Journal of chemical & engineering data

Article

pubs.acs.org/jced

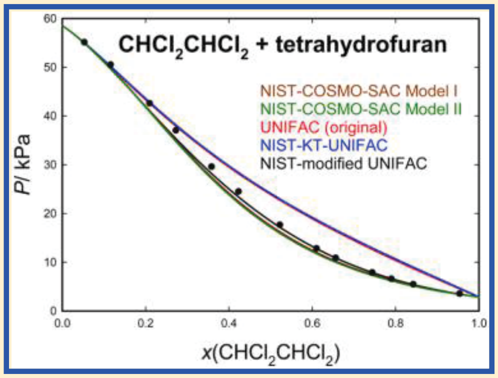
Reparameterization of COSMO-SAC for Phase Equilibrium Properties Based on Critically Evaluated Data

Eugene Paulechka,* Vladimir Diky, Andrei Kazakov, Kenneth Kroenlein, and Michael Frenkel

Applied Chemicals and Materials Division, National Institute of Standards and Technology, Boulder, Colorado 80305-3337, United States

Ⓢ *Supporting Information*

ABSTRACT: COSMO-SAC model was reparameterized with use of the critically evaluated data generated by the NIST ThermoData Engine for vapor−liquid equilibria, excess enthalpies for binary mixtures, and activity coefficients of binary mixture components. The calculated σ-profile library contained 897 individual compounds. The temperature-dependent σ profiles included contributions of up to 40 conformers of a molecule. Splitting of the H-bonding σ profiles into OH and non-OH parts decreased the root-mean square deviation from the experimental data points by about 10% compared to the model using one H-bonding parameter. The original UNIFAC model demonstrated comparable performance with the more advanced COSMO-SAC variation. The challenges of uncertainty evaluation for parameters of the model and the predicted values are discussed.



1. INTRODUCTION

Quantum-chemistry-based statistical thermodynamic calculations are routinely used for reliable prediction of thermodynamic properties of ideal gases. Prediction of thermodynamic properties of pure liquids and liquid solutions is a more complicated problem due to multiple interactions between molecules and nonperiodic structure of these phases. Thermodynamic properties of liquids and solutions are normally predicted with empirical models such as UNIFAC.[1] A bridge between empirical models and quantum chemistry was created when Klamt et al. proposed COSMO-RS (COnductor-like Screening Model for Real Solvents).[2,3]

The quantum chemical component of the model is based on the dielectric continuum model COSMO,[4] in which the solute molecule is placed in a molecule-shaped cavity, whereas the solvent is considered as continuum. The principal difference between COSMO and other dielectric continuum models is that the former uses a scaled-conductor dielectric boundary condition. For evaluation of thermodynamic properties, calculations are performed for the so-called "ideal" conductor with an infinite dielectric constant. This model is coupled with equations of statistical thermodynamics of fluids, providing a tool that enables one to evaluate thermodynamic properties of multicomponent systems and pure fluids.

In addition to the original COSMO-RS, alternative approaches are being developed. They include COSMO-RS(Ol),[5] COSMO-RS(ADF),[6] COSMO-SAC,[7] COSMO-3D,[8] and so forth. Detailed comparison of COSMO-RS, COSMO-RS(Ol), and COSMO-SAC[9,10] reported similar levels of performance among the considered methods.

Molecular parameters required for thermodynamic calculations include σ profile and molecular geometry. σ profile is the probability density function of polarization charge density on the molecular surface, as obtained from the results of quantum chemical computations. The final equations used to predict properties contain a number of empirical parameters. However, many fewer parameters are needed in comparison to the traditional empirical models. Approaches combining COSMO-RS(Ol) and COSMO-SAC with group-contribution schemes[11,12] and with UNIFAC[13] have also been proposed.

The quality of any model containing empirical parameters strongly depends on the reliability of the experimental data set used for parametrization. In this work, the experimental data used for reparameterization of COSMO-SAC were critically evaluated by the NIST ThermoData Engine (TDE),[14] an expert system for thermophysical and thermochemical property evaluation.[15] An expert system consists of an inference engine (algorithmic encoding of the thought process of a data expert) that interrogates a knowledge base (a trusted and comprehensive data archive of relevant facts) in order to develop new knowledge.[16,17] TDE's knowledge base includes the NIST/TRC SOURCE Data Archival System,[18,19] a database that contains the necessary data for pure compounds, binary mixtures, ternary mixtures, and chemical reactions, as derived from experimental measurements described in literature.

In this work, a σ-profile database was generated for 897 compounds, each containing up to eight non-hydrogen atoms, as well as the parameters for estimating thermodynamic properties

Special Issue: Memorial Issue in Honor of Anthony R. H. Goodwin

Received: June 11, 2015
Accepted: September 16, 2015

DOI: 10.1021/acs.jced.5b00483
J. Chem. Eng. Data XXXX, XXX, XXX−XXX

of binary liquid mixtures were determined with two COSMO-SAC type models. The parameters were fitted to critically evaluated data on equilibrium vapor pressure, activity coefficients, and excess enthalpies for binary liquid mixtures. Two different approaches (covariance matrix and Monte Carlo sampling) to uncertainty evaluation for the parameters and the predicted values were deployed.

2. CALCULATIONS

The quantum-chemical calculations were conducted in Gaussian 09 Revision C.01 software.[20] The calculations used the generalized York and Karplus conductor-like screening model,[21] which is based on the original COSMO model.[4] The solvent was a hypothetical water-like one with a static dielectric constant of 1000. The default atomic radii from Gaussian were used. Most of them were taken from ref 22. The radii for Si and Ar were 0.2457 nm and 0.2223 nm, respectively.

