successive substitution is characterized as being both reliable and slowly convergent. Analysis of the successive substitution method and a comparison with the Newton-Raphson method is provided in Possani and de P. Soares.[47]

To solve the segment activity coefficient of Eqs. (28) and (35), initial values have to be specified for $\Gamma_\mathrm{S}(\sigma_n)$ and $\Gamma^s_\mathrm{S}(\sigma^s_n)$, respectively. Therefore all values of Γ are set to unity before initiating the calculation for all intervals of the considered pure molecule or mixture. After the first iteration, the newly calculated Γ will be averaged with the previous values and the differences between the averaged and former values $\Delta\Gamma$ will serve as convergence crite-

ria. Only when $\Delta\Gamma$ of every interval reaches the convergence criterion (here, that the maximum absolute difference between values of Γ is less than 10^{-8}), the successive substitution will be terminated, otherwise the iteration starts again with the averaged Γ substituting the previous initial values.

Furthermore, the equations can be rewritten to remove the evaluation of the logarithm, as the evaluation of log(exp(x)) is computationally much more expensive than division. Therefore, a slightly more efficient implementation (here, demonstrated for the case of the mixture segment activity coefficients with one sigma profile) is

$$\Gamma_{S,\mathrm{new}} = \left\{ \sum_{\sigma_n} \mathbf{A}^{(+)} \Gamma_{\mathrm{S,old}}\left(\sigma_n\right) \right\}^{-1} . \qquad (42)$$

In the present C++ code, the sum on the right hand side is carried out in a vectorized form with matrix-vector operations from the Eigen library. Furthermore, the matrix

$$\mathbf{A}^{(+)} = \exp\left[-\frac{\Delta W\left(\sigma_m, \sigma_n\right)}{RT} \right] p_{\mathrm{S}}\left(\sigma_n\right) \qquad (43)$$

can be precalculated, as it does not depend on the current value of $\Gamma_{S,\mathrm{old}}$. This operation can be carried out by multiplying each row of the matrix $\exp(-\Delta W/(RT))$ by $p_{\mathrm{S}}\left(\sigma_n\right)$ in a coefficient-wise sense.

In the code by Mullins et al.,[34] the sum in Eq. (42) contains all elements of the sigma profile, but this is not necessary as charge densities which do not exist in a molecule do not contribute to the sum. Especially in the case of relatively nonpolar molecules (e.g., the alkanes), only a small fraction of the range of σ is populated. Therefore, it is necessary to, at the time of loading the model, determine the range of σ that is found in any molecule. The range of σ is obtained by considering the non-hydrogen-bonding profile of an equimolar mixture of components. It is not necessary to consider the OT or OH profiles because the NHB will at least have a small contribution from each of the other profiles, according to Eq. (10). The minimum and maximum values of σ with a contribution $p(\sigma)A$ greater than zero are retained, and only these elements are evaluated.

5 Vapor-Liquid Equilibrium Calculations

Pressure-composition diagrams are created by calculating boiling and dew pressures for a selection of compositions to generate the complete boiling and dew point curves. In order to calculate phase equilibria, the fugacity of the vapour phase f_i^{vap} for each component i has to be equal to the fugacity of the liquid phase f_i^{liq} of the same component in a mixture with given mole fractions x_i

$$f_i^{\mathrm{vap}} = f_i^{\mathrm{liq}}. \qquad (44)$$

The fugacity is defined as $f_i = \varphi_i x_i p$ and the activity coefficient for the liquid phase as $\gamma_i^{\mathrm{liq}} = \varphi_i^{\mathrm{liq}}/\varphi_{i,0}^{\mathrm{liq}}$, so that Eq. (44) becomes

$$\varphi_i^{\mathrm{vap}}\ x_i^{\mathrm{vap}}\ p = \gamma_i^{\mathrm{liq}}\ x_i^{\mathrm{liq}}\ \varphi_{i,0}^{\mathrm{liq}}\ p\ , \qquad (45)$$

