I&EC research
Industrial & Engineering Chemistry Research

Article

pubs.acs.org/IECR

An Improvement to COSMO-SAC for Predicting Thermodynamic Properties

Ruichang Xiong, Stanley I. Sandler,* and Russell I. Burnett

Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware 19716, United States

Supporting Information

ABSTRACT: A modified COSMO-SAC model is presented to calculate thermodynamic properties of pure fluids and mixtures using statistical thermodynamics and the surface charge density of each compound obtained from a quantum mechanics (QM) calculation. The main differences from the previous models are that the new model includes a dispersion contribution in the mixture interaction, and is reparametrized using measured pure and mixture thermodynamic data simultaneously. With a single set of universal parameters, the new model provides higher accuracy than our previous models for predicting mixture thermodynamic properties while maintaining the same accuracy for pure compound thermodynamic properties. The overall root-mean-square deviation (RMSD) in the logarithms of partition coefficients for 992 octanol−water partitioning systems and 829 other solvent−water partitioning systems with this new model is reduced by about 10% compared to the results from previous models. Also, the agreement between the predicted and measured partition coefficients over a wide range of values is improved as a result of better activity coefficient predictions at high dilution by inclusion of the dispersive mixture interaction in the model. The accuracy in the vapor−liquid equilibrium (VLE) predictions is comparable to, or better than, the previous model that was developed for phase equilibria calculations only. The new model also provides parameters for use with the Amsterdam Density Functional (ADF) in addition to DMol[3].

1. INTRODUCTION

There is significant interest in the properties of liquids and liquid mixtures in the chemical industry. Molecules in the liquid phase are in relatively close proximity to one another but lack any repeatable long-range structure. Consequently, all the possible interactions must be accounted for in order to accurately model their behavior. This makes the accurate prediction of liquid-phase properties over the composition range difficult, particularly when there are multiple components.

Activity coefficient models, or excess Gibbs free energy models, are typically used in modeling liquid-phase nonidealities in correlating and predicting vapor−liquid, liquid−liquid, and solid−liquid equilibria. The successful semipredictive excess Gibbs free energy models are UNIFAC[1] and modified UNIFAC[2] and they are based on group contribution methods (GCMs), in which a molecule is described by a collection of independent functional groups and a mixture is formed from these functional groups. However, such GCMs require a very large set of empirical parameters for the interactions between all the defined functional groups. Thus, their accuracy is determined by how well the molecules in the mixture of interest are described by the defined functional groups in the model, and how similar the environments and interactions between these functional groups are to those used in the parametrization. Also such group contribution models must be used with caution when the parameters between the functional groups are unavailable, or for groups that have not previously been defined. Consequently, predicting values of the activity coefficients in arbitrary mixtures still remains a challenge.

In addition to GCMs, the quantitative structure−property relationship (QSPR) models[3] and explicit solvent models in molecular simulation[4] have been used to calculate the thermodynamic properties. The QSPR model is an empirical approach in which some thermodynamic properties (generally pure component properties[3] such as vapor pressure) are empirically correlated with descriptors based on molecular structure. The explicit solvent class of models treats the solute and solvent as molecules and uses molecular simulation[4] such as Monte Carlo (MC) and molecular dynamics (MD) to calculate the properties. Such simulations require empirical or semiempirical force fields to describe the interactions between molecules, and they require long computation times that increase dramatically with the size of molecules. This method is especially expensive in the case of large flexible molecules and whenever conformers of significantly different (electronic) structure are involved.

Quantum-based solvation models provide an alternative means of predicting activity coefficients and other thermodynamic properties. In these methods, quantum mechanics (QM) is used to describe the solute with an effective Hamiltonian that is the sum of an electronic Hamiltonian in a vacuum and an interaction potential between the solute and the solvent. These models do not use arbitrarily defined functional groups, and are thus less dependent on fitting to experimental data. There are various quantum-based solvation models available.[5] Of interest is the COSMO-RS (conductor-like screening model for real

Received: December 28, 2013
Revised: March 28, 2014
Accepted: April 8, 2014
Published: April 8, 2014

dx.doi.org/10.1021/ie404410v | *Ind. Eng. Chem. Res.* 2014, 53, 8265−8278

solvents) model developed by Klamt and co-workers,[6,7] which provides a new path for computing solvation energies. In COSMO-RS, molecules are treated as a collection of surface segments. A solute molecule is moved from a vacuum to a perfect conductor, where the interaction energies between segments are determined from a COSMO calculation, i.e., a solvation calculation for the molecule in the conductor. Then the solute is transferred from the perfect conductor to the real solvent, and for this process the chemical potential of each segment is self-consistently determined from statistical mechanics. The chemical potential of each molecule is then obtained by summing the contributions of the individual segments. In this way, thermodynamic properties of liquids are predicted in an a priori manner.

On the basis of the framework of COSMO-RS, Lin and Sandler[8] suggested a variation, the COSMO-SAC (where SAC denotes segment activity coefficient) model by invoking a necessary thermodynamic consistency criterion. Although there are differences, COSMO-RS and COSMO-SAC share some similarities in the calculation of solvation free energy. The first version of the COSMO-SAC model[8] was developed to predict activity coefficients. Then a later model[9] was proposed to also predict the vapor pressures and heats of vaporization by including a dispersion term using mean field theory and a cavity term based on thermodynamic perturbation theory (TPT). To improve the COSMO-SAC method, Wang and Sandler[10] presented a refined model (denoted as COSMO-SAC (2007)) by combining previous approaches to predict both pure and mixture thermodynamic properties. Later Hsieh et al.[11] proposed COSMO-SAC (2010) for improving phase equilibria.

The objective here is to present a further improved COSMO-SAC model for the simultaneous prediction of both pure fluid and liquid mixture properties that provides higher accuracy. The paper is organized as follows: in section 2 we briefly review the last two refined COSMO-SAC models and present the modifications made here. The computational details are presented in section 3. The new parametrization procedure is presented in section 4. The results are then presented and discussed in section 5. Finally, the concluding remarks are given in section 6.