In COSMO calculations, B3LYP and BP functionals are normally coupled with various basis sets including 6-31G(2d,p), 6-311G(d,p), 6-31+G(d,p), TZP, TZVP, TZVP-DGA1, DNP, and so forth. Many functional/basis set combinations have been analyzed,[9] and none of them were found to have a clear advantage in the calculation of the activity coefficients. For the present data set, the total number of structures for which σ profiles were required exceeded 8000 and geometry optimization was the limiting step. Calculations at the B3LYP/6-311G(d,p) level of theory chosen in this work represented a compromise between the computational cost and the quality of predicted charge distributions. Geometry optimization was followed by a single-point calculation of point charges on molecular surfaces in an “ideal solvent” defined in ref 4. A similar procedure is normally used in COSMO-SAC,[7] whereas the original COSMO-RS uses the COSMO-optimized geometries.

For most molecules, σ ranges from about -3 e·nm^{-2} to 3 e·nm^{-2}. The considered range was divided into sections of 0.10 e·nm^{-2} width to create a σ profile. The probability density for each section was found as the area of all segments whose charge density was within ($\sigma \pm 0.05$) e·nm^{-2} divided by the total molecular surface area.

If a compound formed different conformers, the following procedure was applied. The initial structures of conformers were generated using various 3D structure generators and then optimized with MMFF94 force field[23] in the TINKER molecular mechanics package[24] using previously described[25] in-house procedures. The duplicate structures were removed. For each molecule, up to 20 conformers were retained for quantum-chemical calculations, then candidates were again filtered for duplicates. If a species had a chiral counterpart, it was also considered in the subsequent σ-profile generation. Thus, up to 40 conformers could be considered for one molecule.

The nonlinear optimization of model parameters was performed with use of the COBYLA nongradient method[26] from the NLOPT library.[27]

The standard errors of the parameters were estimated as square root of the corresponding diagonal element of the asymptotic covariance matrix[28]

$$\mathbf{V} = s^2(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1} \tag{1}$$

Here, s^2 is the squared standard deviation

$$s^2 = \frac{F_{\min}}{n - p} \tag{2}$$

where $F_{\min}$ is the minimal value of an objective function, n is the number of points, and p is the number of optimized parameters. An A_{ij} element of matrix $\mathbf{A}$ is a partial derivative of the property in the ith point with respect to the jth parameter.

The uncertainty for a prediction of property y was found with the equation[28]

$$u(y) = \sqrt{s_y^2 + \mathbf{a}^{\mathrm{T}}\mathbf{V}\mathbf{a}} \tag{3}$$

where $\mathbf{a}$ is a column matrix of partial derivatives of the calculated property value with respect to parameters. For the data set containing only one property, $s_y = s$. Because the data set under consideration included various properties, the standard deviation for property y was computed as

$$s_y = \sqrt{\frac{\sum_{i=1}^{m} (y_{i,\mathrm{calc}} - y_{i,\mathrm{exp}})^2}{m - p}} \tag{4}$$

Calculation of matrix $\mathbf{V}$ through partial derivatives of properties is based on a linear approximation, and the resulting uncertainties do not include the nonlinear effects. To assess this limitation, the Monte Carlo bootstrap method[29] was also applied to evaluate the uncertainties. In this method, a trial data set is generated by randomly substituting e^{-1} of points with other points from this data set. Thus, those other points would be represented twice. Optimal parameters are then determined for the trial data set. The required number of simulations is performed (more than 800 in this work), and the statistical characteristics are evaluated directly.

The bootstrap method allows one to avoid using the linear approximation altogether. However, in this case, the calculations become very time-consuming (a few minutes per one data point on a single CPU core). Therefore, the use of this method was limited by direct calculation of matrix $\mathbf{V}$. The uncertainties of prediction were calculated according to eqs 3 and 4.

3. THEORY

Two variations of the model were considered in this work. In the first case, the σ profile of a molecule is divided into the H-bonding and non-H-bonding parts. The H-bonding σ profile is formed by segments belonging to F, O, or N atoms, as well as H atoms attached to F, O, or N.[30] Following the COSMO-SAC formalism, the probability density was taken as proportional to the area of the corresponding segments and probability to form an H-bond, $p^{\mathrm{HB}}(\sigma)$. The latter is described by the equation[31]

$$p^{\mathrm{HB}}(\sigma) = 1 - \exp\left(-\frac{\sigma^2}{2\sigma_0^2}\right) \tag{5}$$

where $\sigma_0 = 0.7$ e·nm^{-2} is an empirical parameter.

This methodological variation is addressed as Model I for further discussion in this text.

In a more complex variation (Model II), the H-bonding σ profile is additionally divided into the OH H-bonding σ profile and the non-OH H-bonding σ profile.[32] As demonstrated below, the data on alcohol-containing systems constitute a significant portion of the available thermodynamic data for binary mixtures. Additional splitting of the H-bonding σ profile would therefore improve description of properties for such systems. The segments of atoms of all OH groups (including those in COOH, NOH, etc.) were included into the H-bonding σ profiles.

The surface charge densities were converted into σ profiles using the procedure similar to that by Lin[33] as described in ref 34. In calculation of the σ profile for a mixture, the contribution of each component was taken as proportional to its molecular surface.

If a compound formed several conformers, the average σ profile for the molecule was calculated for each temperature. The contribution of the ith conformer was assumed proportional to its mole fraction x_i. The ratio of the mole fractions of the ith and jth conformers was calculated as follows:

$$\frac{x_i}{x_j} = \frac{s_j}{s_i}\exp\left(-\frac{E_{\text{tot},i} - E_{\text{tot},j}}{kT}\right) \tag{6}$$

where $E_{\text{tot},i}$ and $E_{\text{tot},j}$ are the total energies of ith and jth conformers in the gas phase, s_i and s_j are the corresponding symmetry numbers, k is the Boltzmann constant.

