I&EC research
Industrial & Engineering Chemistry Research

Article

A Comprehensive Assessment of COSMO-SAC Models for Predictions of Fluid Phase Equilibria

Robin Fingerhut, Wei-Lin Chen, Andre Schedemann, Wilfried Cordes, Juergen Rarey, Chieh-Ming Hsieh, Jadran Vrabec, and Shiang-Tai Lin

Ind. Eng. Chem. Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.iecr.7b01360 • Publication Date (Web): 25 Jul 2017

ACS Publications

Industrial & Engineering Chemistry Research is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

A Comprehensive Assessment of COSMO-SAC Models for Predictions of Fluid Phase Equilibria

Robin Fingerhut[a], Wei-Lin Chen[b], Andre Schedemann[c], Wilfried Cordes[c], Jürgen Rarey[c, d], Chieh-Ming Hsieh[e], Jadran Vrabec[a, *] and Shiang-Tai Lin[b, *]

[a] Thermodynamics and Energy Technology, University of Paderborn, 33098 Paderborn, Germany, *E-Mail: jadran.vrabec@upb.de,

[b] Department of Chemical Engineering, National Taiwan University, 10617 Taipei City, Taiwan, *E-Mail: stlin@ntu.edu.tw,

[c] DDBST GmbH, 26129 Oldenburg, Germany,

[d] Carl-von Ossietzky University Oldenburg, 26129 Oldenburg, Germany,

[e] Department of Chemical & Materials Engineering, National Central University, 320 Taoyuan City, Taiwan,

KEYWORDS. Vapor-liquid equilibria, infinite dilution activity coefficient, COSMO-SAC, UNIFAC

ABSTRACT: Two recent and fully open source COSMO-SAC models are assessed for the first time on the basis of very large experimental data sets. The model performance of COSMO-SAC 2010 and COSMO-SAC-dsp (2013) is studied for vapor-liquid equilibrium (VLE) and infinite dilution activity coefficient (γ_i^∞) predictions, and it is benchmarked with respect to the group contribution models UNIFAC and mod. UNIFAC(DO). For this purpose, binary mixture combinations of 2,295 components are investigated. This leads to 10,897 γ_i^∞ and 6,940 VLE mixtures, which corresponds to 29,173 γ_i^∞ and 139,921 VLE data points. The model performance is analyzed in terms of chemical families. A MATLAB program is provided for the interested reader to study the models in detail. The comprehensive assessment shows that there is a clear improvement from COSMO-SAC 2010 to COSMO-SAC-dsp and from UNIFAC to mod. UNIFAC(DO). The mean absolute deviation of γ_i^∞ predictions is reduced from 95 % to 86 % (COSMO-SAC 2010 to COSMO-SAC-dsp) and from 73 % to 58 % (UNIFAC to mod. UNIFAC(DO)). A combined mean absolute deviation is introduced to study the temperature, pressure and vapor mole fraction errors of VLE predictions, and it is reduced from 4.77 % to 4.63 % (COSMO-SAC 2010 to COSMO-SAC-dsp) and from 4.47 % to 3.51 % (UNIFAC to mod. UNIFAC(DO)). Detailed error analyses show that the accuracy of COSMO-SAC models manly depends on chemical family types, but not on the molecular size asymmetry or polarity. The present results may serve as a reference for the reliability of predictions with COSMO-SAC methods and provide directions for future developments.

1. Introduction

Thermodynamic phase equilibrium properties are of crucial importance for process engineering applications. Process design and optimization are only possible with a sufficiently accurate knowledge of thermodynamic properties [1], which are traditionally determined by experiments. Despite longstanding efforts, there is only a very limited database available because experiments are costly, time-consuming or not feasible due to safety aspects [2]. Furthermore, it is impossible to carry out experimental studies for all technically relevant mixtures. In addition, the prediction of phase equilibrium properties is still a broadly open issue. Thus, suitable computational methods are required, which become more accessible with progress in computing power and molecular models [3,4]. One advantage of computational methods over experimental measurements is that there are no limitations due to difficult conditions (e.g. temperature, pressure, toxicity, component stability etc.).

The chemical potential μ_i, as defined by the Gibbs fundamental equation, is the main thermodynamic property in the context of phase equilibria. Since it is the driving force for all phase conversions, it is of central importance. In the liquid phase, the deviation from ideal solution behavior in terms of the chemical potential is given by the activity coefficient. Conventionally, the prediction of activity coefficient data is being made by group contribution methods, such as UNIFAC (universal quasi-chemical functional group activity coefficients) [5-8] and mod. UNIFAC(DO) [9-11]. Activity coefficients estimated by PSRK [12] and VTPR [13] are used in equations of state mixing rules and are then also applicable to supercritical mixtures. Therein, molecules are treated as collections of independent functional groups and mixtures are built up from these groups. The activity coefficient is calculated on the basis of the sum of activity coefficients of the constituent functional groups. It is thus vital to determine the interactions between these groups,

which were parameterized to a large collection of experimental data.

Recently, a quantum chemistry (QC) based thermodynamic equilibrium method, known as COSMO-RS (conductor like screening model for real solvents), was developed by Klamt [14-17]. It presents a remarkable advance in the prediction of fluid phase equilibria because it contains only a few universal, species-independent parameters and does not require experimental data as an input. COSMO-RS divides the molecular surface area into segments and their activity coefficients are determined from screening charges obtained from quantum chemical solvation calculations, assuming that the solvent is a perfect conductor. The activity coefficient of each molecular species is then computed from those of the segments [18-21]. Despite its success, there are several concerns regarding COSMO-RS that promoted the development of other COSMO-based models (such as COSMO-SAC [18-21], COSMO-RS(Ol) [22] or COSMO-vac [23]). E.g., the initial COSMO-RS model fails to satisfy thermodynamic consistency relations (Gibbs-Duhem) [18], which was corrected in the COSMO-SAC (segment activity coefficient) model of Lin and Sandler [18]. Furthermore, the COSMO-RS model is a commercial product and not all calculation details are published. This makes it impossible for others to independently test and further develop this method.

COSMO-based methods are fully predictive and applicable to almost any fluid mixture, but they are not yet as accurate as group contribution methods. Therefore, UNIFAC and mod. UNIFAC(DO) methods remain widely applied in the field of process engineering [5-11]. Nonetheless, due to their strictly predictive character, COSMO-based models are promising candidates to address the scarcity of phase equilibrium properties. Efforts were undertaken to improve their accuracy for different types of fluid mixtures. E.g., the COSMO-SAC 2010 model (which will be referred to as COSMO-SAC10) [18-20] improves the description of associating fluids by recognizing the differing strengths of hydrogen bonding interactions depending on the type of hydrogen bonding donors and acceptors. The more recent version COSMO-SAC-dsp [21] contains a correction term based on molecular simulation data that takes the dispersive intermolecular interactions explicitly into account.

The purpose of this work is to examine the performance of the fully open source COSMO-SAC models based on the world's largest phase equilibrium database, the Dortmund Data Bank (DDB) [24]. Beforehand, those models were integrated into the DDB software package and were applied to all experimental vapor-liquid equilibrium (VLE) and infinite dilution activity coefficient (γ_i^∞) data sets. In particular, the accuracy of QC-based models (COSMO-SAC10 and COSMO-SAC-dsp) and group contribution models (UNIFAC [5-7] and mod. UNIFAC(DO) revision 5 [9]) was assessed to determine for which mixture types (combinations of chemical families) adequate results are obtained. This also includes the identification of poorly described mixtures because the present results will be the basis for future developments of the COSMO-SAC models. Here, the DDB data collection was utilized with its 48,952 pure components. Of those, only 2,295 components are part of the freely available UD-database (University of Delaware) that collects COSMO-SAC σ-profiles. Nevertheless, this results in 10,897 mixtures for which γ_i^∞ and 6,940 for which VLE data are available. This corresponds to 29,173 γ_i^∞ and 139,921 VLE data points. It should be noted that the COSMO-SAC models were analyzed for the first time on such a large data set. Admittedly, the number of γ_i^∞ and VLE data sets is much smaller than all theoretically possible binary combinations. First, experimental data are not available for each binary combination. Second, only the strict subset of data points which can be calculated with all four COSMO-SAC and UNIFAC methods were considsidered for comparison. Moreover, a quality filter was applied to remove unreliable experimental data. Due to the large number of analyzed data sets, a MATLAB program was created to study the model performance and error distribution for chemical main-families and their sub-families and the source code of this program is accessible to the interested reader (cf. Supporting Information).

This work should serve as an orientation for users, e.g. process designers, who are typically not faced with a lack of models. Instead, numerous models exist for which it is unknown whether they yield adequate results for a given mixture. The MATLAB program provided as Supporting Information to this work sheds light on that aspect, considering the largest possible experimental database in a fair way.

2. COSMO-SAC models

In this study, two COSMO-SAC models, i.e. COSMO-SAC10 [20] and COSMO-SAC-dsp [21], were considered. The only difference between them is that the dispersive contribution to the activity coefficient is explicitly taken into account in the COSMO-SAC-dsp model. Therein, the activity coefficient of component *i* in mixture *S* is determined by

$$\ln\gamma_{i/S} = \ln\gamma_{i/S}^{res} + \ln\gamma_{i/S}^{comb} + \ln\gamma_{i/S}^{dsp} \quad (1)$$

where the superscripts *res*, *comb*, and *dsp* indicate the residual, combinatorial, and dispersion contributions, respectively. The details of these three contributions are briefly summarized in the following and all parameter values are listed in Table 1.

The residual contribution is the key element in the COSMO-SAC models, which considers the permanent electrostatic interactions between molecules in the mixture. Such interactions are determined on the basis of the molecular surface screening charges obtained from QC and COSMO solvation calculations [15]. Since the QC/COSMO calculation is the most time-consuming step, that fortunately needs to be carried out only once for every molecular species, it is useful to collect its results in a database (e.g. VT-database [25,26]) for subsequent calculations of thermophysical properties and phase behavior [27-34]. The surface charge distribution of molecule *i* from the QC/COSMO calculation is averaged through a semi-theoretical equation [18] and then used to generate the σ-profile $p_i(\sigma_m)$, i.e. the probability of finding a surface segment with charge density σ_m on molecule *i*. The σ-profile is also known as molecular surface shielding charge density distribution and is unique for every molecule. In order to better describe hydrogen bonding interactions, the molecular surface segments are categorized into three types: nhb (non-hydrogen-bonding surface segments), OH (surface segments on the hydroxyl group), and OT (surface segments on all other hydrogen bonding atoms, i.e. F, O, N, and H bonded to N and F). Consequently, the σ-profile is also separated into three contributions: $p_i(\sigma) = p_i^{\mathrm{nhb}}(\sigma) + p_i^{\mathrm{OH}}(\sigma) + p_i^{\mathrm{OT}}(\sigma)$ [20]. Both $p_i^{\mathrm{OH}}(\sigma)$ and $p_i^{\mathrm{OT}}(\sigma)$ are hydrogen bonding σ-profiles and the probability of hydrogen bonding segments in forming a hydrogen bond is considered by a Gaussian-type function, $p^{HB}(\sigma) = 1 - \exp(\sigma^2/2\sigma_0^2)$ with σ_0 = 0.007 e/Å^2 [19]. The σ-profile of mixture *S* is determined from

$$p_S(\sigma) = \frac{\sum_i x_i A_i p_i(\sigma)}{\sum_i x_i A_i} \quad (2)$$

where A_i and x_i are molecular surface area and mole fraction of component *i*. Once the σ-profiles of all components in the mixture and that of the mixture are established, the segment activity coefficient of a segment with charge density σ_m can be calculated by

$$\ln\Gamma_j^t(\sigma_m^t) = -\left\{\sum_t^{\mathrm{nhb,OH,OT}} \sum_{\sigma_n} p_j^t(\sigma_n^t)\Gamma_j^t(\sigma_n^s)\exp\left[\frac{-\Delta W(\sigma_n^s,\sigma_m^t)}{RT}\right]\right\} \quad (3)$$

where the subscript *j* can denote either the pure component *i* or the mixture *S* and the segment exchange energy ΔW is calculated from the charge density of the interacting segments

$$\Delta W(\sigma_n^s,\sigma_m^t) = \left(A_{ES} + \frac{B_{ES}}{T^2}\right)(\sigma_n^s + \sigma_m^t)^2 - c_{hb}(\sigma_n^s,\sigma_m^t)(\sigma_n^s - \sigma_m^t)^2 \quad (4)$$

where A_{ES} and B_{ES} are electrostatic interaction parameters and c_{hb} is the hydrogen bonding interaction parameter. The first term on the right-hand side considers the general electrostatic interaction between segments and the other term accounts for additional interactions between hydrogen bonding segments. The hydrogen bonding interaction coefficients between different kinds of segment combinations are given by

$$c_{hb}(\sigma_n^s,\sigma_m^t) = \begin{cases} c_{\mathrm{OH}} & \text{if } s = t = \mathrm{OH} \text{ and } \sigma_n^s \cdot \sigma_m^t < 0 \\ c_{\mathrm{OT}} & \text{if } s = t = \mathrm{OT} \text{ and } \sigma_n^s \cdot \sigma_m^t < 0 \\ c_{\mathrm{OH-OT}} & \text{if } s = \mathrm{OH}, t = \mathrm{OT} \text{ and } \sigma_n^s \cdot \sigma_m^t < 0 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Finally, the activity coefficient of the residual contribution for component *i* in a mixture *S* can be determined from the σ-profile of component *i* $p_i(\sigma)$ and the segment activity coefficient of component *i* and that of the mixture *S* by

$$\ln\gamma_{i/S}^{res} = \frac{A_i}{a_{\mathrm{eff}}}\sum_t^{\mathrm{nhb,OH,OT}} \sum_{\sigma_m} p_i^t(\sigma_m^t)\cdot[\ln\Gamma_S^t(\sigma_m^t) - \ln\Gamma_i^t(\sigma_m^t)] \quad (6)$$

where A_i and a_{eff} are the molecular surface area and the surface area of a standard surface segment, respectively.