2. THEORY

2.1. COSMO-SAC Model. The COSMO-SAC approach is based on QM/COSMO solvation calculation followed by a self-consistent statistical thermodynamic analysis to determine the chemical potential of a species in a pure fluid or a mixture. The complete theory of COSMO-SAC has been presented in previous publications.[8−11] Here the major parts of the model are reviewed to understand the changes that have been made.

The model gives an expression for the free energy of solvation of solute i in the solution S, $\Delta G_{i/S}^{*\mathrm{sol}}$; this consists of an electrostatic contribution, as a result of the electrostatic interactions and the polarization between the solute and solvent, and a van der Waals part that includes cavity formation and dispersion contributions. The activity coefficient $\gamma_{i/S}$ of solute i in the solution S quantifies the deviation between the chemical potential of species i in nonideal and ideal liquid mixtures.[12] In previous COSMO-SAC models[8,10,11] it was assumed that the mixture nonideality contribution from the dispersive interactions was negligible, so that $\gamma_{i/S}$ was calculated from

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

where $\ln \gamma_{i/S}^{\mathrm{res}} = [(\Delta G_{i/S}^{*\mathrm{res}} - \Delta G_{i/i}^{*\mathrm{res}})/RT]$ is the contribution from the restoring solvation free energy ($\Delta G^{*\mathrm{res}}$), which is the result of the difference in the electrostatic interactions between the molecule in the solvent ($\Delta G_{i/S}^{*\mathrm{res}}$) and in its own pure fluid ($\Delta G_{i/i}^{*\mathrm{res}}$), and $\ln(\gamma_{i/S}^{\mathrm{comb}})$ is the combinatorial or cavity formation contribution that accounts for molecular size and shape differences between species. The Staverman−Guggenheim (SG) combinatorial term[13,14] was used to calculate $\ln(\gamma_{i/S}^{\mathrm{comb}})$:

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

with $\theta_i = (x_i q_i)/(\sum_j x_j q_j)$, $\phi_i = (x_i r_i)/(\sum_j x_j r_j)$, and $l_i = (z/2)(r_i - q_i) - (r_i - 1)$, where x_i is the mole fraction of component i; r_i and q_i are the normalized volume and surface area parameters for species i; and z is the coordination number, taken to be 10.

The restoring free energy of a solute molecule i from the ideal conductor to the real solvent is obtained as a summation over the segment activity coefficients (SACs)[8]

$$\frac{\Delta G_{i/S}^{*\mathrm{res}}}{RT} = n\sum_{\sigma_m} p(\sigma_m)\ln\Gamma_S(\sigma_m) \tag{3}$$

where $n = A_i/a_{\mathrm{eff}}$ is the number of segments in the molecule, which is the ratio of the total surface area of a single molecule i (A_i) to the area of the standard surface of segments (a_{eff}); $p(\sigma)$ is a two-dimensional histogram produced from the three-dimensional geometric charge density (σ) distribution known as the σ-profile. Because the pairwise interaction between segments is based on contact surfaces of the identical size, the surface charge density distribution from the COSMO output needs to be averaged to find an effective surface charge for each segment of uniform size. The following expression was used for charge averaging in COSMO-SAC (2007)[10] and COSMO-SAC (2010)[11] models.

$$\sigma_\mu = \frac{\sum_\nu \sigma_\nu^* \left(\frac{r_\nu^2 r_{\mathrm{eff}}^2}{r_\nu^2 + r_{\mathrm{eff}}^2}\right) \exp\left[-f_{\mathrm{decay}}\left(\frac{d_{\mu\nu}^2}{r_\nu^2 + r_{\mathrm{eff}}^2}\right)\right]}{\sum_\nu \left(\frac{r_\nu^2 r_{\mathrm{eff}}^2}{r_\nu^2 + r_{\mathrm{eff}}^2}\right) \exp\left[-f_{\mathrm{decay}}\left(\frac{d_{\mu\nu}^2}{r_\nu^2 + r_{\mathrm{eff}}^2}\right)\right]} \tag{4}$$

where σ^* and σ are charge densities before and after the charge averaging process; σ_ν is the charge density of a segment ν; $r_\nu = (a_\nu/\pi)^{1/2}$ is the radius of a segment ν (a_ν is the surface area of the segment ν); $r_{\mathrm{eff}} = (a_{\mathrm{eff}}/\pi)^{1/2}$ is the radius of a standard surface segment. The empirical parameter f_{decay} has been set to 3.57[10,11] and $d_{\mu\nu}$ is the distance between segments μ and ν.

The σ-profile, $p(\sigma)$, is the probability of finding a surface segment with a charge density σ, and is defined as

$$p_i(\sigma) = \frac{A_i(\sigma)}{A_i} \tag{5}$$

where $A_i(\sigma)$ is the surface area with that charge density. For a mixture, the σ-profile is taken to be the linear combination of the σ-profiles for each of the pure species weighted with the product of their mole fractions (x_i) and surface areas (A_i), i.e.

$$p_S(\sigma) = \frac{\sum_i x_i A_i p_i(\sigma)}{\sum_i x_i A_i} \tag{6}$$

To account for the effect of hydrogen-bonding (HB) interactions, the total σ-profile is separated into two

dx.doi.org/10.1021/ie404410v | Ind. Eng. Chem. Res. 2014, 53, 8265−8278

components: one of HB atoms (an N, O, F, or H atom connected to an N, O, or F atom); and the other from non-hydrogen bonding (non-HB) atoms. A Gaussian-like function is used to weight the probability of an HB segment forming a hydrogen bond

$$P^{\mathrm{hb}}(\sigma) = 1 - \exp\left(-\frac{\sigma^2}{2{\sigma_0}^2}\right) \tag{7}$$

where $\sigma_0 = 0.007$ e/Å^2 was chosen for the shape of the Gaussian probability distribution.[10,11]