The activity coefficient of the ith component in the solution (mix) is determined from the equation

$$\ln \gamma_{i/\text{mix}} = \frac{\mu_{i/\text{mix}}^{\text{res}} - \mu_{i/i}^{\text{res}}}{RT} + \ln \gamma_{i/\text{mix}}^{\text{SG}} \tag{7}$$

where μ^{res} is the residual part of pseudochemical potential, $\gamma_{i/\text{mix}}^{\text{SG}}$ is the combinatorial contribution to the activity coefficient calculated using Staverman–Guggenheim approximation[35] in the form

$$\ln \gamma_{i/\text{mix}}^{\text{SG}} = \ln\frac{\phi_i}{x_i} + 1 - \frac{\phi_i}{x_i} - \frac{z}{2}\cdot\frac{A_i}{A_{\text{norm}}}\left(\ln\left(\frac{\phi_i}{\theta_i}\right) + 1 - \frac{\phi_i}{\theta_i}\right) \tag{8}$$

Here

$$\phi_i = \frac{V_i}{\sum_j V_j x_j} \tag{9}$$

$$\theta_i = \frac{A_i}{\sum_j A_j x_j} \tag{10}$$

V_i and A_i are the volume and the surface area of the ith molecule, A_{norm} is the normalized (standard) segment surface area, x_i is the mole fraction of the ith component, and z is the coordination number assumed to be equal to 10. Following the procedure described in ref 7, A_{norm} was set to be 0.7953 nm^2.

The activity coefficients Γ for the segment with the charge density σ_m is

$$\Gamma_j^t(\sigma_m) = \left[\sum_s \sum_{\sigma_n} p_j^s(\sigma_n)\Gamma_j^s(\sigma_n)\exp\left(-\frac{\Delta W_{st}(\sigma_m, \sigma_n)}{RT}\right)\right]^{-1} \tag{11}$$

where j can be either pure liquid (i) or the solution (mix); s and t can correspond to the H-bonding (HB) and non-H-bonding contributions for Model I, or the OH H-bonding (OH), non-OH H-bonding (OT), and non-H-bonding contributions for Model II. The segment exchange energy ΔW_{st} is defined as[30]

$$\Delta W_{st}(\sigma_m, \sigma_n) = C_{\text{es}}(\sigma_m + \sigma_n)^2 - C_{\text{HB},st}(\sigma_m, \sigma_n)(\sigma_m - \sigma_n)^2 \tag{12}$$

where C_{es} and $C_{\text{HB},st}$ are empirical parameters.

For Model I

$$C_{\text{HB},st}(\sigma_m, \sigma_n) = \begin{cases} C_{\text{HB}} \text{ if } s = t = \text{HB},\ \sigma_m\cdot\sigma_n < 0 \\ 0 \end{cases} \tag{13}$$

For Model II

$$C_{\text{HB},st}(\sigma_m, \sigma_n) = \begin{cases} C_{\text{OH-OH}} \text{ if } s = t = \text{OH},\ \sigma_m\cdot\sigma_n < 0 \\ C_{\text{OH-OT}} \text{ if } (s = \text{OH},\ t = \text{OT}) \\ \qquad \text{or } (s = \text{OT},\ t = \text{OH}), \\ \qquad \sigma_m\cdot\sigma_n < 0 \\ C_{\text{OT-OT}} \text{ if } s = t = \text{OT},\ \sigma_m\cdot\sigma_n < 0 \\ 0 \end{cases} \tag{14}$$

Practical application of these equations includes an iterative solution of system of eqs 11 for all segments and types of σ profiles. In order to ensure the optimization process for the segment-wise activity coefficient was well conditioned, the following equation was used for iteration:

$$\Gamma_{j,k}^t = \frac{1}{2}(\Gamma_{j,k}^{t,*} + \Gamma_{j,k-1}^t) \tag{15}$$

where $\Gamma_{j,i}^{t,*}$ is calculated with eq 11, k is the iteration number.

The residual pseudochemical potential is calculated as

$$\frac{\mu_{i/j}}{RT} = \frac{A_i}{a_{\text{eff}}}\sum_s \sum_{\sigma_m} p_i^s(\sigma_m)\ln \Gamma_j^s(\sigma_m) \tag{16}$$

where the effective contact area a_{eff} is an empirical parameter.

The excess enthalpies were found numerically, with use of the equations

$$H_i^{\text{E}} = -RT^2\left(\frac{\partial \ln\gamma_i}{\partial T}\right)_p \tag{17}$$

$$H^{\text{E}} = x_1 H_1^{\text{E}} + (1 - x_1)H_2^{\text{E}} \tag{18}$$

4. RESULTS AND DISCUSSION

Training Set. For this study, the necessary experimental data were extracted from the NIST/TRC SOURCE Data Archival

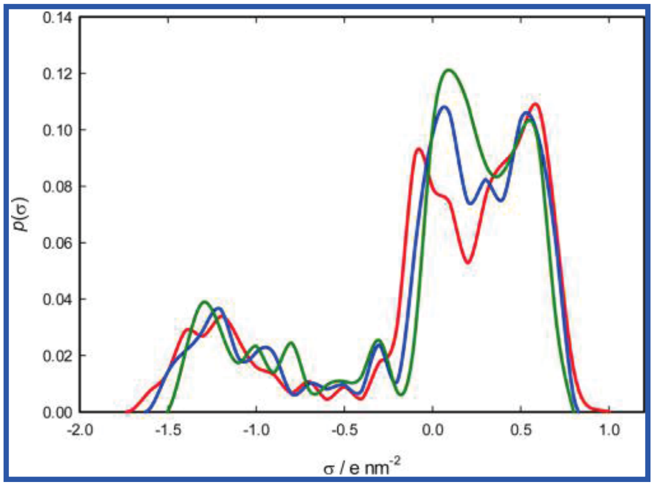


Figure 1. Total σ profile of acetone for r_{av} equal to (a) 0.040 nm (red), (b) 0.058 nm (blue), and (c) 0.079 nm (green).