The combinatorial contribution considers the molecular size and shape effects between molecules in the mixture via the Staverman-Guggenheim (SG) combinatorial term [35,36]

$$\ln\gamma_{i/S}^{comb} = \ln\gamma_{i/S}^{SG} = \ln\frac{\phi_i}{x_i} + \frac{z}{2}q_i\ln\frac{\theta_i}{\phi_i} + l_i - \frac{\phi_i}{x_i}\sum_j x_j l_j \quad (7)$$

where z =10 is the coordination number, $\theta_i = x_i q_i / \sum_j x_j q_j$ $\phi_i = x_i r_i / \sum_j x_j r_j$, and $l_i = (z/2)(r_i - q_i) - (r_i - 1)$ with the coordination number z = 10, and q_i and r_i being the normalized surface area and volume of component *i*.

The dispersion contribution to the activity coefficient was proposed according to molecular simulation results of binary Lennard-Jones model mixtures from the literature [3,4,21]. For component *i* in the mixture *S* it is calculated by the two-suffix Margules equation [37]

$$\ln\gamma_{i/S}^{dsp} = \sum_{j\neq i} x_j A_{ij} - \sum_i \sum_{j>i} x_i x_j A_{ij} \quad (8)$$

where $A_{ij} = A_{ji}$ is the binary interaction parameter between component *i* and *j* and is determined from

$$A_{ij} = w \cdot \left[(\varepsilon_i + \varepsilon_j)/2 - \sqrt{\varepsilon_i \varepsilon_j}\right] \quad (9)$$

where *w* is an empirical parameter and ε is the molecular dispersion parameter calculated from $\varepsilon_{\mathrm{Molecule}} = \left(\sum_{k=1}^{N_\mathrm{n}} N_k \varepsilon_k'\right)/N_\mathrm{n}$. N_i and ε_k' are number of atoms from type *k* and its dispersion parameter, respectively, and N_n is the total number of atoms with a non-zero dispersion parameter. For binary mixtures, Equation (8) can be simplified to the one-constant Margules equation.

3. Infinite dilution activity coefficient data

The non-ideality of a mixture is often characterized by the infinite dilution activity coefficient ${\gamma_i}^\infty$ and its knowledge is important for many industrial applications. Here, the performance of COSMO-SAC10 and COSMO-SAC-dsp for ${\gamma_i}^\infty$ predictions was compared to the UNIFAC and mod. UNIFAC(DO) models. Therefore, it is crucial to calculate the errors of each method in predicting ${\gamma_i}^\infty$ for binary mixtures. Equation (10) introduces the mean absolute deviation (MAD) $|\bar{\delta}|$

$$|\bar{\delta}| = \frac{1}{n}\sum_{a=1}^{n}\left|\ln({\gamma_i}^\infty)_a^{cal} - \ln({\gamma_i}^\infty)_a^{exp}\right|. \quad (10)$$

Here, *n* is the number of considered data points and the abbreviations *cal* and *exp* stand for calculated and experimental ${\gamma_i}^\infty$ data. For the present evaluation, 39,349 data points were available from COSMO-SAC10, 35,837 from COSMO-SAC-dsp, 41,470 from UNIFAC and 42,870 from mod. UNIFAC(DO). Considering only the strict subset of data points for which all four models yielded results, the number of ${\gamma_i}^\infty$ data points was reduced to 29,501. Moreover, two quality filters were used to remove unreliable experimental data and all experimental data measured with liquid-liquid chromatography. Thus, the total number of studied ${\gamma_i}^\infty$ data points was 29,173. These data correspond to an experimental temperature range from 213.13 K to 576.15 K in a ${\gamma_i}^\infty$ interval from -2.53 to 26.40 in natural logarithm, i.e. $\ln {\gamma_i}^\infty$.

3.1 Model performance for chemical families

This section discusses the MAD $|\bar{\delta}|$ of each prediction method according to chemical family combinations for binary mixtures. For this purpose, all mixtures were hierarchically categorized into main-family combinations (cf. Table 2 and Supporting Information SI I for the fully detailed family classification) and the MAD of each combination was calculated with a MATLAB program (cf. Supporting Information SI II). Exemplarily, one selected main-family combination is discussed below in detail in terms of the MAD on its sub-family levels. Here, the model performance is presented for COSMO-SAC-dsp and mod. UNIFAC(DO) only. The other models are shown in the Supporting Information.

Figure 1 (top) depicts the MAD of all binary main-family combinations for the COSMO-SAC-dsp model, which reveals similarities to COSMO-SAC10. This type of diagram was created to give an overview on all main-family deviations (only main-families which contain data are presented). The horizontal axis indicates the main-families of all solvents (mentioned first from now on) and the vertical axis presents all solutes, sorted roughly according to ascending polarity. The symbol size represents the amount of experimental data points in one family combination and its color shows the MAD range. A few family combinations were predicted very adequately ($|\bar{\delta}|$< 10.5 %), e.g. Acids + Multifunctionals or Ethers + Esters. The latter main-family combination shows a significant improvement from COSMO-SAC10 to COSMO-SAC-dsp. This enhanced prediction indicates the strength of the COSMO-SAC-dsp development. Ethers have one oxygen atom with two single bonds (-O-) and Esters have one double bond between oxygen and carbon (C=O). Dispersion energy parameters were developed for these two types of bonded oxygen in the COSMO-SAC-dsp version (cf. Table 1 b). Moreover, Figure 1 points out that numerous main-family combinations are predicted well (10.5 % <$|\bar{\delta}|$<

64.9 %), e.g. Alcohols + Alkanes. A few main-family combinations are predicted decently (64.9 % <$|\bar{\delta}|$< 146.0 %), e.g. Multifunctionals + Alkanes and a few are not described well, e.g. Acids + Carbonyls (146.0 % <$|\bar{\delta}|$< 266.9 %) or Alkanes + Alcohols ($|\bar{\delta}|$> 266.9 %). Moreover, mixtures where water is a solute are not predicted adequately, whereas water as a solvent is less problematic. Hence, the following section focuses on the error distribution for all data, where it is distinguished between aqueous data and non-aqueous data sets, indicating whether water is a component of the mixture or not. Furthermore, the improvement for the combinations Amides + Alkanes and Amides + Alkenes is striking, which is due to the COSMO-SAC-dsp development. The main-family Amides contains molecules with an oxygen double bond (=O) and a nitrogen atom (cf. Table 1 b) for which dispersion parameters were introduced. Additionally, parameters for carbon (sp, sp^2, sp^3) were implemented in COSMO-SAC-dsp and all organic molecules contain carbon. The correlation between errors and polarity is not significant as similarly seen for COSMO-SAC10.

Figure 1 (bottom) gives the MAD of all binary main-family combinations for the mod. UNIFAC(DO) model with the 2006 parameter matrix. As expected, this model shows very accurate descriptions in general. As a result, it was chosen as the standard for comparison to both COSMO-SAC models. Some main-family combinations are described very well ($|\bar{\delta}|$< 10.5 %), e.g. Acids + Esters, and many other combinations show good results (10.5 % <$|\bar{\delta}|$< 64.9 %), e.g. Alcohols + Alkanes. There are several combinations which are described decently (64.9 % <$|\bar{\delta}|$< 146.0 %), e.g. Water + Aromatics, and for a few main-family combinations inaccurate descriptions ($|\bar{\delta}|$> 266.9 %) were found, e.g. Water + Alkanes. Moreover, this figure points out the improvement between UNIFAC (cf. Supporting Information) and mod. UNIFAC(DO) for the main-families Alkanes + Alkanes and Alkanes + Alkenes. It is known that aqueous mixtures are challenging for mod. UNIFAC(DO) and UNIFAC, especially Water + Alkanes and Water + Alkenes. The authors of mod. UNIFAC(DO) explain this issue in Refs. [7,9,38]. Moreover, mod. UNIFAC(DO) performs worse than UNIFAC for some aqueous systems, e.g. Water + Esters. Some of these main-family combinations are predicted more adequately by COSMO-SAC10 and COSMO-SAC-dsp, e.g. Water + Alkenes. Again, there is no obvious correlation between errors and polarity.

From Figures 1 and S.1, S.2 (in the Supporting Information) it can be seen that there is an improvement from COSMO-SAC10 to COSMO-SAC-dsp. Those figures show that both COSMO-SAC models are in general more applicable to aqueous mixtures of aliphatic components, particularly when water is the solvent, compared to both UNIFAC models.

Exemplarily, the first sublevel of the main-family combination Alkanes + Alcohols is discussed which contains 966 experimental data points. Here, both COSMO-SAC models yield large MAD ($|\bar{\delta}|$> 266.9 %). An inadequately predicted main-family combination was chosen to show the purpose of a MATLAB program, which was developed in this work and is provided (cf. Supporting Information). With this program, the user may obtain additional information on the error distribution and the MAD of main- and sub-family combinations. Thus, it is possible to determine functional groups which may cause problems in a model. Based on this, the interested reader can check the performance of the four models for any molecular species in detail by himself. Figures S.5 to S.12 (in the Supporting Information) show exemplarily the MAD for the sub-family combinations of Alkanes + Alcohols. It is obvious that both COSMO-SAC models and in some cases also UNIFAC have issues with alcohols as solutes. The prediction of the sub-families xAlcohol_nointra, e.g. 1,4-Butanediol, and xAlcohol_intra, e.g. 1,2-Propanediol, in Alkanes shows large MAD, whereas mod. UNIFAC(DO) gives better predictions (cf. Table 3 for sub-family abbreviations).

3.2 Error distribution

As discussed above, all considered methods are less accurate for aqueous mixtures, especially when water is the solvent. Moreover, we were interested in the error distribution of very asymmetric mixtures, since the COSMO-SAC methods implicate the Stavermann-Guggenheim combinatorial term (cf. equation (7)). For this purpose, the error distribution over the mixtures' asymmetry was investigated. It is necessary to establish a metric for the system's asymmetry. Here, we used the ratio of the molecular surface area of the solute A_{solute} and the solvent $A_{solvent}$ as a measure

$$\alpha_{\mathrm{sym}} = \log\left(\frac{A_{solute}}{A_{solvent}}\right). \quad (11)$$

In case of a perfectly symmetrical mixture, the solute and the solvent have the same surface area, thus $\alpha_{\mathrm{sym}} = 0$.

Figure 2 shows the error distribution over α_{sym} for COSMO-SAC-dsp and mod. UNIFAC(DO). The error distribution of COSMO-SAC10 and UNIFAC is given in the Supporting Information. Data points are depicted with a different color in case of aqueous mixtures. Those graphics were also created by the MATLAB program that is provided in the Supporting Information. They indicate that the errors of non-aqueous mixtures (81.6 % of all data sets) are slightly shifted to negative values (blue points) for all methods. There is a bulge to negative errors for symmetric mixtures (α_{sym}=[-0.2;0.2], i.e., molecular surface area ratio within about $1.6^{\pm1}$). The error distribution is most uniform for mod. UNIFAC(DO), followed by UNIFAC, COSMO-SAC-dsp and COSMO-SAC10. Apart from that, aqueous mixtures (18.4 % of all data sets) show large errors (red and black points) and it is clear that all prediction methods have some issues with those systems. COSMO-SAC10 yields large errors for mixtures which contain water. The errors tend to have positive values, which means that calculated ${\gamma_i}^{\infty}$ are larger than experimental ${\gamma_i}^{\infty}$. COSMO-SAC-dsp exhibits nearly the same behavior for aqueous mixtures, however, the errors are slightly better distributed around zero. It becomes clear that UNIFAC predicts mixtures with water as a solvent incorrectly, whereas it gives better results for water as a solute. Mod. UNIFAC(DO) shows a more uniform error distribution than UNIFAC, but the predictions for aqueous systems are worse than those of UNIFAC. Initially, it was assumed that the error may correlate with the mixtures' asymmetry due to a potential problem in the Stavermann-Guggenheim combinatorial term in both COSMO-SAC models. However, this cannot be confirmed, as revealed by Figures 1, S.1 and S.2, and it is striking that the performance of all methods mainly depends on chemical families.

Figure 3 shows the error distribution of all data sets for each method. It reflects the same behavior as discussed above, however, featuring a different way of presentation to clarify the statistical circumstances. Moreover, the error distribution is compared to Cauchy and normal distribution functions. It can be seen that all methods show an almost Cauchy distributed error,

but both COSMO-SAC versions are closer to a normal distribution than both UNIFAC models. The stronger Cauchy-like character of mod. UNIFAC(DO) implies the presence of heavy outliers, despite the fact that its error distribution is rather uniform around zero.

3.3 Summary

The performance of COSMO-SAC10 and COSMO-SAC-dsp was studied on the basis of γ_i^∞ predictions for a very large experimental data set of 29,173 data points. For this purpose, MATLAB programs were developed. Table 4 summarizes the γ_i^∞ errors (converted to percentages) of all data, where it is distinguished between non-aqueous and aqueous data sets, as discussed before.