The segment activity coefficient (SAC) of segment m with a charge density of σ_m is calculated from

$$\ln \Gamma_S^t(\sigma_m^t) = -\ln\left\{\sum_s^{\mathrm{nhb,hb}} \sum_{\sigma_n} p_S^s(\sigma_n^s)\Gamma_S^s(\sigma_n^s) \exp\left(\frac{-\Delta W(\sigma_m^t, \sigma_n^s)}{RT}\right)\right\} \tag{8}$$

where superscripts s and t can be either HB or non-HB, representing the contribution from an HB or a non-HB segment. The segment interaction ΔW is taken to be of the following form:

$$\Delta W(\sigma_m^t, \sigma_n^s) = c_{\mathrm{es}}(\sigma_m^t + \sigma_n^s)^2 - c_{\mathrm{hb}}(\sigma_m^t, \sigma_n^s)(\sigma_m^t - \sigma_n^s)^2 \tag{9}$$

In COSMO-SAC (2007),[10] $c_{\mathrm{es}} = f_{\mathrm{pol}}(0.3{a_{\mathrm{eff}}}^{3/2}/2\varepsilon_0)$ is a constant obtained from a_{eff}, the polarization factor f_{pol} and the permittivity of vacuum ε_0, and $c_{\mathrm{hb}}(\sigma_m^t, \sigma_n^s)$ is the following temperature-independent parameter:

$$c_{\mathrm{hb}}(\sigma_m^t, \sigma_n^s) = \begin{cases} c_{\mathrm{hb}} & \text{if } s = t = \mathrm{hb}, \text{ and } \sigma_m^t\sigma_n^s < 0 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

In COSMO-SAC (2010),[11] $c_{\mathrm{es}} = A_{\mathrm{ES}} + (B_{\mathrm{ES}}/T^2)$ is a temperature-dependent parameter based on two universal constants (A_{ES} and B_{ES}), and $c_{\mathrm{hb}}(\sigma_m^t, \sigma_n^s)$ is defined as

$$c_{\mathrm{hb}}(\sigma_m^t, \sigma_n^s) = \begin{cases} c_{\mathrm{OH-OH}} & \text{if } s = t = \mathrm{OH}, \text{ and } \sigma_m^t\sigma_n^s < 0 \\ c_{\mathrm{OT-OT}} & \text{if } s = t = \mathrm{OT}, \text{ and } \sigma_m^t\sigma_n^s < 0 \\ c_{\mathrm{OH-OT}} & \text{if } s = \mathrm{OH},\ t = \mathrm{OT}, \text{ and } \sigma_m^t\sigma_n^s < 0 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

which divides the σ-profile into three components: contributions from surfaces of non-HB atoms, contributions from surfaces of the hydroxyl (OH) group, and contributions from surfaces of all other types of HB donating and accepting atoms (denoted as OT) so that different hydrogen bonds formed by the substances containing hydroxyl groups, amine groups, and nitro groups can be differentiated.

2.2. Modified Model. *2.2.1. Separating the Charge Averaging from the Electrostatics.* The empirical parameter a_{eff} (the standard segment surface area) appears in two separate places in the previous models:[10,11] the sum of the segment chemical potentials to obtain $\Delta G_{i/S}^{*\mathrm{res}}$ (eq 3) and the charge density averaging (eq 4). From a numerical standpoint, this use of a_{eff} in different places is convenient, as it reduces the number of required adjustable parameters. However, it is not related to the basis of the model, as the use of a_{eff} in eq 4 is based on empiricism, not on the result of any theory. Also f_{decay} in eq 4 is purely an empirical parameter.

Therefore, here eq 4 has been simplified to

$$\sigma_\mu = \frac{\sum_\nu \sigma_\nu^* \left(\frac{{r_\nu}^2{r_{\mathrm{avg}}}^2}{{r_\nu}^2 + {r_{\mathrm{avg}}}^2}\right) \exp\left(\frac{-{d_{\mu\nu}}^2}{{r_\nu}^2 + {r_{\mathrm{avg}}}^2}\right)}{\sum_\nu \left(\frac{{r_\nu}^2{r_{\mathrm{avg}}}^2}{{r_\nu}^2 + {r_{\mathrm{avg}}}^2}\right) \exp\left(\frac{-{d_{\mu\nu}}^2}{{r_\nu}^2 + {r_{\mathrm{avg}}}^2}\right)} \tag{12}$$

Also, c_{es} in eq 9 is used as a temperature-independent universal parameter in this work; c_{hb} is a temperature-independent parameter with the same definition (eq 11) used in the COSMO-SAC (2010).[11] The result of this change is that a_{eff} is the only adjustable parameter in eq 3, c_{es} is the only adjustable parameter in eq 9 for non-HB compounds, and r_{avg} is the only adjustable parameter in eq 12. For non-HB compounds, the COSMO-SAC (2007) model has two parameters: f_{decay} and a_{eff}. The COSMO-SAC (2010) model has four parameters: f_{decay} a_{eff}, A_{ES}, and B_{ES}. Separating the charge averaging from the electrostatics, there are three parameters here: r_{avg}, a_{eff}, and c_{es}. Note that eq 12 is then the same equation as used in the published version of COSMO-RS.[7] The use of these same three adjustable parameters was also adopted by Gmehling and co-workers[15−17] in their COSMO-RS(Ol) model, a slight variation of COSMO-RS.

2.2.2. Dispersion in Mixtures. All previous COSMO-SAC models[10,11] assumed that the effect of dispersion would cancel in the calculation of mixture thermodynamic properties. The justification for this was that the magnitude of these interactions would likely be similar for the species in its pure liquid and in the liquid mixture, so that their difference would be negligible when compared to the other contributions. The accuracy of this assumption was verified using only a very limited data set.[10] However, this assumption has now been found not to be valid for all species at any composition in the mixture, which will be discussed in detail later. Consequently, in this work, $\gamma_{i/S}$ is calculated by also including the dispersive interaction contribution,

$$\ln(\gamma_{i/S}) = \ln(\gamma_{i/S}^{\mathrm{res}}) + \ln(\gamma_{i/S}^{\mathrm{disp}}) + \ln(\gamma_{i/S}^{\mathrm{comb}}) \tag{13}$$

where $\ln \gamma_{i/S}^{\mathrm{disp}} = (\Delta G_{i/S}^{*\mathrm{disp}} - \Delta G_{i/i}^{*\mathrm{disp}})/RT$ is the contribution from the dispersive interaction.