DOI: 10.1021/acs.jced.5b00483

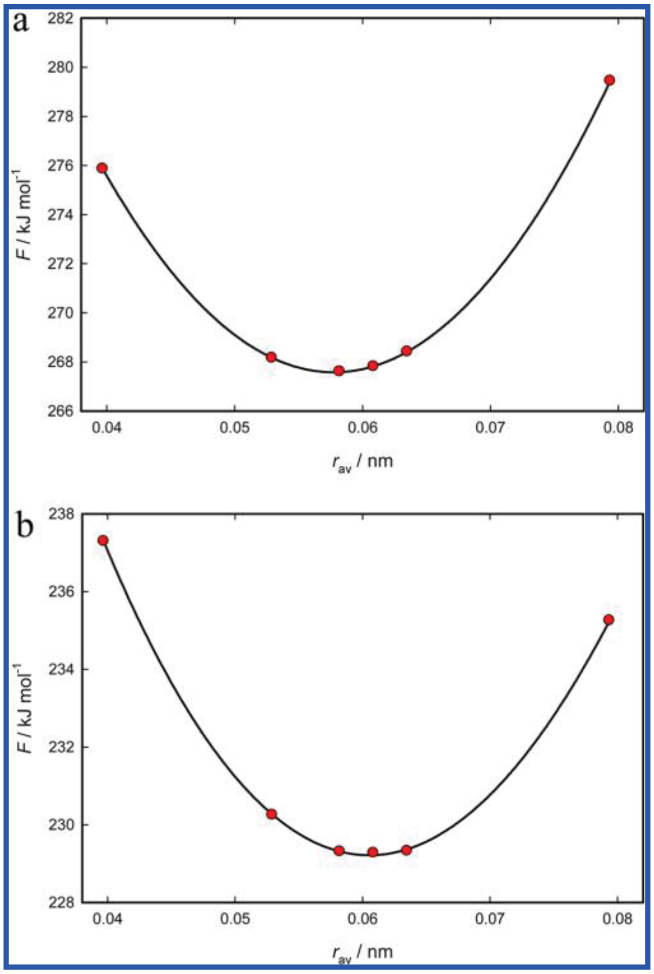


Figure 2. Dependence of the minimum values of the objective function on the averaging radius r_{av} for (a) Model I and (b) Model II.

System. As of May 2015, the archive contains about 380 000 data points on activity coefficients, excess enthalpy, and saturated vapor pressure for binary mixtures. The total number of compounds containing at least one data point of that type was near 1600. Given the substantial computational cost of thermodynamic COSMO calculations relative to empirical models like UNIFAC, efficient parameter evaluation required culling of the target data set. Guidelines on designing a reduction approach included the following considerations: some of the compounds have many more data points (water, benzene, ethanol, etc.) than the others, and COSMO models are primarily used for molecular systems, although the extension for electrolytes has been proposed[36] and an application to ionic liquids has been reported. The training set was limited to the compounds selected according to the following criteria: (i) the number of any binary experimental data points in the archive is ≥60; (ii) the compounds include only those atoms whose COSMO atomic radii could be calculated from literature data; (iii) the compounds are of molecular (and not ionic) nature; (iv) the number of non-hydrogen atoms in a molecule is ≤8. As a result, 897 compounds were selected, and a database of their temperature-dependent σ profiles was generated.

The quality of available data is not consistent across the database. Therefore, the following criteria were used to obtain the most reliable set of the experimental data:

(i) The experimental data on saturated vapor pressure over liquid solutions, excess enthalpy and activity coefficients in binary mixtures were limited to 897 compounds described above.

(ii) Data quality assessment procedures[37,38] were applied.

(iii) In vapor–liquid equilibria (VLE), only compounds with at least 25 pure-compound p_{sat} data points were considered.

(iv) The experimental VLE data were limited to $p_{sat} \leq 150$ kPa for both the mixtures and the pure components, and the estimated gas-phase fugacity coefficients were limited to 1.00 ± 0.05.

(v) The VLE data were considered only for $0.01 \leq x_1 \leq 0.99$.

(vi) The excess enthalpy and activity coefficient data were limited to atmospheric pressure (101.3 kPa).

(vii) The experimental data were sorted according to Chemical Abstract Service Registry Number (CASRN) of the components. One of every 500 data points of liquid properties and one in every 150 data points on saturated vapor pressure were included into the set for parameter optimization.

Criterion (iii) is required to ensure that the vapor pressures of pure compounds evaluated with TDE[14] are based on at least two sources with temperature-dependent p_{sat} values. Criterion (iv) is necessary to consider the gas-phase as ideal. Criterion (v) addresses conditions where a mole fraction of a component is

Table 1. Parameters of COSMO-SAC-Type Models[a,b]

parameter	Model I		Model II	
	NL	MC	NL	MC
z	10^{7}		10^{7}	
A_{norm}/nm^2	0.7953[7]		0.7953[7]	
$\sigma_0/\mathrm{e{\cdot}nm^{-2}}$	0.7[31]		0.7[31]	
	Optimized Parameters			
r_{av}/nm	0.058		0.061	
$C_{es}/\mathrm{kJ{\cdot}mol^{-1}{\cdot}nm^2{\cdot}e^{-2}}$	329.5 ± 8.7	328.9 ± 11.0	344.7 ± 8.5	333.5 ± 11.4
$C_{HB}/\mathrm{J{\cdot}mol^{-1}{\cdot}nm^4{\cdot}e^{-2}}$	955 ± 46	962 ± 71		
$C_{OH-OH}/\mathrm{J{\cdot}mol^{-1}{\cdot}nm^4{\cdot}e^{-2}}$			1309 ± 56	1343 ± 85
$C_{OH\text{-}OT}/\mathrm{J{\cdot}mol^{-1}{\cdot}nm^4{\cdot}e^{-2}}$			1057 ± 45	1102 ± 68
$C_{OT\text{-}OT}/\mathrm{J{\cdot}mol^{-1}{\cdot}nm^4{\cdot}e^{-2}}$			395 ± 73	498 ± 58
$10^2 a_{eff}/\mathrm{nm}^2$	7.80 ± 0.11	7.78 ± 0.19	8.09 ± 0.11	8.05 ± 0.20

[a] NL, nonlinear optimization + linear approximation in uncertainty estimation; MC, Monte Carlo. [b] The reported standard uncertainties were calculated as the standard errors for the corresponding coefficients.