with $\varphi_{i,0}^{\mathrm{liq}}$ being the fugacity coefficient of the liquid phase of pure component i, φ_i^{vap} the fugacity coefficient of the vapor phase of component i in the mixture and γ_i^{liq} the activity coefficient of the liquid phase of component i. The first assumption is $\varphi_i^{\mathrm{vap}} = 1$ for the vapor phase as it is assumed to be an ideal gas. The second assumption is $\varphi_{i,0}^{\mathrm{liq}}(T)$ being independent of pressure p. Equating the fugacity of pure component i in the liquid phase $f_{i,0}(T,p) = \varphi_{i,0}^{\mathrm{liq}}\ p$ to that of the pure fluid at saturation, and then also to that of the vapor phase

$$f_{i,0}(T,p) = p\varphi_{i,0}^{\mathrm{liq}} \approx p_{\mathrm{sat},i}\varphi_{i,0,\mathrm{sat}}^{\mathrm{liq}} \approx p_{\mathrm{sat},i}\cancelto{1}{\varphi_{i,0,\mathrm{sat}}^{\mathrm{vap}}}, \qquad (46)$$

$p\varphi_{i,0}^{\mathrm{liq}}$ becomes the vapor pressure $p_{\mathrm{sat},i}$ of the pure component i. Equation (45) is now given by

$$x_i^{\mathrm{vap}}\ p = \gamma_i^{\mathrm{liq}}\ x_i^{\mathrm{liq}}\ p_{\mathrm{sat},i}. \qquad (47)$$

For a binary mixture this leads to

$$x_1^{\text{vap}}\ p = \gamma_1^{\text{liq}}\ x_1^{\text{liq}}\ p_{\text{sat},1}, \tag{48}$$
$$x_2^{\text{vap}}\ p = \gamma_2^{\text{liq}}\ x_2^{\text{liq}}\ p_{\text{sat},2}. \tag{49}$$

Their sum is equal to

$$p = \gamma_1^{\text{liq}}\ x_1^{\text{liq}}\ p_{\text{sat},1} + \gamma_2^{\text{liq}}\ x_2^{\text{liq}}\ p_{\text{sat},2}, \tag{50}$$

so that the equilibrium pressure p can be directly calculated once the activity coefficients γ_1^{liq} and γ_2^{liq} are known. Then, the mole fraction of component i in the vapor phase is given by

$$x_i^{\text{vap}} = \frac{\gamma_i^{\text{liq}}\ x_i^{\text{liq}}\ p_{\text{sat},i}}{p}. \tag{51}$$

The same process is repeated for a range of mixture compositions to create the boiling and dew point curves. An example of an isothermal phase-equilibrium calculation is presented in Fig. 5. Two isotherms are plotted for the mixture ethanol + water, overlaid with experimental data. The code used to generate the figure is in the jupyter notebook `COSMO-SAC.ipynb` in the repository (and in the archive).

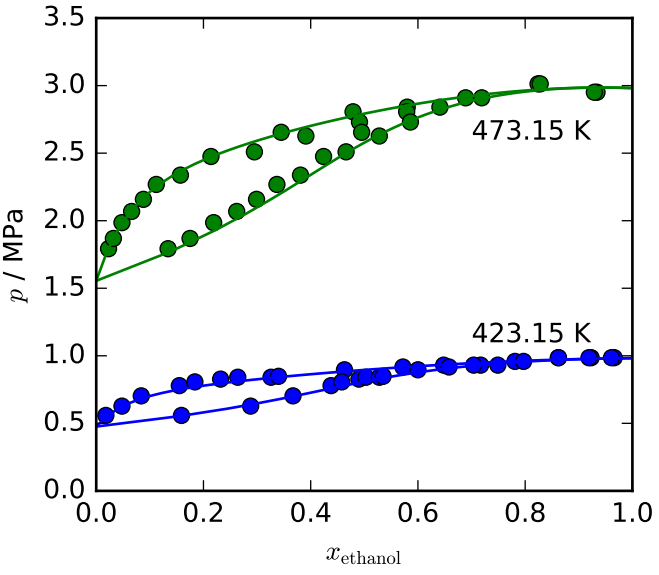

Figure 5: A p-x diagram for the mixture ethanol + water. The experimental data are from Barr-David and Dodge,[48] and the vapor pressure of the pure components are obtained from the ancillary equations of the respective fluid[49]