The dispersion free energy term in a mixture is obtained using a first-order mean field approximation and expressed in terms of the Helmholtz energy:

$$\frac{\Delta \underline{A}_{i/S}^{*\mathrm{disp}}}{RT} = \frac{1}{RT\underline{V}_{\mathrm{mix}}}\left(\frac{\underline{V}_{i/L}}{\underline{V}_{\mathrm{mix}}} \sum_j^{n_c}\sum_k^{n_c}\sum_\tau^{n_a}\sum_\omega^{n_a} x_j x_k m_\tau^j m_\omega^k \varepsilon_{\tau\omega} - 2\sum_j^{n_c}\sum_\tau^{n_a}\sum_\omega^{n_a} x_j m_\tau^i m_\omega^j \varepsilon_{\tau\omega}\right) \tag{14}$$

where n_c is the total number of species in the mixture and n_a is the total number of atom types; $\varepsilon_{\tau\omega} = (\varepsilon_\tau\varepsilon_\omega)^{1/2}$ is the pair-interaction energy between atom types τ and ω, and m_τ^i is the effective number of atoms of type τ within a species i calculated from

$$m_\tau^i = \sum_{a\in i}\left(\frac{S_a}{S_{a0}}\right)^{q_s} \tag{15}$$

dx.doi.org/10.1021/ie404410v | Ind. Eng. Chem. Res. 2014, 53, 8265−8278

Table 1. Parameters Used in the Two COSMO-SAC (2013) Models (This Work), the COSMO-SAC (2007) Model,[10] and the COSMO-SAC (2010) Model[11]

Universal Parameters					
parameter	unit	2007 model	2010 model	2013-Dmol^3	2013-ADF
a_{eff}	Å^2	7.25	7.25	7.90	6.48
f_{decay}	–	3.57	3.57	–	–
r_{avg}	Å	–	–	0.85	0.51
c_{es}	kcal/mol·Å^4/e^2	–	6326 + $[(1.5 \times 10^8)/T^2]$	9254	7877
c_{hb}	kcal/mol·Å^4/e^2	3484	–	–	–
c_{OH-OH}	kcal/mol·Å^4/e^2	–	4014	3947	5787
c_{OH-OT}	kcal/mol·Å^4/e^2	–	3016	3311	4708
c_{OT-OT}	kcal/mol·Å^4/e^2	–	932	2086	2740
q_s	–	0.5	–	0.62	0.57
σ_0	e/Å^2	0.007	0.007	0.006	0.012
q_0	Å^2	79.53	79.53	79.53	79.53
r_0	Å^3	66.69	66.69	–	–
Atom Bonding Specific Parameters					
		ε_i/R^a (K·Å^3)			
atom type	R_i^{el} (Å)	2007 model	2010 model	2013-Dmol^3	2013-ADF
H	1.30	0	–	0	338
C(sp^3)	2.00	36 054	–	42 976	29 161
C(sp^2)	2.00	30 090	–	36 380	30 952
C(sp)	2.00	22 134	–	19 708	20 686
N(sp^3)	1.83	11 860	–	6 753	23 489
N(sp^2)	1.83	12 694	–	16 031	22 663
N(sp)	1.83	2590	–	8252	6390
O(sp^3-H)	1.72	6759	–	7799	8527
O(sp^3)	1.72	8421	–	4979	8484
O(sp^2)	1.72	3290	–	4497	6737
O(sp^2-N)	1.72	10 093	–	13 501	12 145
F	1.72	6742	–	7766	8435
P	2.12	83 439	–	67 738	82 512
S	2.16	60 949	–	59 665	56 068
Cl	2.05	35 624	–	46 518	45 065
Br	2.16	61 074	–	52 577	62 948
I	2.32	100 953	–	115 568	105 911

[a]In the 2007 model, all ε_i values were set to zero for modeling mixtures as described in the text.[10]

where S_a is the exposed surface area (solvent accessible surface area) of atom a of species i, and S_{a0} is the surface area of the bare atom calculated using the same set of atomic radii as in the COSMO calculation for the electrostatic contributions; q_s is a scaling factor in the range 0−1. V_{mix} is the molar volume of the mixture, which for simplicity can be calculated from the pure fluid molar volumes assuming ideal mixing. The detailed derivations are shown in the Supporting Information. Note that for a pure component ($n_c = 1$) this expression is still valid; thus, eq 14 can be used to calculate the dispersion contribution in both pure fluids and mixtures.

Also, an equivalent but simplified combinatorial term (eq 16)[18] is used here to calculate $\ln(\gamma_{i/S}^{comb})$ rather than eq 2:

$$\ln(\gamma_{i/S}^{comb}) = 1 - \frac{\phi_i}{x_i} + \ln\frac{\phi_i}{x_i} - \frac{z}{2}q_i\left(1 - \frac{\phi_i}{\theta_i} + \ln\frac{\phi_i}{\theta_i}\right) \tag{16}$$

In previous models,[8,10,11] the molecular volume and surface area were obtained from the COSMO calculation normalized with a volume of $r_0 = 66.69$ Å^3 and $q_0 = 79.53$ Å^2. In the current model, only the surface area q_0 is needed for normalization as r_0 cancels internally in eq 16.

3. COMPUTATIONAL DETAILS

The activity coefficient is calculated using eq 13 with the dispersion term here and was calculated with eq 1 without dispersion in our previous models. The activity coefficient is used in computing all phase equilibria, including the partition coefficient of a solute i between phases I and II, K_i, which is calculated from[19]

$$K_i = \frac{C_{tot}^{I}}{C_{tot}^{II}}\frac{\gamma_i^{II,\infty}}{\gamma_i^{I,\infty}} \tag{17}$$

Here C_{tot}^{I} and C_{tot}^{II} are the total molar concentrations of phases I and II, respectively. $\gamma_i^{I,\infty}$ and $\gamma_i^{II,\infty}$ are the infinite-dilution activity coefficients of solute i in phases I and II, respectively. In calculating octanol−water partition coefficients, the octanol-rich phase is considered to be phase I and the water-rich phase is considered to be phase II. Water is partly miscible in octanol, so at equilibrium approximately 27.5 mol % water and 72.5 mol % octanol are present in phase I and pure water is assumed to be phase II under ambient conditions. Therefore, $C_{tot}^{I} = 8.37$ mol/L and $C_{tot}^{II} = 55.5$ mol/L are normally used to report octanol−water partition coefficients, which is what we do here. Except for the octanol−water system, the solvent and water are

dx.doi.org/10.1021/ie404410v | Ind. Eng. Chem. Res. 2014, 53, 8265−8278

assumed to be completely immiscible for other solvent−water partitioning systems in this work.