The consideration of all data sets shows that there is a clear improvement from COSMO-SAC10 to COSMO-SAC-dsp (the MAD was reduced from 95 % to 86 %) and from UNIFAC to mod. UNIFAC(DO) (the MAD was reduced from 73 % to 58 %). Moreover, the error becomes more uniformly distributed around zero from COSMO-SAC10 to COSMO-SAC-dsp (the mean signed error (MSD) $\bar{\delta}$ was reduced from -17 % to -1 %) and from UNIFAC to mod. UNIFAC(DO) (MSD was reduced from -30 % to -21 %). For all data sets mod. UNIFAC(DO) gives the smallest error (58 %), followed by UNIFAC (73 %), COSMO-SAC-dsp (86 %) and COSMO-SAC10 (95 %). All prediction methods showed an almost Cauchy distributed error, whereas both COSMO-SAC versions are closer to a normal distribution than both UNIFAC versions. In general, UNIFAC methods are more accurate than the COSMO-SAC methods, but they yield heavy outliers for some systems. As a result, COSMO-SAC is closer to a normal distribution than UNIFAC.

Table 4 shows that all methods are more precise for non-aqueous mixtures (81.6 % of all data). The error distribution for non-aqueous data sets is slightly shifted to negative values, most uniformly distributed for mod. UNIFAC(DO), followed by COSMO-SAC-dsp, UNIFAC and COSMO-SAC10. However, there is a significant progress from COSMO-SAC10 (79 %) to COSMO-SAC-dsp (65 %) and from UNIFAC (49 %) to mod. UNIFAC(DO) (27 %). Aqueous data sets (18.4 % of all data) show very large errors, thus, these systems strongly affect the MAD of all data. Mod. UNIFAC(DO) gives the largest MAD (306 %), followed by UNIFAC (232 %), COSMO-SAC-dsp (203 %), and COSMO-SAC10 (194 %). Hence, COSMO-SAC models should preferably be used for aqueous mixtures. The comparison between aqueous and non-aqueous data indicates that all methods perform better for non-aqueous mixtures. Furthermore, COSMO-SAC10 treats these systems less differently, followed by COSMO-SAC-dsp, UNIFAC and mod. UNIFAC(DO). Mod. UNIFAC(DO) treats aqueous vs. non-aqueous systems most differently (as seen from the MAD ratio of aqueous to non-aqueous mixtures).

The error distribution study has shown that very asymmetric non-aqueous mixtures are well predicted by the COSMO-SAC models and a correlation between the mixtures' asymmetry and error is not present. Instead, the performance of all methods mainly depends on chemical families. Therefore, access to the errors of all family combinations is provided. Moreover, it was assumed that the error may correlate with the polarity of different families. Thus, the main-families were arranged roughly according to their increasing polarity. However, a clear correlation cannot be seen.

In the following, a few main-family combinations are discussed, where COSMO-SAC10 and/or COSMO-SAC-dsp are as good as UNIFAC and/or mod. UNIFAC(DO) or even better, to emphasize the predictive power of the COSMO-SAC models. For that purpose, the MAD from Figures 1, S.1 and S.2 were converted to percentage ranges.

• Water + Alkanes: For this family combination COSMO-SAC10 and COSMO-SAC-dsp (64.9 - 146.0 %) are much better than UNIFAC and mod. UNIFAC(DO) ($\geq$ 266.9 %).

• Esters + Multifunctionals: COSMO-SAC10 and COSMO-SAC-dsp give the most accurate results (10.5 - 64.9 %) and they are better than mod. UNIFAC(DO) (64.9 - 146.0 %) and UNIFAC ($\geq$ 266.9 %).

• Acids + Multifunctionals: All methods are very good for this combination. COSMO-SAC10 and COSMO-SAC-dsp ($\leq$ 10.5 %) are as good as UNIFAC ($\leq$ 10.5 %) and better than mod. UNIFAC(DO) (10.5 - 64.9 %).

• Amides + Carbonyls: COSMO-SAC-dsp ($\leq$ 10.5 %) is better than UNIFAC (64.9 - 146.0 %) and mod. UNIFAC(DO) (10.5 - 64.9 %). COSMO-SAC10 (10.5 - 64.9 %) is as good as mod. UNIFAC(DO) and better than UNIFAC.

Furthermore, some family combinations, e.g. Ethers + Esters, emphasize the potential of the COSMO-SAC-dsp development. This method introduced dispersion energy parameters for the atom types C, O, N, F, Cl and H. The main-families Ethers and Esters contain some of those atoms and they illustrate the improvement from COSMO-SAC10 to COSMO-SAC-dsp. In summary, the results indicate that COSMO-SAC-dsp should be applied instead of mod. UNIFAC(DO) for mixtures where no parameters are available for one of the groups or group combinations because of its predictive character. Moreover, the error distribution of mod. UNIFAC(DO) showed some heavy outliers while COSMO-SAC-dsp is closer to a normal distribution. Finally, COSMO-SAC-dsp should be used for aqueous systems and many other mixture types (cf. Figure 1).

4. Vapor-liquid equilibrium data

Information on the two-phase region, where both components coexist in significant quantities in both vapor and liquid, is crucial for many industrial processes [39]. One VLE state point of a binary mixture has a specific temperature *T*, pressure *p*, liquid mole fraction *x* and vapor mole fraction *y*. At least three of these properties have to be defined by experiment or model calculation and there are various property specifications possible, e.g. *T,p,x,y; T,p,x*; *T,p,y*; *T,x,y*; *p,x,y*. The full experimental *T,p,x,y* data set was utilized here, since those present the most reliable data. The data are distinguished into the common isothermal and isobaric combinations *T,x,y* and *p,x,y*.

This section discusses the performance of COSMO-SAC-dsp for VLE predictions compared to mod. UNIFAC(DO). The COSMO-SAC10 and UNIFAC model results are shown in the Supporting Information. To study the performance, it is crucial to calculate the errors for each method. This error calculation differs from the γ_i^∞ study because VLE data contain more properties and thus different error types (e.g. either in *p,y* or *T,y*). To preserve a good overview for the main- and sub-family figures, a combined and weighted MAD was defined. The weight was based on the smallest MAD which was calculated over all isothermal or isobaric data sets, respectively. The MAD $|\bar{\delta}|$ calculation for each VLE property is given by

$$\left|\bar{\delta}\right|_p [\%] = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{p_{cal}-p_{exp}}{p_{exp}}\right|_i \cdot 100, \quad (12)$$

$$\left|\bar{\delta}\right|_y [\%] = \frac{1}{n}\sum_{i=1}^{n}\left|y_{cal}-y_{exp}\right|_i \cdot 100, \quad (13)$$

$$\left|\bar{\delta}\right|_T \text{ [K]} = \frac{1}{n}\sum_{i=1}^{n}\left|T_{cal} - T_{exp}\right|_i, \quad (14)$$

where *n* is the number of data points. The smallest MAD was reached by mod. UNIFAC(DO). Table 5 shows the MAD for all isothermal and isobaric VLE data. These values were used in equations (15) and (16) to determine the weight between pressure and vapor mole fraction errors for the isothermal or temperature and vapor mole fraction errors for the isobaric VLE, respectively. The weight factors *a* and *b* for the isothermal VLE data were calculated by

$$\frac{a}{b} = \frac{|\bar{\delta}|_p}{|\bar{\delta}|_y} = \frac{3.88\ \%}{1.45\ \%} = 2.6759, \quad (15)$$

where the vapor mole fraction weight factor *a* was set to unity and thus the pressure weight factor *b* results in 0.3737. The weight factors *c* and *d* for the isobaric VLE data were determined by

$$\frac{c}{d} = \frac{|\bar{\delta}|_T}{|\bar{\delta}|_y} = \frac{1.37\ \text{K}}{1.91\ \%} = 0.7173\ \text{K}/\%, \quad (16)$$

where the vapor mole fraction weight factor *c* was set to *a* and the temperature weight factor *d* results in 1.3941 % / K. On the basis of these weight factors, the combined mean absolute deviation (CMAD) $|\bar{\Delta}|$ for VLE data is given by

$$|\bar{\Delta}|[\%] = \begin{cases} a \cdot \left|\bar{\delta}\right|_y + b \cdot \left|\bar{\delta}\right|_p, & \text{isothermal} \\ c \cdot \left|\bar{\delta}\right|_y + d \cdot \left|\bar{\delta}\right|_T, & \text{isobaric.} \end{cases} \quad (17)$$

Those CMAD $|\bar{\Delta}|$ were used to assess the performance of each model for different chemical family combinations.

For the assessment, 336,291 experimental data points were available for COSMO-SAC10, 316,046 for COSMO-SAC-dsp, 299,078 for UNIFAC and 298,032 for mod. UNIFAC(DO). After filtering for the strict subset of data points which can be calculated with all four considered methods, 268,629 VLE data points remained. In addition, two filters were used to remove the edges of the two phase region ($x, y = 0$ or 1, i.e. pure substance data) and data with a pressure above 1000 kPa. The latter filter was applied because all methods treat the vapor as an ideal gas. As a consequence, 139,921 data points (6,940 mixtures) were evaluated, containing 45,456 isothermal and 94,465 isobaric points. The data can also be divided into 125,888 (90.0 %) non-aqueous and 14,033 (10.0 %) aqueous data points. These are in a temperature range from 233 K to 633.4 K and a pressure range from 0.084 kPa to 1000 kPa.

4.1 Model performance for chemical families

The model performance is discussed here for COSMO-SAC-dsp and mod. UNIFAC(DO) for chemical families. COSMO-SAC10 and UNIFAC are shown in the Supporting Information. All mixtures were again categorized in the same main-family combinations and their CMAD was calculated with a MATLAB program. Figure 4 (top) shows the CMAD in percentage ranges for all main-family combinations for the COSMO-SAC-dsp model, which reveals similarities to COSMO-SAC10. It should be noted that the horizontal axis indicates the main-family of the low boiling component and the vertical axis shows the main-family of the high boiling component of a given mixture. A few main-family combinations were predicted very adequately ($|\bar{\Delta}|$< 1 %), e.g. (Iso)Nitriles + (Iso)Nitriles. Many combinations show accurate results (1 % < $|\bar{\Delta}|$ < 3 %), e.g. Esters + Aromatics or Amides + Carbonyls. Moreover, these combinations exemplify improvements from COSMO-SAC10 to COSMO-SAC-dsp. This again shows the potential of the COSMO-SAC-dsp development. Esters have one oxygen atom with a double bond to carbon (C=O) as Amides and Carbonyls have. Furthermore, Amides contain bonds between N and H. Exactly for those atoms, dispersion energy parameters were developed in the COSMO-SAC-dsp version (cf. Table 1 b). Several main-family combinations are predicted decently (3 % <$|\bar{\Delta}|$< 5 %), e.g. Alcohols + Multifunctionals, and some are not described well (5 % <$|\bar{\Delta}|$< 10 %), e.g. Acids + Esters. The number of poorly predicted ($|\bar{\Delta}|$> 10 %) main-family combinations decreases from COSMO-SAC10 to COSMO-SAC-dsp. A few main-family combinations show inconclusive results ($|\bar{\Delta}|$> 10 %) (many aqueous mixtures), e.g. Water + Alcohols. Hence, the next section will focus on the error distribution for all data, distinguished between aqueous and non-aqueous data. It is clear that there is no obvious correlation between the CMAD and polarity.

Figure 4 (bottom) gives the CMAD of all main-family combinations for the mod. UNIFAC(DO) model with the 2006 parameter matrix. As expected, mod. UNIFAC(DO) shows accurate descriptions for numerous main-family combinations, e.g. Amines + Esters ($|\bar{\Delta}|$< 1 %) or Aromatics + Carbonyls (1 % < $|\bar{\Delta}|$ < 3 %). Due to this reliable behavior, it was chosen as the standard to compare both COSMO-SAC models. Moreover, this diagram points out the improvement from UNIFAC (cf. Supporting Information) to mod. UNIFAC(DO) in general because the number of well described ($|\bar{\Delta}|$< 1 %) main-family combinations increased. There are still a few decently described (3 % <$|\bar{\Delta}|$< 5 %) combinations, e.g. Alcohols + Carbonyls, and it can be seen that mod. UNIFAC(DO) has problems ($|\bar{\Delta}|$> 10 %) with aqueous mixtures, especially with Water + Alcohols and Water + Multifunctionals (as UNIFAC does). The authors of mod. UNIFAC(DO) explained this issue in Refs. [7,9,38]. Some family combinations are more adequately predicted by COSMO-SAC10 and COSMO-SAC-dsp, e.g. Ethers + Esters or Alkenes + Amines.

Figures 4 and S.3 show that there is an improvement from COSMO-SAC10 to COSMO-SAC-dsp. Furthermore, Figures 4, S.3 and S.4 show that both COSMO-SAC models are more suited for some mixture types, e.g. Ethers + Esters, than both UNIFAC models. A presentation on sub-family levels is omitted here, however, the interested reader can assess the model performance in detail himself by using the provided MATLAB program.

4.2 Error distribution

The error distribution over the molecular surface area ratio α_{sym} was studied following Equation (11). Because it is not meaningful to show the error distribution using absolute error values, the combined VLE error (CMAD) consideration is not applicable here. As a consequence, the signed error distributions of VLE properties temperature, pressure and vapor mole fraction are more complex and they are given for all models in the Supporting Information.