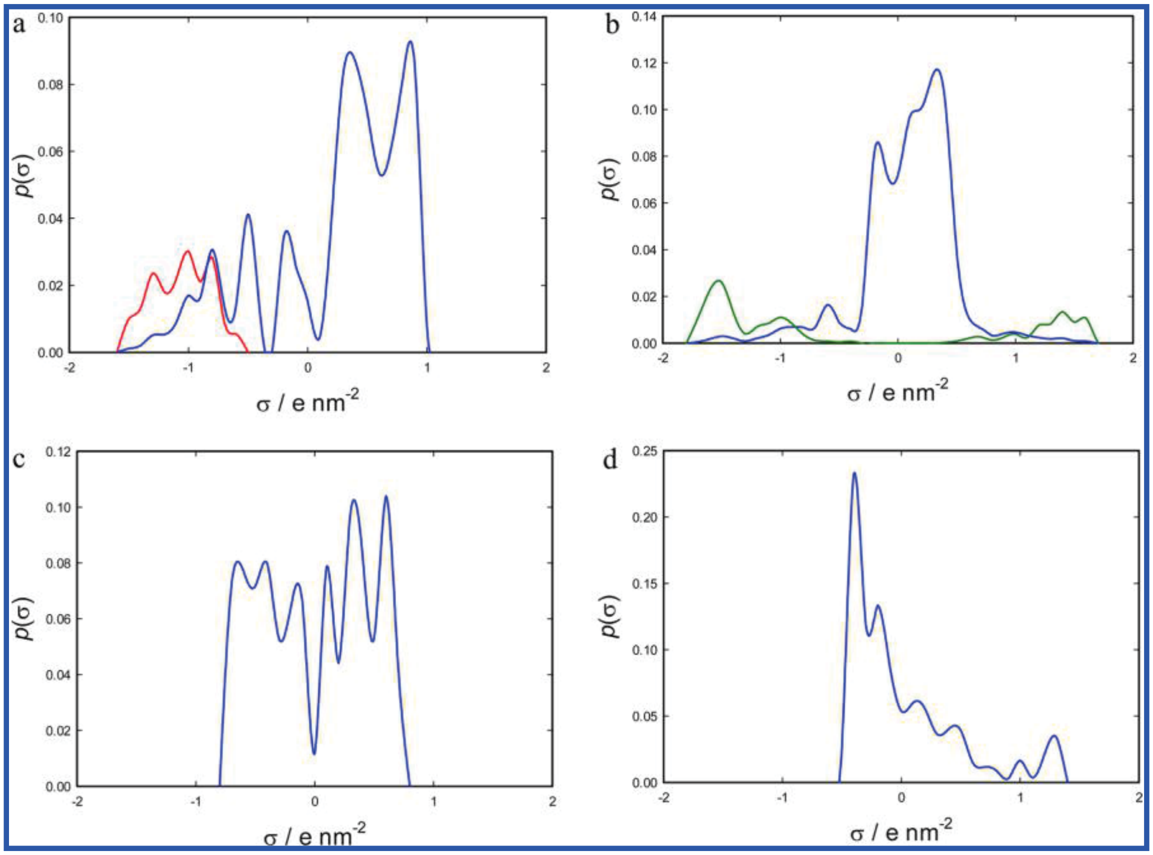


Figure 3. σ profiles ($r_{\text{av}} = 0.058$ nm) for (a) acetonitrile, (b) most stable conformer of ethanol, (c) benzene, and (d) trichloromethane: red, non-OH H-bonding σ profile; green, OH H-bonding σ profile; blue, non-H-bonding σ profile.

Table 2. Standard Uncertainties of Vapor Pressure over Liquid Binaries p, Excess Enthalpies H^E, and Activity Coefficients γ Obtained with Different Models

uncertainty	Model I	Model II
$u_r(p)$[a]	0.14	0.13
$u(H^E)$/kJ·mol^{-1}	0.62	0.57
$u(\ln\gamma)$	0.71	0.66

[a] Relative standard uncertainty.

very low, and the resulting pressure over the mixture will not significantly change compared to the pure-component limit. Criterion (vi) precludes the need to include pressure adjustments when a pure-component p_{sat} exceeds 1 bar. An empirical balance between VLE and liquid phase data defined by criterion (vii) corresponds to the relative importance of the corresponding data accepted in this work. The relative abundance of different compounds in the generated data set corresponds to the actual distribution of the experimental data.

The data set obtained as described above is presented in Supporting Information. It contains 1032 data points for the pressure over a solution, 262 data points for the excess enthalpy, and 12 data points for the activity coefficients. The total number of compounds covered by this set is 326. The temperature range considered is (15 to 457) K and the VLE pressure range considered is (0.3 to 132) kPa. The number of data points involving various compounds reflects the actual distribution of experimental efforts. The most abundant compound is water (152 points). Alcohols and hydroxy derivatives appear 647 times in the data set. Therefore, it is very important that the model provides adequate description of O−H hydrogen bonding.

Objective Function for Optimization of COSMO-SAC Parameters. Various forms of an objective function were considered. The use of relative deviations $[(p_{\text{sat},i,\text{calc}}/p_{\text{sat},i,\text{exp}}) - 1]$ or the similar in effect $RT\ln(p_{\text{sat},k,\text{calc}}/p_{\text{sat},k,\text{exp}})$ term allows one to avoid overestimation of importance for higher pressure data. Similar terms were used for the activity coefficients. The absolute deviation term $(H^E_{\text{m},i,\text{calc}} - H^E_{\text{m},i,\text{exp}})$ for the excess enthalpy was selected because the excess enthalpy can be close to zero; in this scenario, the relative deviation term will tend to overestimate the significance of the corresponding point. The uncertainty-based statistical weights for the solution were not applied since large sensitivity to assigned uncertainties appeared to lead to unphysical parameter sets. The objective function used for optimization had the form

$$\frac{F}{\text{kJ}\cdot\text{mol}^{-1}} = \sum_i \left(\frac{H^E_{\text{m},i,\text{calc}} - H^E_{\text{m},i,\text{exp}}}{\text{kJ}\cdot\text{mol}^{-1}}\right)^2 + \sum_j \left(\frac{R(T/\text{K})}{1000}\ln\frac{\gamma_{j,\text{calc}}}{\gamma_{j,\text{exp}}}\right)^2 + \sum_k \left(\frac{R(T/\text{K})}{1000}\ln\frac{p_{\text{sat},k,\text{calc}}}{p_{\text{sat},k,\text{exp}}}\right)^2 \tag{19}$$

where $R = 8.314462$ J·K^{-1}·mol^{-1}.