The vapor pressure P_i^{vap} is determined using[9]

$$\ln P_i^{vap} = \frac{\Delta \underline{G}_{i/i}^{*sol}}{RT} + \ln \frac{RT}{\underline{V}_{i/L}} - \ln \phi \tag{18}$$

where $\Delta \underline{G}_{i/i}^{*sol}$ is the self-solvation Gibbs energy, and $\phi = f(T, P_i^{vap})/P_i^{vap}$ is the fugacity coefficient accounting for the nonideality of the vapor phase. $\Delta \underline{G}_{i/i}^{*sol}$ is calculated using the same procedure as in the previous models[9,10] for predicting pure compound thermodynamic properties.

The enthalpy of vaporization is obtained from the derivative of the vapor pressure with respect to temperature:

$$\frac{\Delta H_i^{vap}}{RT} = \frac{\underline{V}_{i/V} - \underline{V}_{i/L}}{RT} \frac{dP_i^{vap}}{dT} \tag{19}$$

where $\underline{V}_{i/V}$ is the vapor phase molar volume of species i. If the vapor phase is assumed to be ideal, then

$$\frac{\Delta H_i^{vap}}{RT} = \left(T - \frac{P_i^{vap}\underline{V}_{i/L}}{R}\right) \frac{d(\ln P_i^{vap})}{dT} \tag{20}$$

The normal boiling temperature (NBT) is calculated by solving eq 18 at a pressure of 1 atm, using Newton's iterative method.

The procedure for the COSMO calculations at the QM level is same as that used earlier[10,20] and is not reproduced here. In the previous models,[10,11] the calculations were done using only the COSMO database created using the DMol^3 QM package.[21,22] Now we have also created a COSMO database using the Amsterdam Density Functional (ADF) QM package.[23−25] The detailed settings for COSMO calculation at the QM level to create the databases are provided in the Supporting Information.

The liquid molar volume for each compound is required for the dispersive interaction and also in the calculation of the pure compound thermodynamic properties. The liquid molar volumes were directly taken from the Design Institute for Physical Property Data (DIPPR)[26] database, the AspenTech[27] database, or the NIST[28] database if available; otherwise, a relationship established by fitting the molecular volume from COSMO calculation to the experimental data was used to estimate $\underline{V}_{i/L}$. The fitting details are shown in the Supporting Information. It is noted that the fitting shows an average 5% variation in the value of $\underline{V}_{i/L}$, which could affect the prediction accuracy. The advantage of using this correlation is that $\underline{V}_{i/L}$ can be predicted within the model if experimental data are not available.

There are a total of seven universal parameters in COSMO-SAC (2007),[10] 10 universal parameters in COSMO-SAC (2010),[11] and nine universal parameters in our modified model here (denoted now as COSMO-SAC (2013)), in addition to the dispersive parameters. The parameter values for use separately with DMol^3 and ADF in the COSMO-SAC (2013) model and those in the COSMO-SAC (2007)[10] and (2010)[11] models can be found in Table 1.

4. PARAMETER OPTIMIZATION

The parametrization procedure in this work is different from that in our earlier work.[8−11] Previously, some parameters in the models were determined separately from correlating mixture data and then kept fixed while the remaining parameters were determined using pure compound data. In this work, all the parameters were determined simultaneously for both pure compounds and mixtures. The objective function was the root-mean-square deviation (RMSD) in the sum of the vapor pressures, heats of vaporization, activity coefficients, and octanol−water partition coefficients as follows:

$$\text{Obj} = \text{Objl}_{pure} + W_m \cdot \text{Objl}_{mix} \tag{21}$$

where

$$\text{Objl}_{pure} = \sqrt{\frac{1}{N_1}\sum_{i=1}^{N_1} (\ln P_i^{expt} - \ln P_i^{calc})^2} + W_h \sqrt{\frac{1}{N_1}\sum_{i=1}^{N_1} \left(\frac{\Delta H_i^{expt} - \Delta H_i^{calc}}{\Delta H_i^{expt}}\right)^2} \tag{22}$$

$$\text{Objl}_{mix} = \sqrt{\frac{1}{N_2}\sum_{i=1}^{N_2} (\ln \gamma_i^{expt} - \ln \gamma_i^{calc})^2} + \sqrt{\frac{1}{N_3}\sum_{i=1}^{N_3} (\log K_{ow_i}^{expt} - \log K_{ow_i}^{calc})^2} \tag{23}$$

In eq 22, the weighting factor of $W_h = 2$ for the heats of vaporization was used to achieve an optimum simultaneous correlation for vapor pressures and heats of vaporization.[9,10] In eq 21, the weighting factor of $W_m = 1.5$ was set to balance the accuracy of the pure and mixture properties. N_1, N_2, and N_3 are numbers of data points used in the optimization for the pure compounds (vapor pressures and heats of vaporization), activity coefficients, and octanol−water partition coefficients, respectively.

An optimization procedure was used in which a genetic algorithm[29] was first used to identify the vicinity of the global minimum, after which the simplex algorithm[30] was used to locate the minimum more precisely. The COSMO-SAC (2007)[10] parameters were used as initial guesses. The optimization was also run multiple times with different initial guesses to ensure repeatability and validity. Since the number of experimental data points and the number of parameters being optimized are relatively large, an in-house COSMO-SAC program for parallel computing was developed and run in a high-speed Linux cluster to ensure that the optimization could be completed in a short period of time.