Figures S.18 to S.20 and S.24 to S.26 show the temperature, pressure and vapor mole fraction error distribution over the asymmetry rate of the COSMO-SAC-dsp and mod. UNIFAC(DO) models. Data points of aqueous mixtures (10.0 % of all data sets) are depicted in red color. The blue symbols show non-aqueous data (90.0 %). For the COSMO-SAC-dsp model, the temperature, pressure and vapor mole fraction errors are all nearly uniformly distributed around zero, although there are considerable errors. The temperature errors tend to exhibit

negative values, whereas pressure and vapor mole fraction errors are slightly shifted towards positive values. Most of the non-aqueous mixtures are symmetric (α_{sym}=[-0.2;0.2]), as seen by the accumulation of blue points. The majority of those systems have temperature errors in the ranges -3 K to 3 K, and pressure and vapor mole fraction errors between -10 % to 10 %. Sizable errors were found for all three VLE properties in case of aqueous mixtures. The mod. UNIFAC(DO) model has qualitatively almost the same error distribution for all three VLE properties as COSMO-SAC-dsp, however, its error distribution maximum is closer to zero. Aqueous mixtures also exhibit considerable errors, although they are smaller than in case of COSMO-SAC-dsp. Aside from this, those figures underline that the Stavermann-Guggenheim combinatorial term within the COSMO-SAC models performs well. As revealed by Figures 4, S.3 and S.4, it is clear that the performance of all methods mainly depends on chemical families.

Figure 5 shows for COSMO-SAC-dsp and mod. UNIFAC(DO) the temperature, pressure and vapor mole fraction error distribution of all data sets. They reflect the same behaviors of the models as shown in Figures S.18 to S.20 and S.24 to S.26, featured in a different way of presentation to clarify the statistical circumstances. The error distribution is compared to Cauchy (black line) and normal (blue line) distribution functions. They point out that all methods lead an almost Cauchy distributed error, but COSMO-SAC-dsp is slightly closer to a normal distribution than mod. UNIFAC(DO).

4.3 Summary

The performance of COSMO-SAC10 and COSMO-SAC-dsp was studied for VLE predictions on the basis of a very large experimental data set of 139,921 data points. For that purpose, the CMAD was introduced to allow for a condensed overview of the VLE properties temperature, pressure and vapor mole fraction. Table 6 summarizes the VLE errors of all data, non-aqueous and aqueous data.

The consideration of all data sets shows that there is an improvement from COSMO-SAC10 to COSMO-SAC-dsp (CMAD was reduced from 4.77 % to 4.63 %) and from UNIFAC to mod. UNIFAC(DO) (CMAD was reduced from 4.47 % to 3.51 %). Mod. UNIFAC(DO) gives the smallest CMAD (3.51 %), followed by UNIFAC (4.47 %), COSMO-SAC-dsp (4.63 %) and COSMO-SAC10 (4.77 %). All prediction methods exhibit almost Cauchy distributed temperature, pressure and vapor mole fraction errors, whereas both COSMO-SAC versions are slightly closer to a normal distribution than both UNIFAC versions. As expected, UNIFAC methods are more accurate than COSMO-SAC methods in general.

All methods are more accurate for non-aqueous mixtures (90.0 % of all data) than for aqueous mixtures (10.0 %). Mod. UNIFAC(DO) shows the smallest CMAD (3.30 %), followed by COSMO-SAC-dsp (4.25 %), UNIFAC (4.30 %) and COSMO-SAC10 (4.37 %) for non-aqueous mixtures. Again, these data reveal model improvements. For aqueous mixtures very large CMAD were found, thus, these systems affect the CMAD of all data sets. However, this effect is less substantial here due to the smaller number of aqueous mixtures. Mod. UNIFAC(DO) gives the smallest CMAD (5.41 %), followed by UNIFAC (5.98 %), COSMO-SAC-dsp (8.00 %), and COSMO-SAC10 (8.36 %).

Moreover, no correlation between error and molecular asymmetry was found (cf. Figures S.15 to S.26). Instead, the performance of all methods mainly depends on chemical families. Furthermore, a clear dependence between error and polarity of different families is not present (cf. Figures 4, S.3 and S.4).

In the following, some main-family combinations are discussed where COSMO-SAC10 and/or COSMO-SAC-dsp are as good as UNIFAC and/or mod. UNIFAC(DO) or even better, to underline the predictive capabilities of the COSMO-SAC models. For that purpose, the according CMAD $|\bar{\Delta}|$ ranges were used.

- Ethers + Halogenated Hydrocarbons: For this main-family combination COSMO-SAC10 and COSMO-SAC-dsp (1 - 3 %) are better than UNIFAC (≥ 10 %) and mod. UNIFAC(DO) (3 - 5 %).
- Ethers + Esters: COSMO-SAC10 and COSMO-SAC-dsp (1 - 3 %) are both as good as UNIFAC (1 - 3 %) and better than mod. UNIFAC(DO) (3 - 5 %).
- Iso(Nitriles) + Alkanes: COSMO-SAC10 and COSMO-SAC-dsp (3 - 5 %) are better than UNIFAC and mod. UNIFAC(DO) (5 - 10 %).
- Alkenes + Amines: COSMO-SAC10 and COSMO-SAC-dsp (1 - 3 %) are better than UNIFAC and mod. UNIFAC(DO) (3 - 5 %).

Some main-family combinations, e.g. Esters + Aromatics or Amides + Carbonyls, show the potential of the COSMO-SAC-dsp development (compared to COSMO-SAC10). For some main-families, e.g. Esters, Amides or Carbonyls, dispersion parameters were developed. Finally, COSMO-SAC-dsp should be applied instead of mod. UNIFAC(DO) for mixtures where no parameters are available for one of the groups or group combinations because of its predictive character (cf. Figure 4).

5. Comparison of γ_i^{∞} and VLE studies

The results of the γ_i^{∞} and VLE analyses were compared. 29,173 data sets were considered for the γ_i^{∞} study, which were divided into 23,816 (81.6 %) non-aqueous and 5,357 (18.4 %) aqueous data points. In the VLE study 139,921 data points were analyzed with 125,888 (90.0 %) non-aqueous and 14,033 (10.0 %) aqueous data points.

The vertical axis of Figure 6 shows the MAD $|\bar{\delta}|$ of the γ_i^{∞} (cf. Table 4) and CMAD $|\bar{\Delta}|$ of the VLE assessment (cf. Table 6). The horizontal axis itemizes γ_i^{∞} and VLE data, and all data sets are distinguished between non-aqueous and aqueous systems. On this basis, the performance of each method can be seen. First, all models show nearly the same tendencies for γ_i^{∞} and VLE predictions, such as model improvements and similar behavior for aqueous vs. non-aqueous mixtures. It is striking that the errors of VLE calculations are much smaller than the γ_i^{∞} errors. However, this was expected because experimental γ_i^{∞} data scatter more. Furthermore, this figure documents the model improvement from COSMO-SAC10 to COSMO-SAC-dsp and from UNIFAC to mod. UNIFAC(DO) for all types of considered data sets. The sole exception are γ_i^{∞} of aqueous systems. Here, the model development shows no improvement, however, COSMO-SAC10 and COSMO-SAC-dsp are better than UNIFAC and mod. UNIFAC(DO).

Each method gives the most accurate predictions for non-aqueous data sets and they are less efficient for aqueous mixtures. The model behavior of γ_i^{∞} and VLE calculations for aqueous mixtures differs remarkably between UNIFAC and COSMO-SAC models. For these systems, γ_i^{∞} predictions are

most accurately done by COSMO-SAC10, whereas mod. UNIFAC(DO) showed significant errors for some combinations. Again, it should be noted that UNIFAC and mod. UNIFAC(DO) have issues with aqueous mixtures [7,9,38]. On the contrary, both UNIFAC methods yield better VLE predictions than both COSMO-SAC methods for aqueous mixtures. In general, the difference between COSMO-SAC-dsp and mod. UNIFAC(DO) is smaller for VLE than for γ_i^∞ calculations.

6. COSMO-SAC10 on a larger data set

This section discusses COSMO-SAC10 without the restriction to the subset of experimental data to which all four models can be applied. It is studied by how much the prediction accuracy is changed when more data are included. Therefore, the largest possible COSMO-SAC10 data set was utilized, increasing the γ_i^∞ data set by about 33 % from 29,173 to 39,014 data points. However, the quality filters to remove little trustworthy experimental data were still active (cf. section 3). This larger data set corresponds to an experimental temperature range from 126 K to 576.15 K and an γ_i^∞ interval from -3.91 to 26.40 in natural logarithm. The VLE data increased by about 18 % from 139,921 to 165,943 data points. Again, two more filters were still active to remove pure fluid data and data with a pressure above 1000 kPa (cf. section 4). The VLE data cover a temperature range from 183 K to 638.15 K and a pressure range from 0.02 kPa to 1000 kPa. Figure 7 depicts the MAD $|\bar{\delta}|$ of the γ_i^∞ data and Figure 8 the CMAD $|\bar{\Delta}|$ of the VLE data in percentage ranges for all binary main-family combinations for the COSMO-SAC10 model. The solid circles show the original data set discussed above and the open circles the larger data set.

In general, the γ_i^∞ and VLE results are very similar for most family combinations, e.g. Alkanes + Esters or Alcohols + Carbonyls. For a few family combinations, improvements can be seen for γ_i^∞, e.g. Ethers + Esters or Amines + Esters, and for VLE, e.g. Carbonyls + (Iso)Nitriles or Water + OtherNitrogens, when the larger data set is applied. However, there are also some family combinations where the accuracy is reduced for both γ_i^∞ and VLE when the larger data set is used, e.g. Multifunctionals + Aromatics or Amines + Ethers. A few combinations that are not present in the original data set are well predicted, e.g. Alkanes + Thiols/Thioethers or Carbonates + Carbonyls, and some are not adequately predicted, e.g. Water + Sulfoxides&Sulfonyls. In some cases, it becomes clear that the method is inapplicable, e.g. Acids + Alcohols/Water or Multifunctionals/Carbonyls + Water, which are foremost acid and base mixtures. However, it should be noted that the experimental data for Acids + Alcohols are unreliable because esterification may occur during the measurement.

It was confirmed that consistent results can be achieved when a larger data set is considered. Therefore, the above analyses based on the original data set seem to be representative.

7. Conclusion

The QC based models COSMO-SAC10 [20] and COSMO-SAC-dsp [21] were analyzed with respect to their accuracy for γ_i^∞ and VLE predictions. For that purpose, the COSMO-SAC models were evaluated for the first time on a very large experimental data set. Both models yield predictions for the chemical potential without binary parameters. They rely on a few global and atomic parameters only, which are independent from experiments (complete independence for COSMO-SAC10, whereas COSMO-SAC-dsp includes 13 global parameters). This is a major advantage over group contribution methods, like UNIFAC [5-7] or mod. UNIFAC(DO) [9], which require 1270 parameters or 2484 parameters (public parameter file from 2007 [24]), respectively. These semi-empirical models are highly accurate because their parameters were fitted to all available experimental data in a longstanding effort.

One aim of this work was to establish a rationale for the further development of both COSMO-SAC models, particularly COSMO-SAC-dsp. The COSMO-SAC-dsp model differs from its preceding version by consideration of the dispersive intermolecular interactions on the basis of molecular simulation data. In this study, all binary mixtures, which are available in the DDB, were utilized and both COSMO-SAC models were compared to UNIFAC and mod. UNIFAC(DO) based on chemical families. Therefore, 29,173 γ_i^∞ and 139,921 VLE data points were studied and divided into aqueous and non-aqueous systems as well as into chemical families. The 16 main-families can be studied in more detail on two additional hierarchical levels of sub-families. In order to do this, MATLAB programs were created for this type of analysis that can be undertaken by the interested reader.

In general, the γ_i^∞ and VLE studies showed the same trends. A clear improvement from COSMO-SAC10 to COSMO-SAC-dsp as well as from UNIFAC to mod. UNIFAC(DO) was found. Mod. UNIFAC(DO) performed best, followed by UNIFAC, COSMO-SAC-dsp and COSMO-SAC10 with only slight differences. This assessment emphasized that the COSMO-SAC-dsp model development was meaningful and that the dispersive interactions should be taken into account, even though they are just a small part of the total intermolecular interaction energy. Mixtures for which dispersion parameters were established showed an improvement. As a result, the 13 dispersion energy parameters for the atoms C, O, N, F, Cl and H should be extended to more atom or bonding types, respectively.

The analysis between aqueous and non-aqueous data sets showed that each method is more accurate for non-aqueous mixtures. Mod. UNIFAC(DO) performs poorly in case of γ_i^∞ calculations for aqueous mixtures, however, its accuracy is much better for VLE calculations. As a consequence, both COSMO-SAC models should be applied for γ_i^∞ predictions of aqueous systems and they should give better predictions than the UNIFAC models for mixtures where no parameters are available for one of the groups or group combinations because of their strictly predictive character.

The large errors in predicting data for aqueous systems with COSMO-SAC10 and COSMO-SAC-dsp occur for highly polar compounds, most of which are either hydrogen bond donors or acceptors. Hydrogen bonds in strongly associating fluids, including water, are directional. Such directional interactions should restrict the range of interaction surfaces and are not considered in the present model. However, the inclusion of these geometrical constraints in the COSMO-SAC model is in progress and preliminary results show that the consideration of directional hydrogen bonding does improve the prediction accuracy for associating fluids. As a result, there are further possible improvements for the COSMO-SAC model development. First, the explicit consideration of directional hydrogen bonding (in progress). Second, a more sophisticated model for the dispersive interaction, which, however, may result in additional empirical parameters that must be determined by regression to experimental data. Third, the dispersive interactions may be obtained from molecular dynamics or Monte Carlo simulations with suitable force fields to describe the interactions. Our recent

study showed that the dispersive term can be directly obtained from the van der Waals component of the solvation free energy derived from thermodynamic integration [40].