6 Code and Validation

The development code including the sigma profiles and COSMO-SAC post-processing is contained in a git repository at `https://github.com/usnistgov/COSMOSAC`. The archival version of the code used in this paper is furthermore stored at the DOI of `https://doi.org/10.5281/zenodo.3669311`. Permission from BioVia was obtained to make the .cosmo files available for academic and non-commercial use. Additional information about the .cosmo file database is given in a README file in the file `profiles/UD/README.txt` relative to the root of the code.

The code workflow mirrors the analysis described in this paper. As a pre-processing step, the sigma profiles are generated from each of the .cosmo files according to the Python script in `profiles/to_sigma.py`. This script has a command-line interface that allows for selection of the charge averaging scheme, how many contributions the sigma profiles should be subdivided into (1 or 3), and from this script a single .sigma profile is generated.

Once the sigma profile has been generated, the COSMO-SAC analysis is applied. This code allows for the calculation of the activity coefficients, among other outputs. Either the C++ or Python interfaces may be used, according to the user's preference. A wide range of other numerical analysis programming environments now support calling Python in a nearly-native fashion, so between C++ and Python most users should be able to find a way to call the COSMO code.

Examples of the use of the COSMO-SAC implementation are provided in jupyter notebooks provided with the code, along with the calculation of phase equilibrium calculations, for which the saturation pressure curves for the pure fluids provided by CoolProp[49] are used. A limitation of this method is that vapor pressure curves must be available for the given fluid. Alternatively, vapor pressure curves could be obtained with the consistent alpha function parameters for the Peng-Robinson equation of state of Bell et al.[50]

Further verification is provided by a large

set of calculated values from our model. Users should first ensure that they can precisely regenerate these values prior to making use of the library. The script that generates the verification data is in the file `profiles/generate_validation_data.py` relative to the root of the code.

7 Conclusions

Since Lin and Sandler[10] proposed the original COSMO-SAC model, many modifications and improvements for this model have been proposed. The reproduction of COSMO-SAC models from the literature is often challenging, because on the one hand the model results strongly depend on the sigma profiles, which themselves depend on the program used to calculate them. Therefore, it is crucial to use the same sigma profiles as the authors of the COSMO-SAC model in order to reproduce their model. On the other hand, some misunderstandings regarding the description of COSMO-SAC models exist in the literature, which further complicate the reimplementation of COSMO-SAC models. In this work, we provide an open source C++ and python implementation of three different COSMO-SAC models[10,23,24], together with a detailed documentation of the implemented models. Furthermore, we provide a consistent set of sigma profiles calculated with the software DMol^3 based on the database provided by Mullins et al.[34]. The corresponding COSMO output files and computer code to calculate the sigma-profiles from the COSMO output files is also provided. Thus, this work intends to provide an open-source reference implementation of state-of-the-art COSMO-SAC models.

Acknowledgement IB would like to thank the German Excellence Initiative for funding the research stay at TU Dresden.

References

(1) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. *AIChE J.* **1975**, *21*, 1086–1099.

(2) Fredenslund, A.; Gmehling, J.; Michelsen, M. L.; Rasmussen, P.; Prausnitz, J. M. Computerized Design of Multicomponent Distillation Columns Using the UNIFAC Group Contribution Method for Calculation of Activity Coefficients. *Ind. Eng. Chem. Process Des. Dev.* **1977**, *16*, 450–462.

(3) Gmehling, J.; Li, J.; Schiller, M. A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties. *Ind. Eng. Chem. Res.* **1993**, *32*, 178–193.