5. RESULTS AND DISCUSSION

The revised model (COSMO-SAC (2013)) in this work was validated by comparing it with the refined COSMO-SAC (2007)[10] and (2010)[11] models for the prediction of various thermodynamic properties for both mixtures (partition coefficients, activity coefficients, and vapor−liquid equilibrium (VLE)) and pure compounds (vapor pressures, heats of vaporization, and normal boiling point temperatures). For simplicity, hereafter the COSMO-SAC (2013) model for use with DMol^3 and ADF is referred to as 2013-DMol^3 and 2013-ADF, respectively. Also, COSMO-SAC (2007) and COSMO-SAC (2010) are referred to as the 2007 and 2010 models, respectively.

5.1. Prediction of Partition Coefficients. In this study, partitioning of various solutes between a number of different solvents at room temperature (298.15 K) were considered, including 992 octanol−water systems, 277 tetrachloromethane−water systems, 276 chlorobutane−water systems, 88

dx.doi.org/10.1021/ie404410v | Ind. Eng. Chem. Res. 2014, 53, 8265−8278

Table 2. Accuracy of the COSMO-SAC Models for the Prediction of the Selected 992 Octanol−Water Partitioning Systems[a]

	group	N_{sys}	model	a	r	RMSD
class 1	monofunctional	343	2013-DMol^3	0.98	0.95	0.59
			2013-ADF	1.03	0.95	0.61
			2010	0.98	0.93	0.67
			2007	0.98	0.94	0.67
	multifunctional	649	2013-DMol^3	0.96	0.96	0.65
			2013-ADF	0.94	0.96	0.63
			2010	0.93	0.94	0.74
			2007	0.92	0.95	0.70
class 2	non-HB	376	2013-DMol^3	0.95	0.97	0.50
			2013-ADF	0.95	0.96	0.53
			2010	0.91	0.96	0.65
			2007	0.92	0.96	0.63
	HB	616	2013-DMol^3	1.00	0.90	0.70
			2013-ADF	1.02	0.91	0.68
			2010	1.05	0.88	0.76
			2007	1.00	0.90	0.72
overall		992	2013-DMol^3	0.96	0.95	0.63
			2013-ADF	0.97	0.96	0.62
			2010	0.95	0.93	0.72
			2007	0.94	0.95	0.69

[a] a and r are, respectively, the slope of the regression line and the correlation coefficient determined from the goodness of fit ($\log K_{calc} = a \log K_{expt}$) between predicted and experimental log K values assuming a zero intercept for correct scaling; N_{sys} is the number of partitioning systems.

hexane−water systems, 53 dichloromethane−water systems, 51 chlorobutane−water systems, 42 benzene−water systems, and 42 diethyl ether−water systems. There are a total of 1024 unique solute species in these partitioning systems (from the smallest, molecular hydrogen (two atoms), to the largest, dexamethasone (62 atoms) and tetracosane (74 atoms)). Two types of solute species were considered here: monofunctional and multifunctional. Monofunctional compounds contain only one nonalkyl, strong functional group, e.g., species in the homologous series, cyclic alkanes, alkylbenzenes, and branched, secondary, and tertiary species. Species containing more than one strong functional group, the multifunctional compounds, include $X(CH_2)_nY$ type species, where X and Y are strong functional groups (OH, COOH, CHO, $COCH_3$, CH_3COO, CH_3O, NO_2, CN, $CONH_2$, C_6H_5, $CHCH_2$, CCH, F, Cl, Br, I, and S), chlorofluorocarbons, multiaromatic rings, multichlorinated benzenes, phenol, aniline, toluene, pyridine, furan, and their derivatives, and complex pharmaceutical compounds such as caffeine, theophylline, and aspirin. The complete list of the systems and the predictions can be found in the Supporting Information. The values of the partition coefficients (K_i) range over more than 12 orders of magnitude from less than 10^{-4} to greater than 10^9. Therefore, the logarithm of K_i, log K, is normally used, and the predicted errors are measured using the RMSD of log K as follows:

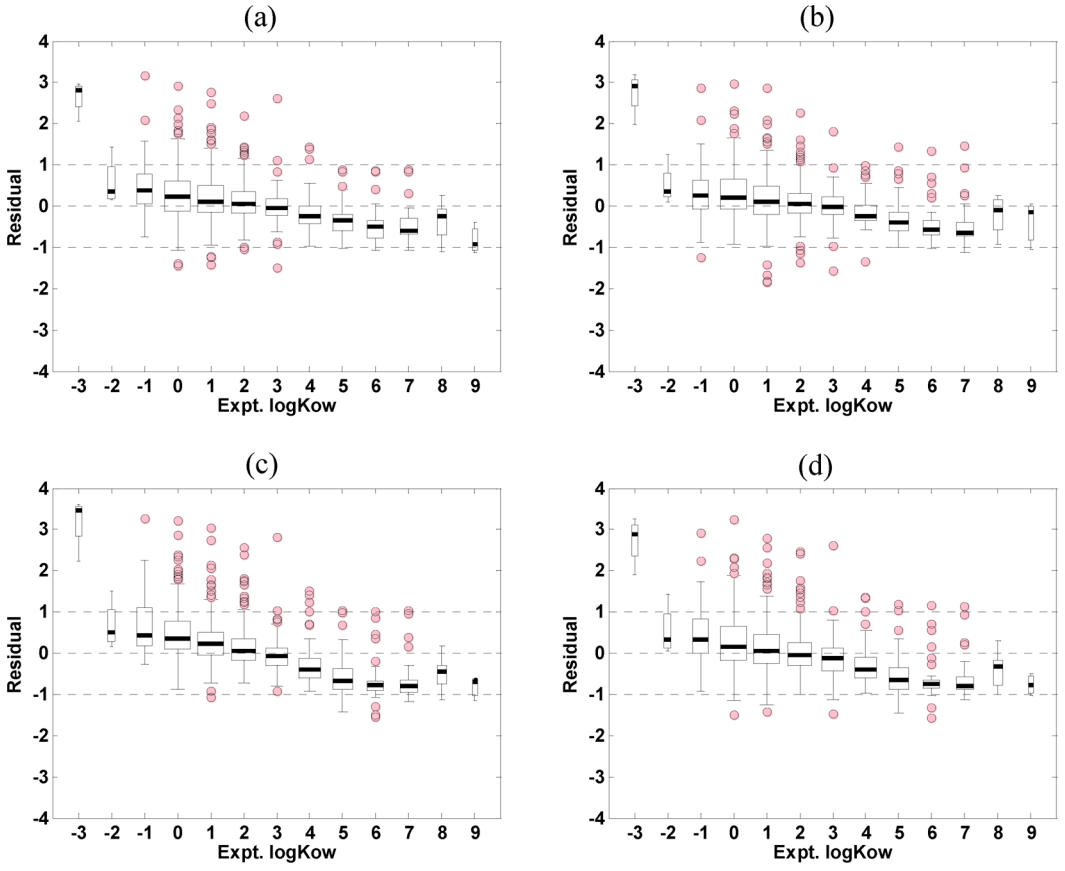
$$\mathrm{RMSD} = \left[\frac{1}{N}\sum_{i=1}^{N}(\log K_i^{\mathrm{calc}} - \log K_i^{\mathrm{expt}})^2\right]^{1/2} \quad (24)$$