This assessment indicates that model efficiency strongly correlates with the type of chemical family. No significant correlation between model accuracy and polarity as well as between error and molecular size asymmetry was found. The γ_i^{∞} study on the main-family level showed that COSMO-SAC-dsp gives good predictions for almost all binary mixtures. Admittedly, several main-family combinations containing Water as a solute are challenging. The performance of Alcohols as solutes in combination with the main-families Alkanes, Alkenes and Acids also indicates issues. Mixtures with Water as a solvent are predicted decently. UNIFAC and mod. UNIFAC(DO) yield inaccurate descriptions for aqueous systems with aliphatic hydrocarbons. The VLE study showed that mixtures with main-families OtherNitrogens (several different bonding types of N), Acids and Water are not predicted well by COSMO-SAC-dsp. Apart from that, all mixtures with the remaining hydrocarbons are well predicted. UNIFAC and mod. UNIFAC(DO) perform well, except for aqueous mixtures, as well as for some family combinations with HalogenatedHydrocarbons, Ethers, Amines and Amides (notably UNIFAC).

For the first time both COSMO-SAC models were analyzed on a very large experimental data set. The COSMO-SAC-dsp development is encouraging because it requires only a few global and atomic parameters. With a minor computing intensity, the COSMO-SAC models are capable to provide convincing phase equilibrium property predictions for a wide range of mixtures.

Acknowledgement

This research was financially supported by BMBF under the grants 01IH13005G and 01IH13005I SkaSim: Skalierbare HPC-Software für molekulare Simulationen in der chemischen Industrie" and computational support was given by the High Performance Computing Center Stuttgart (HLRS) under the grant MMHBF2. Furthermore, we gratefully acknowledge the Paderborn Center for Parallel Computing (PC2) for the generous allocation of computer time on the OCuLUS cluster. It was also partially supported by the Ministry of Science and Technology of Taiwan (104-2221-E-002-186-MY3 and 105-2221-E-008-106). R. F. gratefully acknowledges support by Ernest-Solvay-Stiftung. The computation resources of the National Center for High-Performance Computing of Taiwan and the Computing and Information Networking Center of the National Taiwan University are acknowledged as well.

ASSOCIATED CONTENT

Supporting Information. Main-family and sub-family figures, error distributions, MATLAB code. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

* Shiang Tai-Lin, Department of Chemical Engineering, National Taiwan University, 10617 Taipei City, Taiwan, E-mail: stlin@ntu.edu.tw

Co-corresponding Author

* Jadran Vrabec, Thermodynamics and Energy Technology, University of Paderborn, 33098 Paderborn, Germany, E-mail: jadran.vrabec@upb.de

Funding Sources

Federal Ministry of Education and Research (BMBF) and Ernest-Solvay-Stiftung

REFERENCES

(1) Hendriks, E.; Kontogeorgis, G. M.; Dohrn, R.; Hemptinne, J.-C.; Economou, I. G.; Zilnik, L. F.; Vesovic, V. Industrial requirements for thermodynamics and transport properties. *Ind. Eng. Chem. Res.* **2010**, 49, 11131.

(2) Gubbins, K. E.; Moore, J. D. Molecular modeling of matter: Impact and prospects in engineering. *J. Chem. Phys.* **2010**, 49, 3026.

(3) Deublein, S.; Eckl, B.; Stoll, J.; Lishchuk, S. V.; Guevara-Carrion, G.; Glass, C. W.; Merker, T.; Bernreuther, M.; Hasse, H.; Vrabec, J. ms2: A molecular simulation tool for thermodynamic properties. *Comput. Phys. Commun.* **2011**, 182, 2350.

(4) Glass, C. W.; Reiser, S.; Rutkai, G.; Deublein, S.; Köster, A.; Guevara Carrion, G.; Wafai, A.; Horsch, M.; Bernreuther, M.; Windmann, T.; Hasse, H.; Vrabec, J. ms2: A molecular simulation tool for thermodynamic properties, new version release. *Comput. Phys. Commun.* **2014**, 185, 3302.

(5) Fredenslund, A.; Gmehling, J.; Rasmussen, P. *Vapor-Liquid Equilibria Using UNIFAC - A Group Contribution Method*; Elsevier: Amsterdam, 1977.

(6) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. *AIChE J.* **1975**, 21, 1086.

(7) 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. Proc. Des. Dev.* **1977**, 16, 450.

(8) Wittig, R.; Lohmann, J.; Gmehling, J. Vapor−Liquid Equilibria by UNIFAC Group Contribution. 6. Revision and Extension. *Ind. Eng. Chem. Res.* **2003**, 42, 183.

(9) Jakob, A.; Grensemann, H.; Lohmann, J.; Gmehling, J. Further development of modified UNIFAC (Dortmund): Revision and extension 5. *Ind. Eng. Chem. Res.* **2006**, 45, 7924.

(10) Gmehling, J. Phase-Equilibrium Models in the Synthesis and Design of Separation Processes. *Chem. Ing. Tech.* **1994**, 66, 792.

(11) Constantinescu, D.; Gmehling, J. Further development of modified UNIFAC (Dortmund): Revision and extension 6. *J. Chem. Eng. Data* **2016**, 61, 2738.

(12) Horstmann, S.; Jabloniec, A.; Krafczyk, J.; Fischer, K.; Gmehling, J. PSRK group contribution equation of state: comprehensive revision and extension IV, including critical constants and α-function parameters for 1000 components. *Fluid Phase Equilib.* **2005**, 227, 157.

(13) Schmid, B.; Schedemann, A.; Gmehling, J. Extension of the VTPR Group Contribution Equation of State: Group Interaction Parameters for Additional 192 Group Combinations and Typical Results. *Ind. Eng. Chem. Res*. **2014**, 53, 3393.

(14) 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.

(15) Klamt, A.; Schuurmann, G. COSMO - A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc.-Perkin Trans. 2*, **1993**, 799.

(16) Klamt, A.; Jonas, V.; Burger, T.; Lohrenz, J.C.W. Refinement and Parametrization of COSMO-RS. *J. Phys. Chem. A* **1998**, 102, 5074.

(17) Klamt, A. *COSMO-RS from Quantum Chemistry to Fluid Phase Thermodynamicsand Drug Design*, Elsevier: Amsterdam 2005.

(18) Lin, S.-T.; Sandler, S. I. A priori phase equilibrium prediction from a segment contribution solvation model. *Ind. Eng. Chem. Res.* **2002**, 41, 899.

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

(20) 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.

(21) 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 Equilibr.* **2014**, 367, 109.

(22) Grensemann, H.; Gmehling, J. Performance of a Conductor-Like Screening Model for Real Solvents Model in Comparison to Classical Group Contribution Methods. *Ind. Eng. Chem. Res.* **2005**, 44, 1610.

(23) 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. Supercrit. Fluids* **2009**, 50, 210.

(24) Dortmund Data Bank, http://www.ddbst.com (February 20, 2017).

(25) 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.

(26) 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.

(27) 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.

(28) Wang, L.-H.; Hsieh, C.-M.; Lin, S.-T. Improved prediction of vapor pressure for pure liquids and solids from the PR plus COSMOSAC equation of state. *Ind. Eng. Chem. Res.* **2015**, 54, 10115.

(29) Hsieh, C.-M.; Lin, S.-T. First-principles prediction of phase equilibria using the PR+COSMOSAC equation of state. *Asia-Pac. J. Chem. Eng.* **2012**, 7, S1-S10.

(30) Lin, S.-T.; Wang, L.-H.; Chen, W.-L.; Lai, P.-K.; Hsieh, C.-M. Prediction of miscibility gaps in water/ether mixtures using COSMO-SAC model. *Fluid Phase Equilib.* **2011**, 310, 19.

(31) Hsieh, C.-M.; Wang, S.; Lin, S.-T.; Sandler, S. I. A predictive model for the solubility and octanol-water partition coefficient of pharmaceuticals. *J. Chem. Eng. Data* **2011**, 56, 936.

(32) Hsieh, C.-M.; Lin, S.-T. Prediction of liquid-liquid equilibrium from the Peng-Robinson plus COSMOSAC equation of state. *Chem. Eng. Sci.* **2010**, 65, 1955.

(33) 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.

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

(35) Staverman, A. J. The entropy of high polymer solutions - Generalization of formulae. *Recl. Trav. Chim. Pays-Bas-J. Roy. Neth. Chem. Soc.* **1950**, 69, 163.

(36) Guggenheim, E. A. *Mixtures*. Oxford University Press: Oxford, 1952.

(37) Prausnitz, J. M.; Lichtenthaler, R. N.; de Azevedo, E. G. *Molecular Thermodynamics of Fluid-Phase Equilibria*. 3rd ed.; Pearson Education Taiwan: Taipei 2004.

(38) Gmehling, J.; Kolbe, B.; Kleiber, M.; Rarey, J. *Chemical Thermodynamics for Process Simulation*; Wiley-VCH Verlag: Weinheim 2012.

(39) Baehr, H. D.; Kabelac, S. *Thermodynamik - Grundlagen und technische Anwendungen*; Springer Vieweg Press: Berlin Heidelberg 2012.

(40) Yang, L.; Chang, C.-W.; Lin, S. T. A Novel Multiscale Approach for Rapid Prediction of Phase Behaviors with Consideration of Molecular Conformations. *AIChE J.* **2016**, 62, 4047.

(41) Kurzweil, P. *Chemie - Grundlagen, Aufbauwissen, Anwendungen und Experimente*; Springer Vieweg Verlag: Berlin Heidelberg 2015.

(42) Nannoolal, Y.; Rarey, J.; Ramjugernath, D.; Cordes, W. Estimation of pure component properties: Part 1. Estimation of the normal boiling point of non-electrolyte organic compounds via group contributions and group interactions. *Fluid Phase Equilib.* **2004**, 226, 45.

Table 1. Parameter values of the COSMO-SAC models.

(a) Universal Parameters

Parameter	Value
a_{eff} (Å^2)	7.25
f_{decay} (-) [a]	3.57
σ_0 (e/Å^2)	0.007
r (Å^3)	66.69
q (Å^2)	79.53
A_{ES} (kcal/mol)(A^4/e^2)	6525.69
B_{ES} (kcal/mol)(A^4/e^2)K^2	1.4859×10^8
$c_{\text{OH-OH}}$ (kcal/mol)(A^4/e^2)	4013.78
$c_{\text{OT-OT}}$ (kcal/mol)(A^4/e^2)	932.31
$c_{\text{OH-OT}}$ (kcal/mol)(A^4/e^2)	3016.43
w (-) [b]	±0.27027

(b) Atomic Parameters

Atom type (Hybridization type)	ε'_k/k_B (K)	Radius (Å) [c]
C (sp)	66.0691	2.00
C (sp2)	117.4650	
C (sp3)	115.7023	
O (sp3, -O-)	95.6184	1.72
O (sp2, =O)	-11.0549	
N (sp3)	15.4901	1.83
N (sp2)	84.6268	
N (sp)	109.6621	
F	52.9318	1.72
Cl	104.2534	2.05
H (OH)	19.3477	1.30
H (NH)	141.1709	
H (H_2O/COOH) [d]	58.3301	

[a.] The empirical parameter f_{decay} is used in the semi-theoretical equation of the molecular surface charge averaging process [18].

[b.] Substances are categorized into three groups in the COSMO-SAC-dsp model: non-hydrogen-bonding (*nhb*), hb-only-acceptor (*hb-a*), and hb-donor–acceptor (*hb-da*) [21]. *w* is negative for systems of H_2O + *hb-a*, COOH + *nhb* or *hb-da*, and H_2O + COOH.

[c.] The atomic radii are used in the COSMO solvation calculations.

[d.] H2O represents water and COOH represents molecules with a carboxyl group, e.g. carboxylic acids.

Table 2. Main-family structure that was utilized here; roughly sorted in the order of ascending polarity.

main-family	members	examples
Gases	7	carbon dioxide, methane, carbonyl sulfide
Multifunctionals	477	3-hydroxy-2-butanone, tetrachloroethylene, vinyl chloride
OtherNitrogens	58	1-nitrobutane, 1,1-dimethylhydrazine, o-nitrotoluene
Alkanes	264	n-butane, biphenyl, cyclohexane
Alkenes	175	1-hexene, cyclohexane, isoprene
Allenes	6	1,2-butadiene, propadiene, dimethylallene
Ketenes	1	ketene
Alkynes	21	2-hexyne, diphenylacetylene, 1,6-heptadiyne
Aromatics	323	benzene, 2-fluorotoluene, anthracene
Carbonates	9	phosgene, ethyl chloroformate, carbonyl fluoride
Epoxies	11	1,2-epoxyhexane, ethylene oxide, 2,2-dimethyloxirane
Esters	154	ethyl acetate, propylacrylate, tetrahydropyran-2-one
Halogenated hydrocarbons	147	butyl chloride, pentachlorofluoroethane, ethyl bromide
Halogens	6	chlorine, iodine, bromine trifluoride
Ethers	78	oxetane, methyl pentyl ether, dibenzo-p-dioxin
Peroxy (no acids)	10	hydrogen peroxide, cyclohexanhydroperoxide, n-butylhydroperoxide
Acids	71	acetic acid, dehydroabietic acid, hydrogen bromide
Anhydrides	7	acetic anhydride, tetrahydropyran-2,6-dione
Amines	126	butylamine, 1,3-diaminopropane, 1-aminononane
Carbonyls	96	3-pentanone, n-decanal, 2-methylbenzaldehyde
Thiols	27	methanethiol, 1-decanethiol, undecyl mercaptan
Thioethers	22	tetrahydrothiophene, diethyl sulfide, tetrahydrothiopyran
Alcohols	127	1-butanol, cyclooctanol, D-glucitol
Amides	29	acrylamide, N-methylformamide, acetanilide
(Iso)Nitriles	28	butanenitrile, benzylcyanide, 1,5-dicyanopentane
Sulfoxides & Sulfonyls	14	dipropylsulfoxide, ethyl isopropyl sulfoxide, dipropyl sulfone
Water	1	water

Table 3. Explanations for sub-family abbreviations.

symbol / abbreviation	explanation
+	combinations, e.g. Alkanes+Cyclic: hydrocarbons formed by chains and rings
_	more combinations are allowed, e.g. Alkane_Cyclic: chains and rings are allowed in a family
chain	functional group on a chain
ring	functional group on a ring
conj	conjugated double bond
noconj	non conjugated double bond
_aromat	functional group on an aromatic
prim, sec, tert	functional group is primarily, secondarily, tertiary arranged
single	only one functional group
x	more than one functional group
intra	direct intramolecular interaction, e.g. hydrogen bridge bond in ethandiol
nointra	no direct intramolecular interaction between two functional groups
&	or, e.g. ring&chain: family with rings or chains

Table 4. γ_i^{∞} error comparison, distinguishing between all data, non-aqueous and aqueous data; the errors are converted from natural logarithm and given in relative terms.