Radius for σ Profile Averaging. The charge densities for each segment obtained from quantum-chemical calculations were averaged as described in ref 3 to give the apparent charge densities

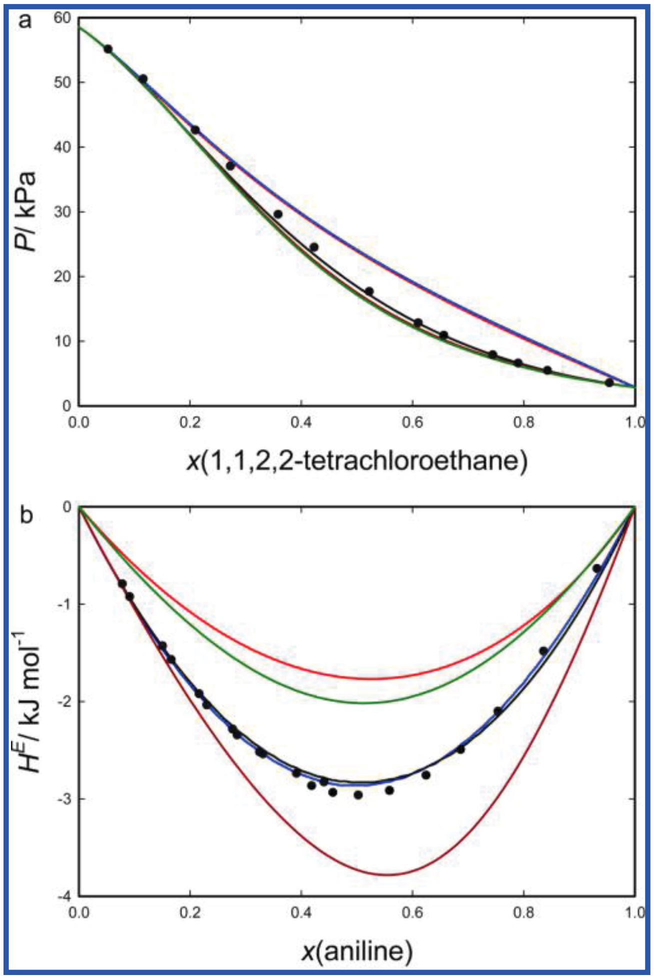


Figure 4. Comparison of experimental binary mixture properties with those predicted by different models. (a) Vapor pressure over 1,1,2,2-tetrachloroethane + tetrahydrofuran at T = 323.15 K and (b) excess enthalpies of aniline + *N,N*-dimethylformamide at T = 298.15 K: red, original UNIFAC;[40] blue, NIST-KT-UNIFAC;[42] black, NIST-modified UNIFAC;[41] brown, Model I; green, Model II. Experimental data on P are taken from ref 43 and on H^E from ref 44.

$$\sigma_m = \frac{\sum_n \sigma_n^* \frac{r_n^2 r_{\mathrm{av}}^2}{r_n^2 + r_{\mathrm{av}}^2} \exp\left(-\frac{d_{mn}^2}{r_n^2 + r_{\mathrm{av}}^2}\right)}{\sum_n \frac{r_n^2 r_{\mathrm{av}}^2}{r_n^2 + r_{\mathrm{av}}^2} \exp\left(-\frac{d_{mn}^2}{r_n^2 + r_{\mathrm{av}}^2}\right)} \qquad (20)$$

where σ^* and σ are charge densities before and after the averaging procedure, respectively; r_n is the radius for segment n; r_{av} is the averaging radius; d_{mn} is the distance between segments m and n. The optimal r_{av} values will be specific for different software packages and different levels of theory. The increase of r_{av} makes the σ profiles narrower (Figure 1). For low r_{av}, smoothing of charges is not as strong as for high r_{av} values. As a result, charge densities exceeding the limits of (−3 to 3) e·nm^{-2} applied for σ profiles in this work may remain on some segments.

The original COSMO-RS[3] uses r_{av} = 0.05 nm and (in addition to σ_m) a descriptor based on r_{av} = 0.10 nm. r_{av} is even lower in the COSMO-RS(ADF) version,[6] which uses similar equations for the residual pseudochemical potential. Historically in COSMO-SAC, a value of r_{av} close to 0.15 nm was applied and the exponential terms in eq 20 contained a decay factor equal to 3.57. Under normal circumstances $r_n \ll 0.15$ nm, and so the apparent r_{av} was close to 0.15 nm/(3.57)$^{1/2}$ ≈ 0.08 nm. Recent COSMO-SAC versions use eq 20 for averaging and r_{av} = 0.085 nm or 0.051 nm depending on the σ profiles used.[39] Because r_n is small compared to r_{av}, eq 20 can be simplified as follows:

$$\sigma_m = \frac{\sum_n \sigma_n^* r_n^2 \exp\left(-\frac{d_{mn}^2}{r_{\mathrm{av}}^2}\right)}{\sum_n r_n^2 \exp\left(-\frac{d_{mn}^2}{r_{\mathrm{av}}^2}\right)} \qquad (21)$$

The optimal F values obtained with σ profiles generated using eqs 20 and 21 did not differ significantly. In further calculations, eq 20 was used to keep consistency with previous implementations though eq 21 could be used as well.