where N is the number of partitioning systems (one solute per system). The experimental partition coefficient values were taken from the literature[31−36] that were compilations of measured data.

Table 2 summarizes the accuracy of the predicted octanol−water partition coefficients. To evaluate the predictions of the new model, in addition to the RMSD error, the following information was also considered: (1) how close the slope of the regression line between the predicted and measured values is to unity, (2) how close the intercept with the ordinate axis is to 0, and (3) how close the correlation coefficient is to 1. Therefore, assuming a zero intercept for correct scaling, the slope of the regression line (a) and the correlation coefficient (r) determined from the goodness of the fit between the predicted and experimental log K values ($\log K_{calc} = a \log K_{expt}$) are shown in Table 2 along with the RMSD values. The overall RMSD of log K for the 992 octanol−water partitioning systems (log K_{ow}) from 2013-DMol^3 and 2013-ADF are similar (0.63 and 0.62, respectively), which is about 10% reduction from 2007 and 2010 models (0.69 and 0.72, respectively). Similar improvements are observed for both monofunctional and multifunctional solutes. Furthermore, the 2013 models are most successful in terms predicting log K_{ow} for non-HB solutes since their RMSD from 2013 models is about 20% less than that of the 2007 and 2010 models, though the RMSD of log K_{ow} for HB solutes from the 2013 models is only about 5% less than that from 2007 and 2010 models. In addition, the slope of the regression line of 2013 models is closer to unity and the correlation coefficient is generally higher compared to that from 2007 and 2010 models. Overall, the 2013 models provide a better accuracy of log K_{ow} than the previous two models. While the 2010 model was optimized to liquid−liquid and vapor−liquid equilibrium data, it is the worst of these models for predicting log K_{ow}. The reason for that will be discussed later.

Figure 1 is a statistical box plot that shows the distribution of residuals as a function of the experimental log K_{ow} values. The residual is the difference between the predicted log K_{ow} from COSMO-SAC models and the experimental values. The bold solid line inside each box indicates the median value of the residual for the entire data set under each bin. The lower and upper lines for each box represent the average values of the lower half and upper half of the entire data set separated by the median value, denoted as Q_1 and Q_2, respectively. The "interquartile range", abbreviated IQR, is the width of the box (IQR = $Q_2 - Q_1$). The IQR can be used as a measure of how

dx.doi.org/10.1021/ie404410v | Ind. Eng. Chem. Res. 2014, 53, 8265−8278



Box/Bin	1	2	3	4	5	6	7	8	9	10	11	12	13
Expt. $\log K_{ow}$	[-4, -3]	[-3, -2]	[-2, -1]	[-1, 1]	[1, 2]	[2, 3]	[3, 4]	[4, 5]	[5, 6]	[6, 7]	[7, 8]	[8, 9]	[9, 10]
Number of systems	4	4	28	236	213	173	104	87	59	32	41	8	3

Figure 1. Box plots showing the residual ($\log K_{calc} - \log K_{expt}$) distribution as a function of $\log K_{expt}$ for the selected 992 octanol–water partitioning systems calculated using the COSMO-SAC with (a) 2013-DMol[3], (b) 2013-ADF, (c) 2010, and (d) 2007 models. The legend at the bottom shows the number of partitioning systems used for each box/bin.

spread out are the residual values. The lower and upper bounds outside the box are defined by $Q_1 - 1.5(IQR)$ and $Q_2 + 1.5(IQR)$, respectively. The circular points out of the lower and upper bounds are the statistical outliers.

Figure 1 clearly shows that 2013 models generally overpredict the log K_{ow} values when the experimental values are less than 2, and underpredict the values when the experimental values are greater than 4. The possible reason could be that, as the absolute value of log K_{ow} becomes larger, the measurement error becomes higher since the solute concentration becomes extremely small either in the octanol phase or in the water phase. Similar behavior is observed in the predictions with the 2007 and 2010 models, but the trends for those are worse. As shown in Table 3, the RMSD of log K_{ow} for the 2013 models decreases by over 20% compared with the 2007 and 2010 models when examining the data points whose experimental log K_{ow} values are larger than 4. For the data points whose experimental log K_{ow} values are less than 4 (excluding the outstanding residuals from the four partitioning systems with experimental log K_{ow} less than −3), the RMSD from 2013 models is decreased by 5% from the 2007 model and 8% from the 2010 model.