(4) Gmehling, J.; Wittig, R.; Lohmann, J.; Joh, R. A Modified UNIFAC (Dortmund) Model. 4. Revision and Extension. *Ind. Eng. Chem. Res.* **2002**, *41*, 1678–1688.

(5) Hansen, H. K.; Rasmussen, P.; Fredenslund, A.; Schiller, M.; Gmehling, J. Vapor-Liquid Equilibria by UNIFAC Group Contribution. 5. Revision and Extension. *Ind. Eng. Chem. Res.* **1991**, *30*, 2355–2358.

(6) Weidlich, U.; Gmehling, J. A modified UNIFAC Model. 1. Prediction of VLE, h^{E}, and y^{∞}. *Ind. Eng. Chem. Res.* **1987**, *26*, 1372–1381.

(7) Klamt, A. Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. *J. Phys. Chem.* **1995**, *99*, 2224–2235.

(8) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. Refinement and parametrization of COSMO-RS. *J. Phys. Chem. A* **1998**, *102*, 5074–5085.

(9) Klamt, A.; Eckert, F. COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids. *Fluid Phase Equilib.* **2000**, *172*, 43–72.

(10) Lin, S.-T.; Sandler, S. I. A Priori Phase Equilibrium Prediction from a Segment Contribution Solvation Model. *Ind. Eng. Chem. Res.* **2002**, *41*, 899–913.

(11) Lee, M.-T.; Lin, S.-T. Prediction of mixture vapor–liquid equilibrium from the combined use of Peng–Robinson equation of state and COSMO-SAC activity coefficient model through the Wong–Sandler mixing rule. *Fluid Phase Equilib.* **2007**, *254*, 28–34.

(12) Hsieh, C.-M.; Lin, S.-T. Determination of cubic equation of state parameters for pure fluids from first principle solvation calculations. *AIChE J.* **2008**, *54*, 2174–2181.

(13) Hsieh, C.-M.; Lin, S.-T. First-Principles Predictions of Vapor-Liquid Equilibria for Pure and Mixture Fluids from the Combined Use of Cubic Equations of State and Solvation Calculations. *Ind. Eng. Chem. Res.* **2009**, *48*, 3197–3205.

(14) Wang, L.-H.; Hsieh, C.-M.; Lin, S.-T. Improved Prediction of Vapor Pressure for Pure Liquids and Solids from the PR+COSMOSAC Equation of State. *Ind. Eng. Chem. Res.* **2015**, *54*, 10115–10125.

(15) Wang, L.-H.; Hsieh, C.-M.; Lin, S.-T. Prediction of Gas and Liquid Solubility in Organic Polymers Based on the PR+COSMOSAC Equation of State. *Ind. Eng. Chem. Res.* **2018**, *57*, 10628–10639.

(16) Hsieh, C.-M.; Lin, S.-T. First-Principles Prediction of Vapor-Liquid-Liquid Equilibrium from the PR+COSMOSAC Equation of State. *Ind. Eng. Chem. Res.* **2011**, *50*, 1496–1503.

(17) Silveira, C. L.; Sandler, S. I. Extending the range of COSMO-SAC to high temperatures and high pressures. *AIChE Journal* **2017**, *64*, 1806–1813.

(18) Jäger, A.; Bell, I. H.; Breitkopf, C. A theoretically based departure function for multi-fluid mixture models. *Fluid Phase Equilib.* **2018**, *469*, 56–69.

(19) Jäger, A.; Mickoleit, E.; Breitkopf, C. A Combination of Multi-Fluid Mixture Models with COSMO-SAC. *Fluid Phase Equilib.* **2018**, *476*, 147–156.

(20) Jäger, A.; Mickoleit, E.; Breitkopf, C. Accurate and predictive mixture models applied to mixtures with CO_2. Proceedings 3rd European supercritical CO2 conference, Paris, France. 2019.