The dependence of the objective functions on r_{av} was determined for the considered models (Figure 2). In both cases, the optimal r_{av} values were close to 0.06 nm (Table 1), which are in turn in agreement with the values used in the most recent implementations of both COSMO-SAC and COSMO-RS. It should also be noted that the minimum values of the objective functions only slightly changed for r_{av} differing from the optimal value by ± 0.005 nm. σ profiles for some compounds are presented in Figure 3.

Parameter Optimization. The optimized parameters for Model I and Model II along with the corresponding uncertainties are presented in Table 1. With one exception, the uncertainties from Monte Carlo are higher relative to those calculated using a linear approximation by a factor of 1.2 to 1.9. At the same time, the mathematical background of the Monte Carlo method is less constrained by various approximations. The parameters obtained with two different approaches agree within their uncertainties. This fact confirms that the direct nonlinear optimization of the objective function (19) can be used in this case. This is important because the nonlinear optimization is more computationally intensive than the Monte Carlo approach.

Evaluation of Uncertainties. To evaluate the uncertainties in the predicted values, one needs to know s_y in eq 3 for each corresponding property. These values were estimated based upon the target training set. The second term in eq 3 caused by the parameter uncertainty was typically 2 orders of magnitude smaller than s_y^2. Thus, the resulting uncertainties were virtually independent of the applied mathematical method (Table 2) and were close for all points within the same property for the considered T and P ranges. The uncertainties are defined by s_y, which characterizes scatter of the calculated data relative to the experimental values. Because the typical experimental uncertainties for the considered properties are significantly lower than the values in Table 2, one can conclude that these uncertainties are due to the nature of the models. Further increase in the size of training set should not improve the quality of the model.

As follows from the results in Table 2, Model II outperforms Model I for all of the considered properties. The calculations were also compared with those from the original UNIFAC model[40] in terms of the weight-averaged root-mean-square error (r.m.s.) as described in ref 41; at the same time, performance of the latter is somewhat less accurate than that of NIST-KT-UNIFAC[42] by 15% and that of NIST-Modified UNIFAC[41] by 44%. Model II demonstrated comparable results with the original UNIFAC model, and Model I had about 10% larger r.m.s.

Some representative examples comparing the results from various methods are presented in Figure 4. Supporting Information for this article provides comparison of original experimental data from 1067 sources for 1089 binary mixtures and predicted values by deployment of Model I and Model II described here for vapor pressure (Table S1), excess enthalpy (Table S2), and activity coefficients (Table S3).

5. CONCLUSION

Two sets of parameters for COSMO-SAC implementation were derived with the use of critically evaluated experimental data points on saturated vapor pressure, excess enthalpy, and activity coefficients for 1089 binary mixtures available in the NIST/TRC SOURCE Data Archival System. The uncertainty of the implemented COSMO-SAC version was evaluated using this data set. It was demonstrated that the uncertainty of prediction is limited by the nature of the model.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jced.5b00483.

> Tables of deviations of the calculated vapor pressures, excess enthalpies, and activity coefficients from the experimental values. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: yauheni.paulechka@nist.gov.

Notes

The authors declare no competing financial interest.
This is a contribution of the U.S. National Institute of Standards and Technology and not subject to copyright in the United States. Trade names are provided only to specify procedures adequately and do not imply endorsement by the National Institute of Standards and Technology. Similar products by other manufacturers may be found to work as well or better.

■ REFERENCES

(1) 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*, 2352−2355.

(2) 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.

(3) Klamt, A.; Jonas, V.; Burger, T.; Lohrenz, J. C. W. Refinement and Parametrization of COSMO-RS. *J. Phys. Chem. A* **1998**, *102*, 5074−5085.

(4) Klamt, A.; Schüürmann, 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**, 799−805.

(5) Grensemann, H.; Gmehling, J. Performance of a Conductor-like Screening Model for Real Solvents in Comparison to Classical Group Contribution Methods. *Ind. Eng. Chem. Res.* **2005**, *44*, 1610−1624.

(6) Pye, C. E.; Ziegler, T.; van Lenthe, E.; Louwen, J. N. An implementation of the conductor-like screening model of solvation within the Amsterdam density functional package − Part II. COSMO for real solvents. *Can. J. Chem.* **2009**, *87*, 790−797.

(7) 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.

(8) Gutierrez-Sevillano, J. J.; Leonhard, K.; van der Eerden, J. P. J. M.; Vlugt, T. J. H.; Krooshof, G. J. P. COSMO-3D: Incorporating Three-Dimensional Contact Information into the COSMO-SAC Model. *Ind. Eng. Chem. Res.* **2015**, *54*, 2214−2226.

(9) 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.

(10) Constantinescu, D.; Rarey, J.; Gmehling, J. Application of COSMO-RS Type Models to the Prediction of Excess Enthalpies. *Ind. Eng. Chem. Res.* **2009**, *48*, 8710−8725.

(11) Mu, T.; Rarey, J.; Gmehling, J. Group contribution prediction of surface charge density profiles for COSMO-RS(Ol). *AIChE J.* **2007**, *53*, 3231−3240.

(12) Mu, T.; Rarey, J.; G, J. Group contribution prediction of surface charge density distribution of molecules for COSMO-SAC. *AIChE J.* **2009**, *55*, 3298−3300.

(13) Jakob, A.; Grensemann, H.; Lohmann, J.; Gmehling, J. Further Development of Modified UNIFAC (Dortmund): Revision and Extension 5. *Ind. Eng. Chem. Res.* **2006**, *45*, 7924−7933.

(14) Frenkel, M.; Chirico, R. D.; Diky, V.; Kroenlein, K.; Muzny, C. D.; Kazakov, A. F.; Magee, J. W.; Abdulagatov, I. M.; Lemmon, E. W. *NIST ThermoData Engine, NIST Standard Reference Database 103b - Pure Compounds, Binary Mixtures, and Chemical Reactions, version 9.0; Standard Reference Data Program*; National Institute of Standards and Technology: Gaithersburg, MD, 2014.