		COSMO-SAC10	COSMO-SAC-dsp	UNIFAC	mod.UNIFAC(DO)
all data	data points	29,173	29,173	29,173	29,173
	$\bar{\delta}$ / %	**-17.30**	**-1.00**	**-29.53**	**-20.55**
	min. error / %	-99.26	-98.98	-100.00	-99.94
	max. error / %	$13.26 \cdot 10^{8}$	$82.03 \cdot 10^{7}$	$32.36 \cdot 10^{7}$	$13.59 \cdot 10^{10}$
	$\lvert\bar{\delta}\rvert$ / %	**95.42**	**85.89**	**73.33**	**58.41**
non-aqueous data (81.6%)	data points	23,816	23,816	23,816	23,816
	$\bar{\delta}$ / %	**-29.53**	**-18.13**	**-24.42**	**-10.42**
	min. error / %	-99.26	-98.98	-99.10	-98.53
	max. error / %	$49.91 \cdot 10^{2}$	$56.40 \cdot 10^{2}$	$32.36 \cdot 10^{7}$	$88.75 \cdot 10^{1}$
	$\lvert\bar{\delta}\rvert$ / %	**78.60**	**64.87**	**49.18**	**27.12**
aqueous data (18.4%)	data points	5,357	5,357	5,357	5,357
	$\bar{\delta}$ / %	**66.53**	**120.34**	**-49.84**	**-52.76**
	min. error / %	-99.22	-98.39	-100.00	-99.94
	max. error / %	$13.26 \cdot 10^{8}$	$82.03 \cdot 10^{7}$	$15.79 \cdot 10^{6}$	$13.59 \cdot 10^{10}$
	$\lvert\bar{\delta}\rvert$ / %	**194.47**	**203.44**	**232.01**	**305.52**

Table 5. Mean absolute deviations for all isothermal and isobaric VLE data.

	isothermal VLE		isobaric VLE	
	$\lvert\bar{\delta}\rvert_p$ / %	$\lvert\bar{\delta}\rvert_y$ / %	$\lvert\bar{\delta}\rvert_T$ / K	$\lvert\bar{\delta}\rvert_y$ / %
COSMO-SAC10	6.34	2.11	1.77	2.46
COSMO-SAC-dsp	5.72	1.98	1.74	2.46
UNIFAC	5.29	1.95	1.69	2.39
mod. UNIFAC(DO)	3.88	1.45	1.37	1.91

Table 6. VLE error comparison, distinguishing between all data, non-aqueous and aqueous data.

		COSMO-SAC10	COSMO-SAC13dsp	UNIFAC	mod.UNIFAC(DO)
all data	data points	139,921	139,921	139,921	139,921
	min. error / %	0	0	0	0
	max. error / %	741.28	946.60	396.18	399.20
	$\lvert\bar{\Delta}\rvert$ / %	**4.77**	**4.63**	**4.47**	**3.51**
non-aqueous data (90%)	data points	125,888	125,888	125,888	125,888
	min. error / %	0	0	0	0
	max. error / %	152.06	257.87	152.14	180.32
	$\lvert\bar{\Delta}\rvert$ / %	**4.37**	**4.25**	**4.30**	**3.30**
aqueous data (10%)	data points	14,033	14,033	14,033	14,033
	min. error / %	0	0	0	0
	max. error / %	741.28	946.60	396.18	399.20
	$\lvert\bar{\Delta}\rvert$ / %	**8.36**	**8.00**	**5.98**	**5.41**

Table 7. Sub-families of the main-family Alcohols.

main-family	sub-families (level 1)	sub-families (level 2)
Alcohols	Methanol	Methanol
	Alchol_nointra_single	Alcohol_chain
		Alcohol_ring
		Alcohol_sec
		Alcohol_tert
	xAlcohol_intra	xAlcohol_intra_chain(1,2)
		xAlcohol_intra_chain(1,3)
		xAlcohol_intra_ring(1,2)
		xAlcohol_intra_ring(1,3)
		Enol
	xAlcohol_nointra	Diol
		Triol

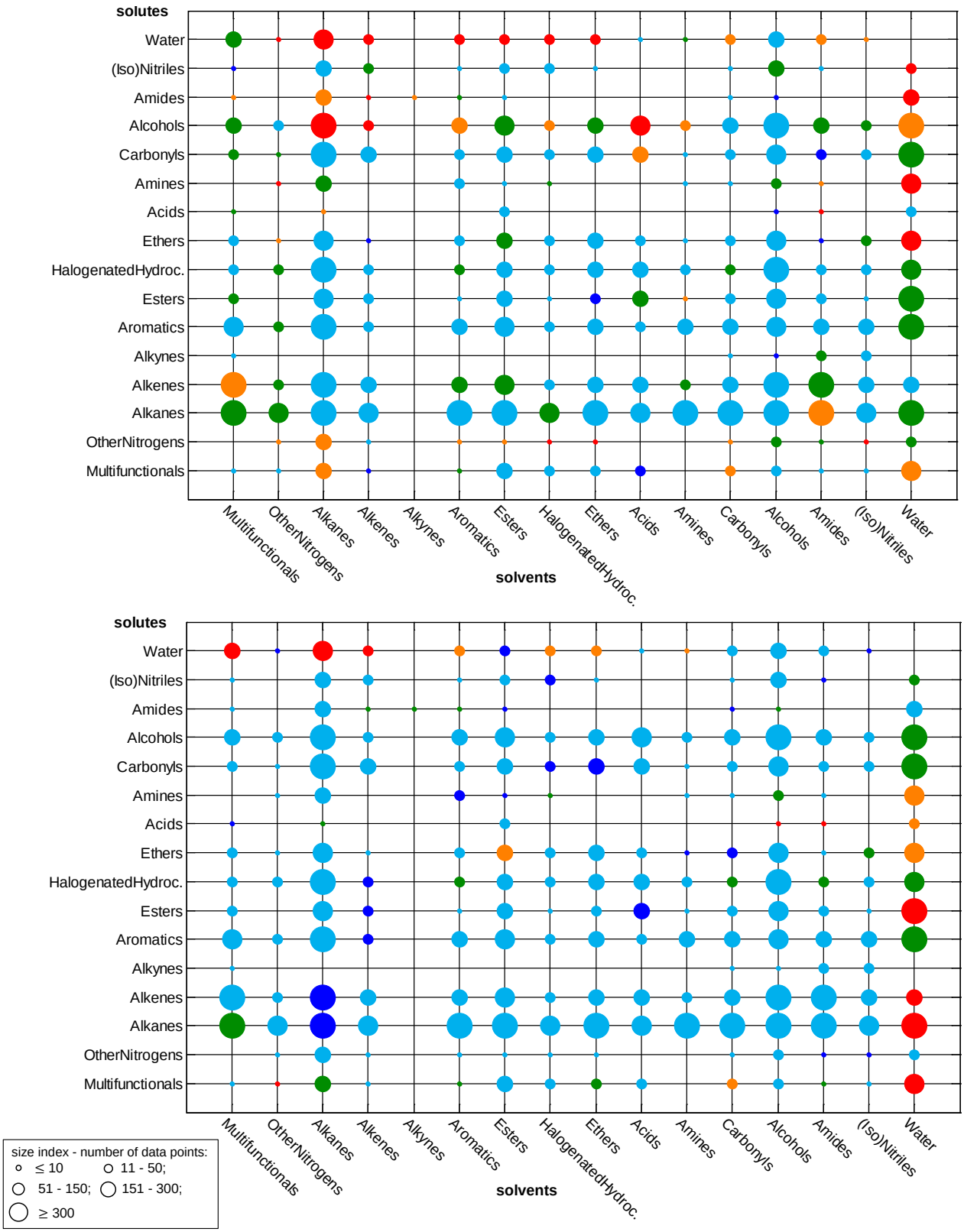


Figure 1. Mean absolute deviation $|\overline{\delta}|$ (eq. 10) for infinite dilution activity coefficients of all main-family combinations for COSMO-SAC-dsp (top) and mod. UNIFAC(DO) (bottom); color index: ● $|\overline{\delta}| \leq 0.1$ (10.5%); ● $|\overline{\delta}|$ 0.1-0.5 (10.5-64.9%); ● $|\overline{\delta}|$ 0.5-0.9 (64.9-146.0%); ● $|\overline{\delta}|$ 0.9-1.3 (146.0-266.9%); ● $|\overline{\delta}| \geq 1.3$ (266.9%).

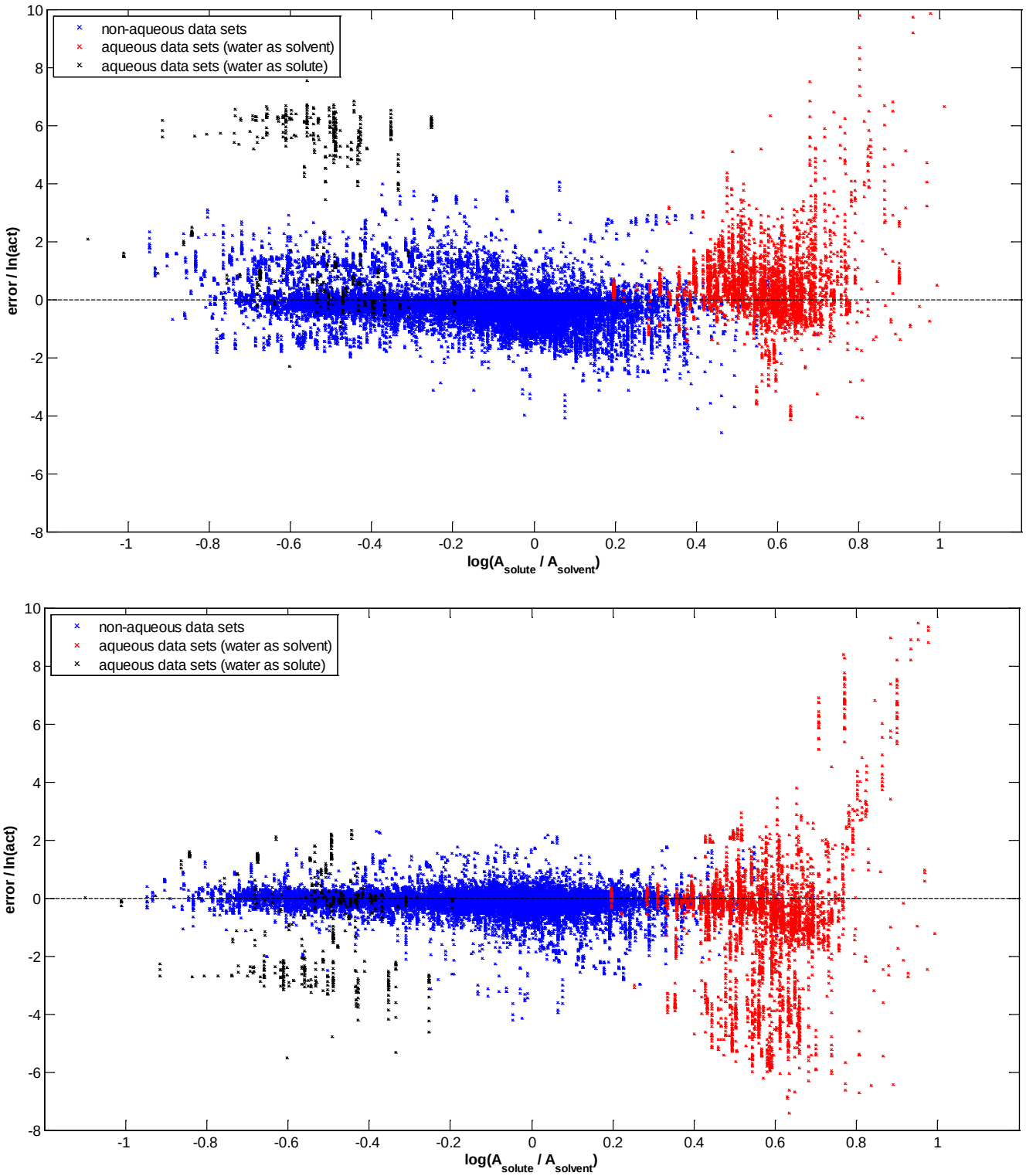


Figure 2. γ_i^{∞} error distribution over the ratio of molecular surface area for COSMO-SAC-dsp (top) and mod. UNIFAC(DO) (bottom).