(21) Shimoyama, Y.; Iwai, Y. Development of activity coefficient model based on COSMO method for prediction of solubilities of solid solutes in supercritical carbon dioxide. *J. Supercritic. Fluid.* **2009**, *50*, 210–217.

(22) de P. Soares, R.; Baladão, L. F.; Staudt, P. B. A pairwise surface contact equation of state: COSMO-SAC-Phi. *Fluid Phase Equilib.* **2019**, *488*, 13–26.

(23) Hsieh, C.-M.; Sandler, S. I.; Lin, S.-T. Improvements of COSMO-SAC for vapor–liquid and liquid–liquid equilibrium predictions. *Fluid Phase Equilib.* **2010**, *297*, 90–97.

(24) Hsieh, C.-M.; Lin, S.-T.; Vrabec, J. Considering the dispersive interactions in the COSMO-SAC model for more accurate predictions of fluid phase behavior. *Fluid Phase Equilib.* **2014**, *367*, 109–116.

(25) Fingerhut, R.; Chen, W.-L.; Schedemann, A.; Cordes, W.; Rarey, J.; Hsieh, C.-M.; Vrabec, J.; Lin, S.-T. Comprehensive Assessment of COSMO-SAC Models for Predictions of Fluid-Phase Equilibria. *Ind. Eng. Chem. Res.* **2017**, *56*, 9868–9884.

(26) Kuo, Y.-C.; Hsu, C.-C.; Lin, S.-T. Prediction of Phase Behaviors of Polymer–Solvent Mixtures from the COSMO-SAC Activity Coefficient Model. *Ind. Eng. Chem. Res.* **2013**, *52*, 13505–13515.

(27) Hsieh, M.-T.; Lin, S.-T. A predictive model for the excess gibbs free energy of fully dissociated electrolyte solutions. *AIChE Journal* **2010**, *57*, 1061–1074.

(28) Wang, S.; Song, Y.; Chen, C.-C. Extension of COSMO-SAC Solvation Model for Electrolytes. *Ind. Eng. Chem. Res.* **2011**, *50*, 176–187.

(29) Lee, B.-S.; Lin, S.-T. A Priori Prediction of Dissociation Phenomena and Phase Behaviors of Ionic Liquids. *Ind. Eng. Chem. Res.* **2015**, *54*, 9005–9012.

(30) Lee, B.-S.; Lin, S.-T. Prediction of phase behaviors of ionic liquids over a wide range of conditions. *Fluid Phase Equilib.* **2013**, *356*, 309–320.

(31) Mohr, P. J.; Newell, D. B.; Taylor, B. N. CODATA Recommended Values of the Fundamental Physical Constants: 2014. *J. Phys. Chem. Ref. Data* **2016**, *45*, 043102.

(32) Cordero, B.; Gómez, V.; Platero-Prats, A. E.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S. Covalent radii revisited. *Dalton Trans.* **2008**, 2832–2838.

(33) Xiong, R.; Sandler, S. I.; Burnett, R. I. An Improvement to COSMO-SAC for Predicting Thermodynamic Properties. *Ind. Eng. Chem. Res.* **2014**, *53*, 8265–8278.

(34) Mullins, E.; Oldland, R.; Liu, Y. A.; Wang, S.; Sandler, S. I.; Chen, C.-C.; Zwolak, M.; Seavey, K. C. Sigma-Profile Database for Using COSMO-Based Thermodynamic Methods. *Ind. Eng. Chem. Res.* **2006**, *45*, 4389–4415.

(35) Lin, S.-T.; Sandler, S. I. A Priori Phase Equilibrium Prediction from a Segment Contribution Solvation Model. - Additions and Corrections. *Ind. Eng. Chem. Res.* **2004**, *43*, 1322–1322.

(36) Mu, T.; Rarey, J.; Gmehling, J. Performance of COSMO-RS with Sigma Profiles from Different Model Chemistries. *Ind. Eng. Chem. Res.* **2007**, *46*, 6612–6629.