(15) Frenkel, M.; Chirico, R. D.; Diky, V.; Xinjian, Y.; Qian, D.; Muzny, C. ThermoData Engine (TDE): Software Implementation of the Dynamic Data Evaluation Concept. *J. Chem. Inf. Model.* **2005**, *45*, 816−838.

(16) Frenkel, M. Global Information Systems in Science: Application to the Field of Thermodynamics. *J. Chem. Eng. Data* **2009**, *54*, 2411−2428.

(17) Frenkel, M. A Never-Ending Search for the Truth: Thermodynamics in the Uncertain Era of the Internet. *J. Chem. Thermodyn.* **2015**, *84*, 18−40.

(18) Frenkel, M.; Dong, Q.; Wilhoit, R. C.; Hall, K. R. TRC SOURCE Database: A Unique Tool for Automatic Production of Data Compilations. *Int. J. Thermophys.* **2001**, *22*, 215−226.

(19) Kazakov, A.; Muzny, C. D.; Kroenlein, K.; Diky, V.; Chirico, R. D.; Magee, J. W.; Abdulagatov, I. M.; Frenkel, M. NIST/TRC SOURCE Data Archival System: The Next-Generation Data Model for Storage of Thermophysical Properties. *Int. J. Thermophys.* **2012**, *33*, 22−33.

(20) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenb *Gaussian 09*, Revision C.01; Gaussian, Inc.: Wallingford, CT, 2009.

(21) York, D. M.; Karplus, M. A Smooth Solvation Potential Based on the Conductor-Like Screening Model. *J. Phys. Chem. A* **1999**, *103*, 11060−11079.

(22) Klamt, A. *COSMO-RS: From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*; Elsevier: Amsterdam, 2005.

(23) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(24) Ren, P.; Wu, C.; Ponder, J. W. Polarizable Atomic Multipole-based Molecular Mechanics for Organic Molecules. *J. Chem. Theory Comput.* **2011**, *7*, 3143−3161.

(25) Carande, W. H.; Kazakov, A.; Muzny, C.; Frenkel, M. Quantitative Structure−Property Relationship Predictions of Critical Properties and Acentric Factors for Pure Compounds. *J. Chem. Eng. Data* **2015**, *60*, 1377−1387.

(26) Powell, M. J. D. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in Optimization and Numerical Analysis*; Hennart, S. G. a. J.-P., Ed.; Kluwer Academic: Dordrecht, 1994; pp 51−67.

(27) Johnson, S. G. The NLopt nonlinear-optimization package. http://ab-initio.mit.edu/nlopt (accessed April, 2015).

(28) Ruckstuhl, A. Introduction to Nonlinear Regression, 2010. http://stat.ethz.ch/~stahel/courses/cheming/nlreg10E.pdf (accessed April 26, 2015).

(29) Efron, B. *The Jacknife, the Bootstrap, and Other Resampling Plans*; Stanford University: Stanford, CA, 1980.

(30) Lin, S. T.; Chang, J.; Wang, S.; Goddard, W. A., III; Sandler, S. A. Prediction of Vapor Pressures and Enthalpies of Vaporization Using a COSMO Solvation Model. *J. Phys. Chem. A* **2004**, *108*, 7429−7288.

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

(32) Hsiech, 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.

(33) Lin, S.-T. *Quantum Mechanical Approaches to the Prediction of Phase Equilibria: Solvation Thermodynamics and Group Contribution Methods*; Unversity of Delaware: Newark, 2000.

(34) VT Sigma Profile Databases. http://www.design.che.vt.edu/COSMO/Program_Files/Sigma-average_v2.txt (accessed September 8, 2015).

(35) Staverman, A. J. The Entropy of High Polymer Solutions. *Rec. Trav. Chim. Pays-Bas* **1950**, *69*, 163−174.

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

(37) Kang, J. W.; Diky, V.; Chirico, R. D.; Magee, J. W.; Muzny, C. D.; Abdulagatov, I.; Kazakov, A. F.; Frenkel, M. Quality Assessment Algorithm for Vapor-Liquid Equilibrium Data. *J. Chem. Eng. Data* **2010**, *55*, 3631−3640.

(38) Kang, J. W.; Diky, V.; Chirico, R. D.; Magee, J. W.; Muzny, C. D.; Kazakov, A. F.; Kroenlein, K.; Frenkel, M. Algorithmic Framework for Quality Assessment of Phase Equilibrium Data. *J. Chem. Eng. Data* **2014**, *59*, 2283−2293.

(39) 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.

(40) Poling, B. E.; Praunitz, J. M.; O'Connell, J. P. *The Properties of Gases and Liquids*, 5th ed.; McGraw-Hill: New York, 2001.

(41) Kang, J. W.; Diky, V.; Frenkel, M. New modified UNIFAC parameters using critically evaluated phase equilibrium data. *Fluid Phase Equilib.* **2015**, *388*, 128−141.

(42) Kang, J. W.; Diky, V. V.; Chirico, R. D.; Magee, J. W.; Muzny, C. D.; Abdulagatov, I.; Kazakov, A. F.; Frenkel, M. A new method for evaluation of UNIFAC interaction parameters. *Fluid Phase Equilib.* **2011**, *309*, 68−75.

(43) Garriga, R.; Perez, P.; Gracia, M. Isothermal (vapour + liquid) equilibrium for binary mixtures of (tetrahydrofuran + 1,1,2,2-tetrachloroethane or tetrachloroethene) at nine temperatures. *J. Chem. Thermodyn.* **2006**, *38*, 348−358.

(44) Ramadevi, R. S.; Venkatesu, P.; Rao, M. V. P.; Krishna, M. R. Excess enthalpies of binary mixtures of N,N-dimethylformamide with substituted benzenes at 298.15 K. *Fluid Phase Equilib.* **1996**, *114*, 189−197.

DOI: 10.1021/acs.jced.5b00483