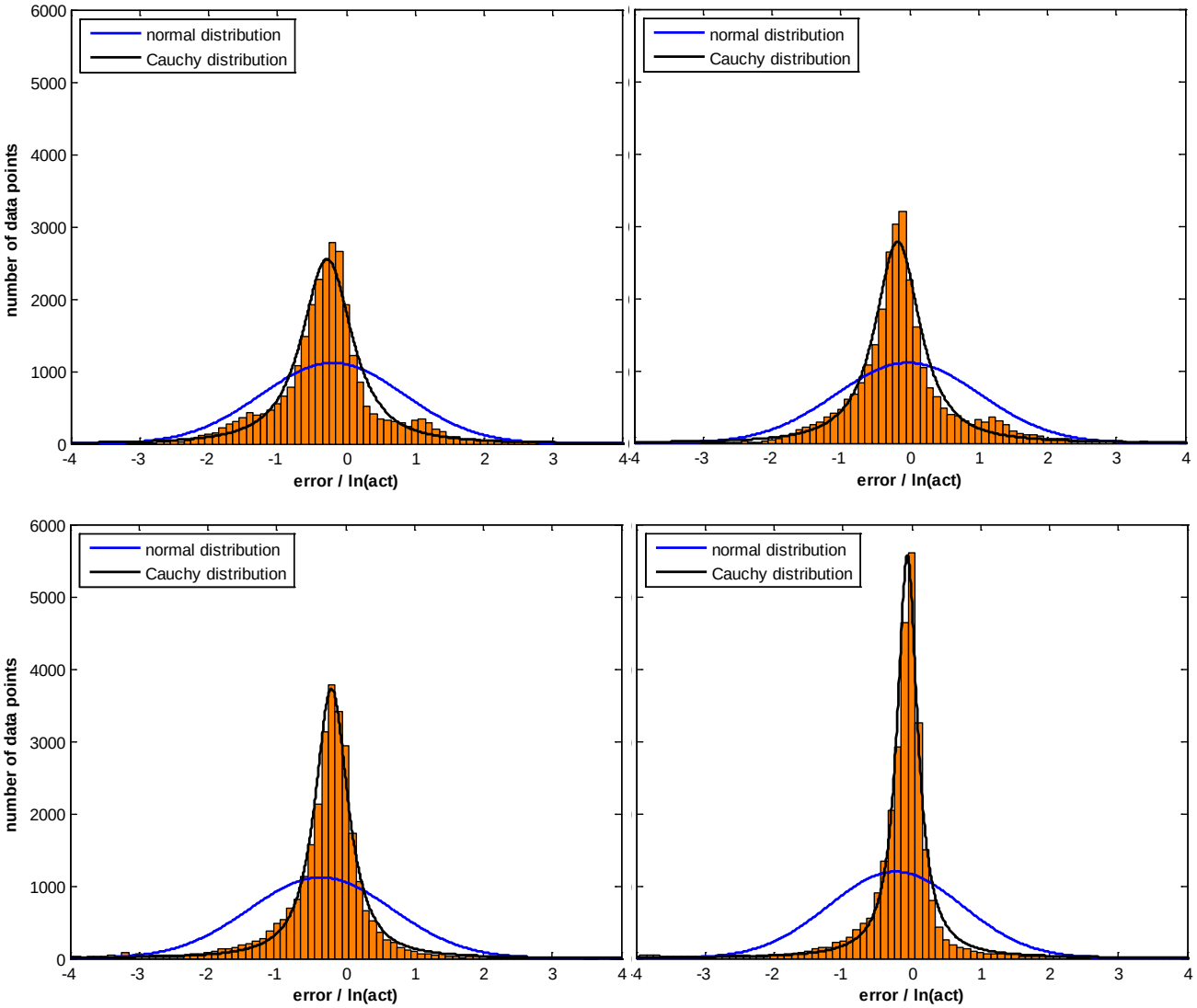


Figure 3. γ_i^{∞} error distribution, i.e. number of data points over model error; top left: COSMO-SAC10; top right: COSMO-SAC-dsp; bottom left: UNIFAC; bottom right: mod. UNIFAC(DO).

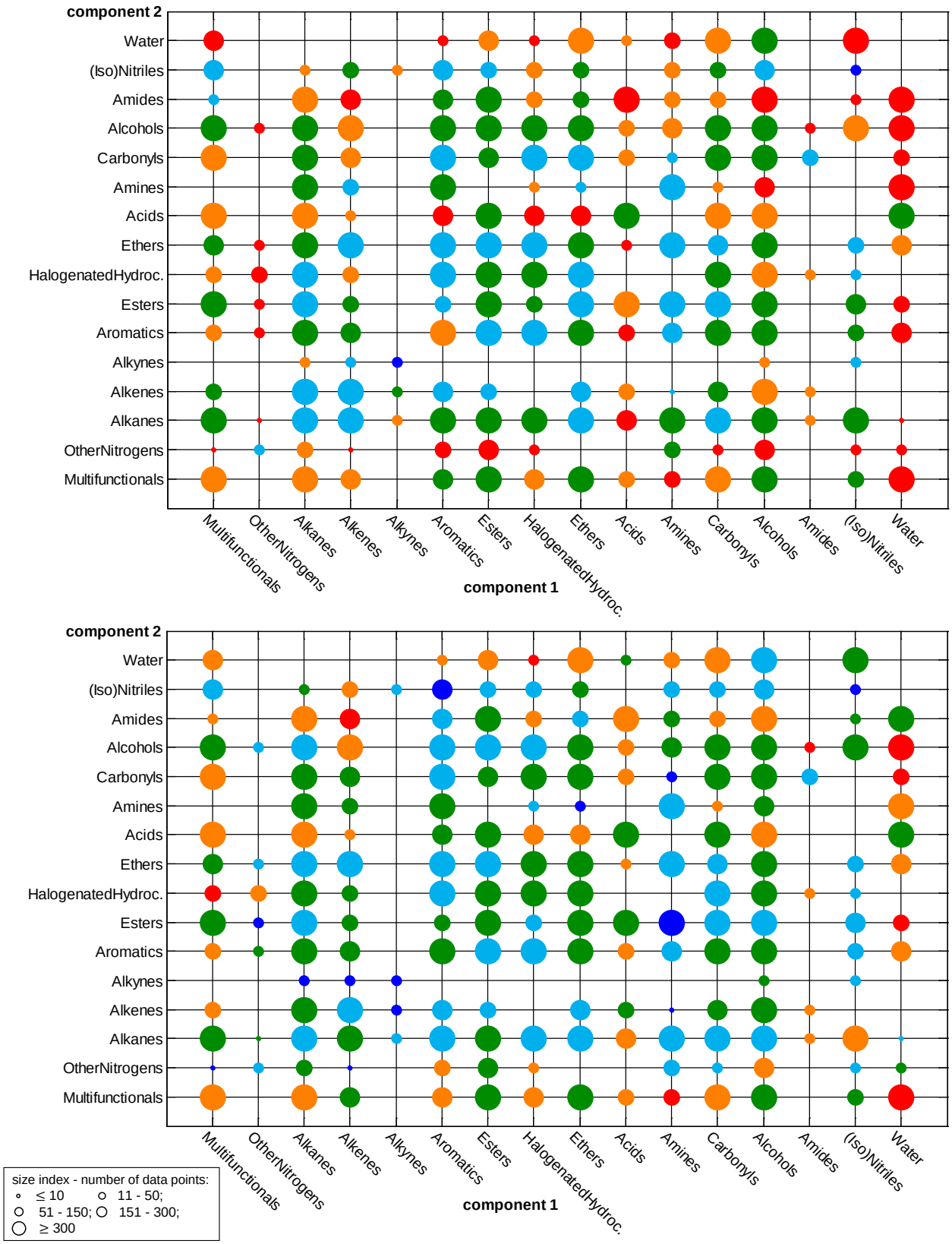


Figure 4. Combined mean absolute deviation $|\bar{\Delta}|$ (eq. 17) for vapor-liquid equilibrium properties of all main-family combinations for COSMO-SAC-dsp (top) and mod. UNIFAC(DO) (bottom); component 1 is low boiling and component 2 is high boiling; color index: ● $|\bar{\Delta}| \leq 1$ %; ● $|\bar{\Delta}|$ 1-3 %; ● $|\bar{\Delta}|$ 3-5 %; ● $|\bar{\Delta}|$ 5-10 %; ● $|\bar{\Delta}| \geq 10$ %.

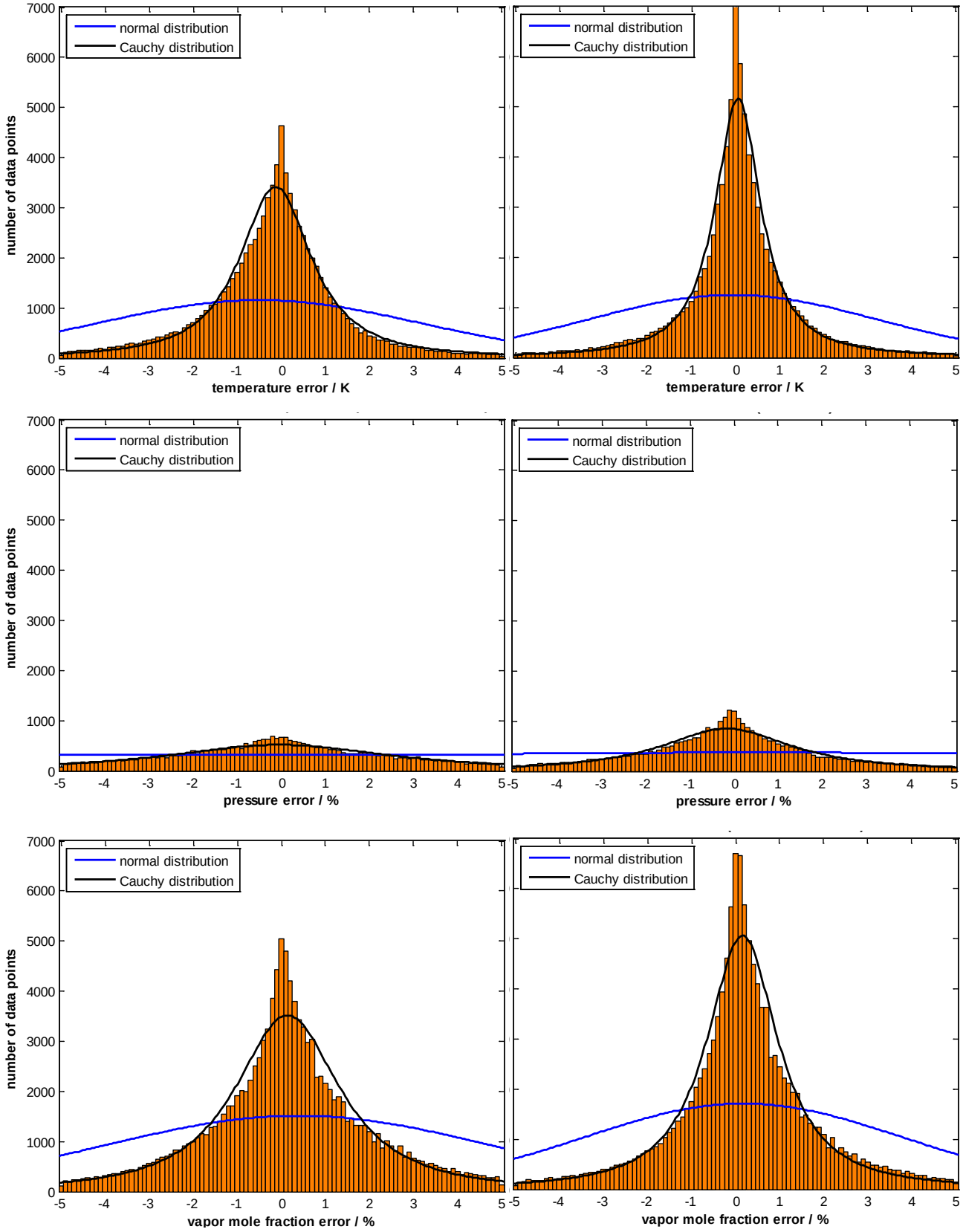


Figure 5. VLE error distribution, i. e. number of data points over model error; left: COSMO-SAC-dsp; right: mod. UNIFAC(DO); top: temperature error; center: pressure error; bottom: vapor mole fraction error.