(37) Lin, S.-T.; Chang, J.; Wang, S.; Goddard, W. A.; Sandler, S. I. Prediction of Vapor Pressures and Enthalpies of Vaporization Using a COSMO Solvation Model. *J. Phys. Chem. A* **2004**, *108*, 7429–7439.

(38) Wang, S.; Sandler, S. I.; Chen, C.-C. Refinement of COSMO–SAC and the Applications. *Ind. Eng. Chem. Res.* **2007**, *46*, 7275–7288.

(39) Mullins, E.; Liu, Y. A.; Ghaderi, A.; Fast, S. D. Sigma Profile Database for Predicting Solid Solubility in Pure and Mixed Solvent Mixtures for Organic Pharmacological Compounds with COSMO-Based Thermodynamic Methods. *Ind. Eng. Chem. Res.* **2008**, *47*, 1707–1725.

(40) Hsieh, C.-M.; Lin, S.-T.; Vrabec, J. Corrigendum to: Considering the dispersive interactions in the COSMO-SAC model for more accurate predictions of fluid phase behavior [Fluid Phase Equilib. 367 (2014) 109-116]. *Fluid Phase Equilib.* **2014**, *384*, 14–15.

(41) Mohr, P. J.; Newell, D. B.; Taylor, B. N.; Tiesinga, E. Data and analysis for the CODATA 2017 special fundamental constants adjustment. *Metrologia* **2018**, *55*, 125–146.

(42) Chen, W.-L.; Hsieh, C.-M.; Yang, L.; Hsu, C.-C.; Lin, S.-T. A Critical Evaluation on the Performance of COSMO-SAC Models for Vapor–Liquid and Liquid–Liquid Equilibrium Predictions Based on Different Quantum Chemical Calculations. *Ind. Eng. Chem. Res.* **2016**, *55*, 9312–9322.

(43) Ferrarini, F.; Flôres, G. B.; Muniz, A. R.; de Soares, R. P. An open and extensible sigma-profile database for COSMO-based models. *AIChE Journal* **2018**, *64*, 3443–3455.

(44) Soares, R. D. P.; Flôres, G. B.; Xavier, V. B.; Pelisser, E. N.; Fabrício Ferrarini,; Staudt, P. B. lvpp/sigma:

LVPP sigma-profile database (18.07); doi: 10.5281/ZENODO.3613786. 2017.

(45) Gerber, R. P.; Soares, R. P. Assessing the reliability of predictive activity coefficient models for molecules consisting of several functional groups. *Braz. J. Chem. Eng.* **2013**, *30*, 1–11.

(46) Wang, S.; Lin, S.-T.; Watanasiri, S.; Chen, C.-C. Use of GAMESS/COSMO program in support of COSMO-SAC model applications in phase equilibrium prediction calculations. *Fluid Phase Equilib.* **2009**, *276*, 37–45.

(47) Possani, L. F. K.; de P. Soares, R. Numerical and Computational Aspects of COSMO-based Activity Coefficient Models. *Braz. J. Chem. Eng.* **2019**, *36*, 587–598.

(48) Barr-David, F.; Dodge, B. F. Vapor-Liquid Equilibrium at High Pressures. The Systems Ethanol-Water and 2-Propanol-Water. *J. Chem. Eng. Data* **1959**, *4*, 107–121.

(49) Bell, I. H.; Wronski, J.; Quoilin, S.; Lemort, V. Pure and Pseudo-pure Fluid Thermophysical Property Evaluation and the Open-Source Thermophysical Property Library CoolProp. *Ind. Eng. Chem. Res.* **2014**, *53*, 2498–2508.

(50) Bell, I. H.; Satyro, M.; Lemmon, E. W. Consistent Twu Parameters for More than 2500 Pure Fluids from Critically Evaluated Experimental Data. *J. Chem. Eng. Data* **2018**, *63*, 2402–2409.

Graphical TOC Entry