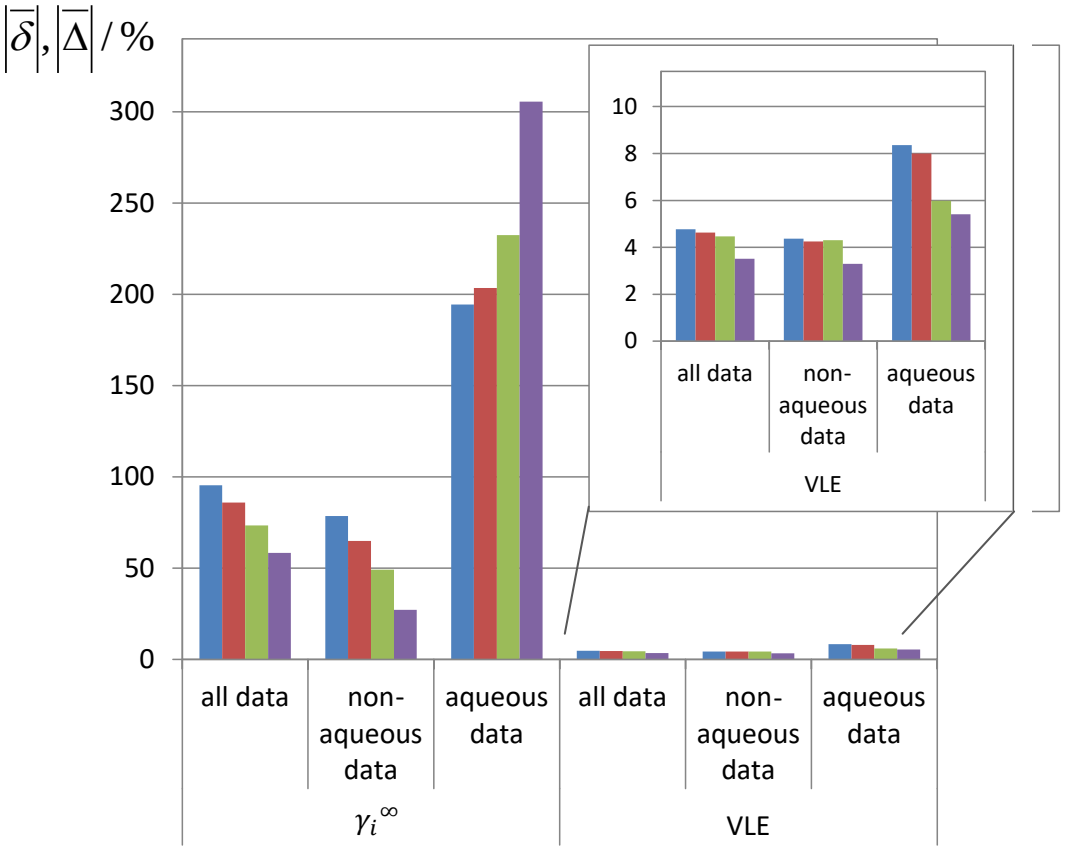


Figure 6. Comparison of γ_i^{∞} and VLE analyses for all models; blue: COSMO-SAC10; red: COSMO-SAC-dsp; green: UNIFAC; purple: mod. UNIFAC(DO).

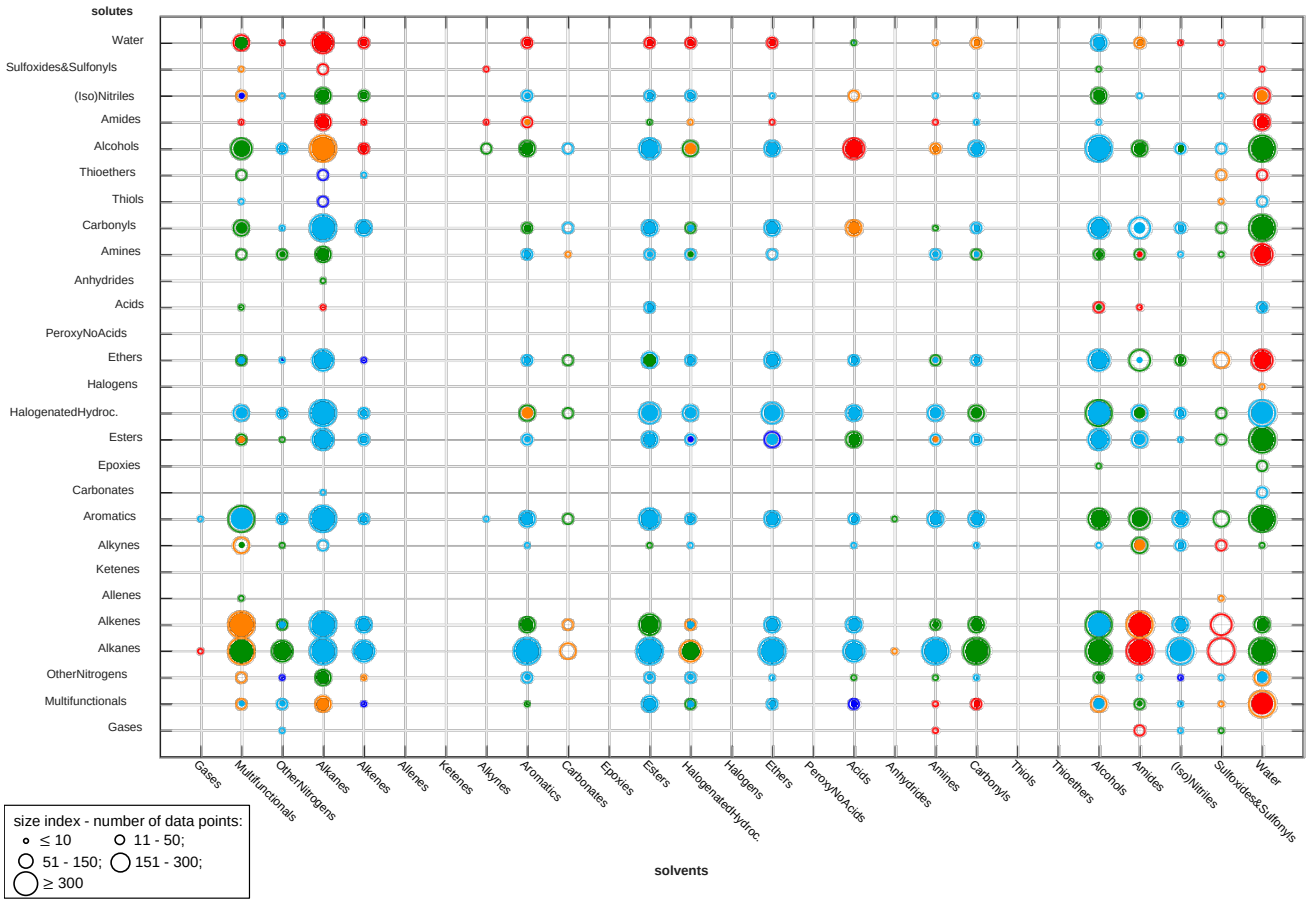


Figure 7. Mean absolue deviation $|\overline{\delta}|$ (eq. 10) for infinite dilution activity coefficients of all main-family combinations for COSMO-SAC10; color index: ● $|\overline{\delta}| \leq 0.1$ (10.5%); ● $|\overline{\delta}|$ 0.1-0.5 (10.5-64.9%); ● $|\overline{\delta}|$ 0.5-0.9 (64.9-146.0%); ● $|\overline{\delta}|$ 0.9-1.3 (146.0-266.%); ● $|\overline{\delta}| \geq 1.3$ (266.9%); solid circles: original data set (29,173 data points); open circles: larger data set (39,014 data points).

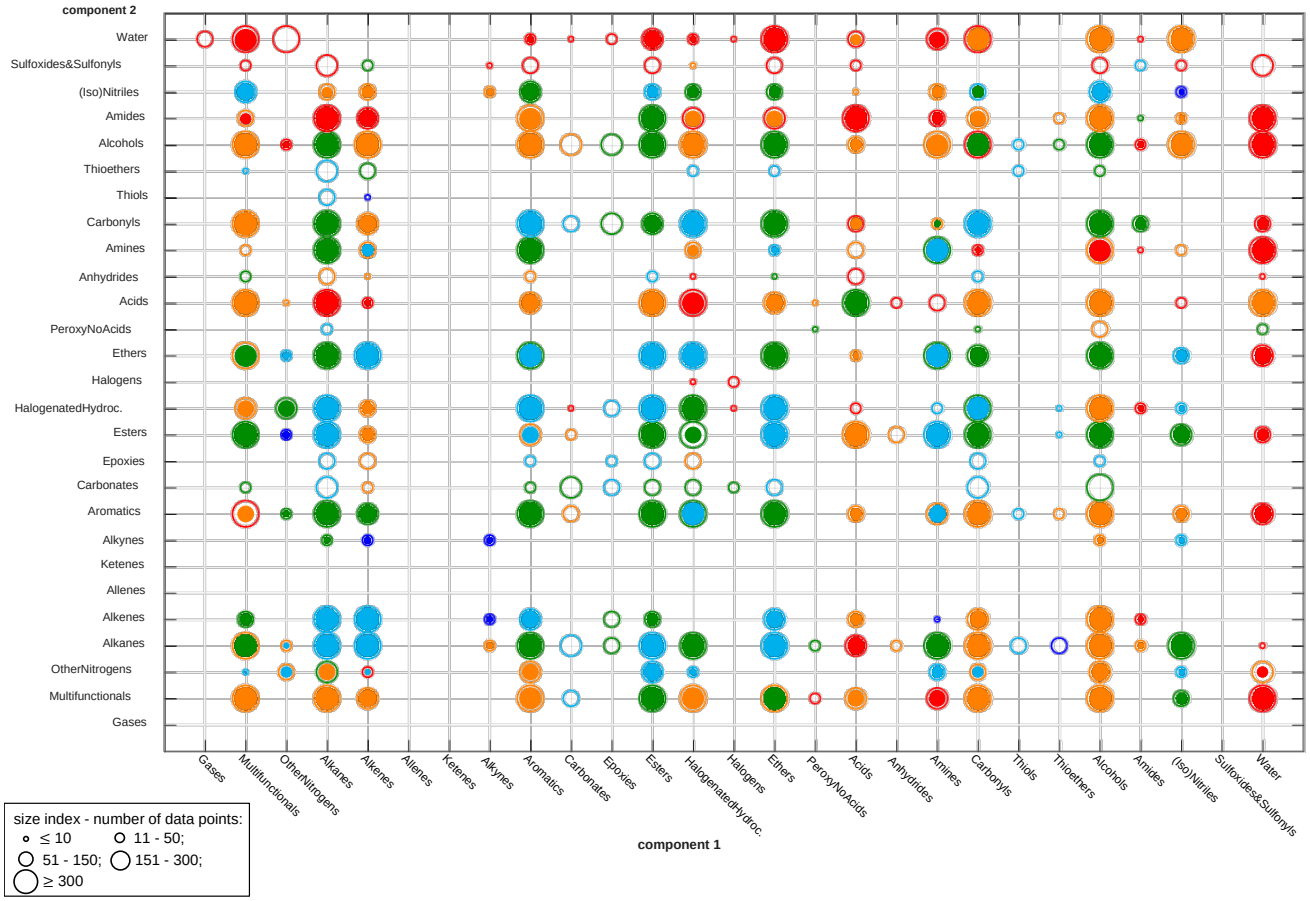
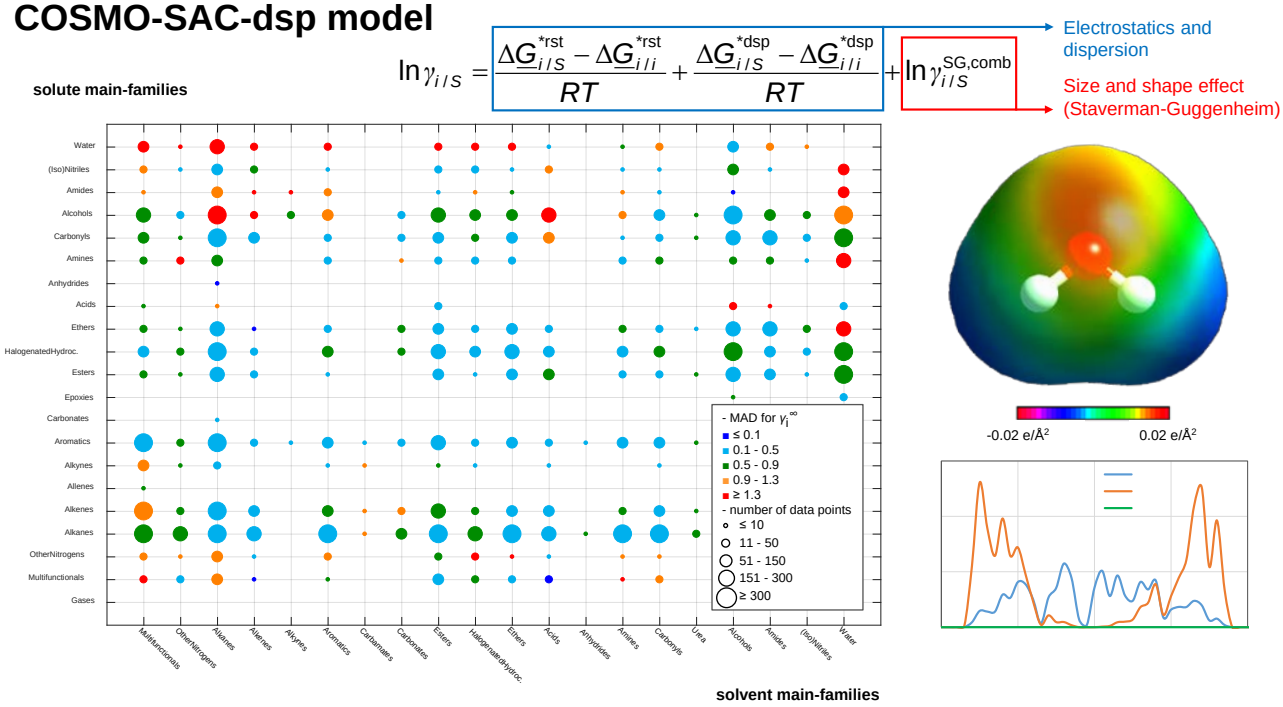


Figure 8. Combined mean absolute deviation $|\bar{\Delta}|$ (eq. 17) for vapor-liquid equilibrium properties of all main-family combinations for COSMO-SAC10; component 1 is low boiling and component 2 is high boiling; color index: ● $|\bar{\Delta}| \leq 1$ %; ● $|\bar{\Delta}|$ 1-3 %; ● $|\bar{\Delta}|$ 3-5 %; ● $|\bar{\Delta}|$ 5-10 %; ● $|\bar{\Delta}| \geq 10$ %; solid circles: original data set (139,921 data points); open circles: larger data set (165,943 data points).



For Table of Contents only