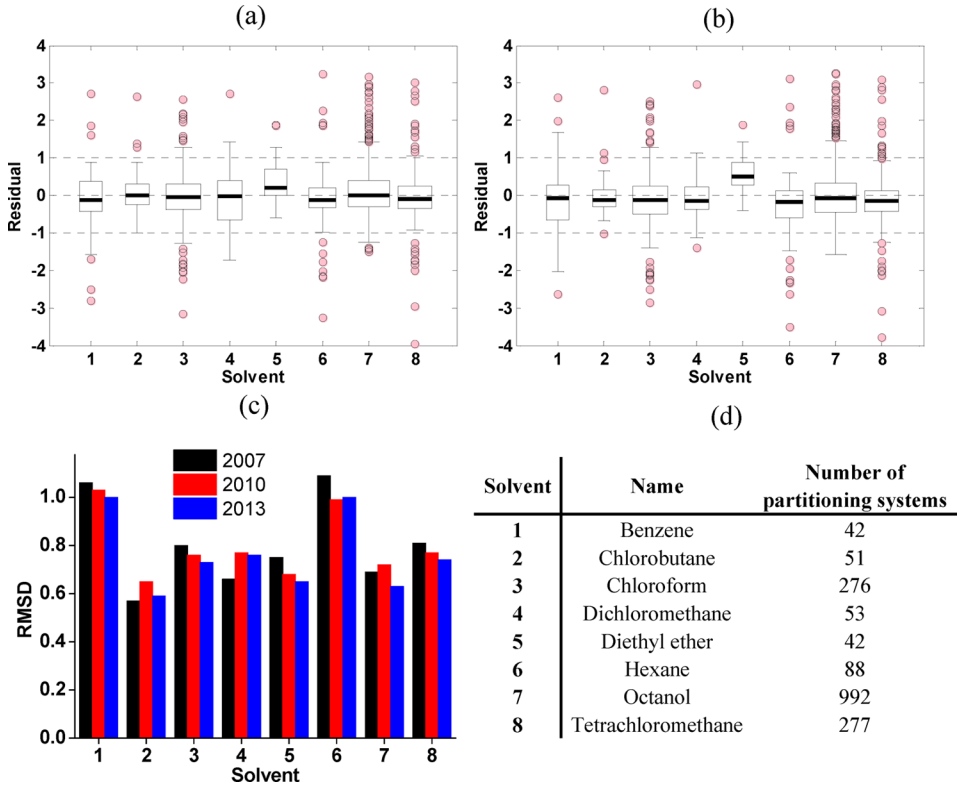
Table 3. Comparison of RMSD Values for Selected Octanol–Water Partitioning Systems Classified by Experimental log *K* Values

		RMSD			
	N	2013-DMol[3]	2013-ADF	2010	2007
$\log K_{expt} \leq 4$	761	0.63	0.62	0.68	0.66
$\log K_{expt} > 4$	227	0.53	0.54	0.72	0.70

While the parameters for the 2013 models introduced in this work are obtained from fitting to log K_{ow} data, the accuracy of the predictions for other solvent–water partition coefficient predictions is also improved. Figure 2 shows the residual distribution and the RMSD for eight different solvents consisting of 1821 solvent–water partitioning systems. The residuals were calculated in the same manner as mentioned above, and are more evenly distributed around 0 for the different solvent systems from the 2013 model (Figure 2a) than from the previous model (Figure 2b). The RMSD values from 2013 models for the different solvent systems are generally lower than those from the previous two models with one exception (the 2007 model has lower RMSD for the 53 dichloromethane–water partitioning systems).

5.2. Prediction of Activity Coefficients. The increased accuracy of the partition coefficient predictions is a result of the improvement in the accuracy of the infinite dilution activity

dx.doi.org/10.1021/ie404410v | Ind. Eng. Chem. Res. 2014, 53, 8265–8278



Solvent	Name	Number of partitioning systems
1	Benzene	42
2	Chlorobutane	51
3	Chloroform	276
4	Dichloromethane	53
5	Diethyl ether	42
6	Hexane	88
7	Octanol	992
8	Tetrachloromethane	277

Figure 2. COSMO-SAC predictions for solvent–water partitioning systems: (a) box plot showing the residual ($\log K_{calc} - \log K_{expt}$) distribution for the COSMO-SAC (2013) model, (b) box plot showing residual distribution for the COSMO-SAC (2007) model, (c) RMSD comparisons for the 2007, 2010, and 2013 models, and (d) legend showing the name of each solvent and number of systems used for each solvent box.

coefficient. Table 4 summarizes the RMSD of the activity coefficient predictions from the 2007, 2010, and 2013 COSMO-SAC models for the data set of 4388 binary mixture systems containing 339 unique compounds. Unlike the partitioning systems studied at room temperature, the temperature for the binary mixture systems ranges from 213.15 to 523.15 K. Due to the large variations in the values of activity coefficients, their RMSD is calculated in the natural logarithm unit. The experimental data for the activity coefficients are from the DECHEMA[37,38] database, and the complete list of systems and the calculated results can be found in the Supporting Information. The entire data set was also divided into nondilute and dilute subsets according to the concentration. The dilute subset contains data for those systems with a mole fraction less than 0.01. Since the data set contains many aqueous systems, it was also divided into nonaqueous and aqueous subsets so that the accuracy of the predictions can be evaluated from different perspectives.

Table 4. RMSD of Activity Coefficients for the Two COSMO-SAC (2013) Models (This Work), the COSMO-SAC (2007) Model,[10] and the COSMO-SAC (2010) Model[11] for the Entire Data Set (4388 Binary Systems), as Well as Two Different Subsets: One for Aqueous and Nonaqueous Data and the Other One for Dilute (Mole Fraction <0.01) and Nondilute Data

	all data (6092)[a]	nonaqueous (4500)	aqueous (1592)	nondilute (3415)	dilute (2677)
2007	0.66	0.69	0.55	0.32	0.93
2010	0.67	0.71	0.54	0.29	0.96
2013-DMol^3	0.61	0.65	0.50	0.28	0.87
2013-ADF	0.58	0.61	0.47	0.26	0.82

[a]Numbers in parentheses are numbers of data points.

As shown in Table 4, the overall RMSDs from the 2007, 2010, 2013-DMol^3, and 2013-ADF COSMO-SAC models for the entire data set are 0.66, 0.67, 0.61, and 0.58, respectively. The RMSD of an 2013 model has an overall decrease of about 10% relative to that of the 2007 or 2010 models. The RMSDs from 2007, 2010, 2013-DMol^3, and 2013-ADF models for dilute systems are 0.93, 0.96, 0.87, and 0.82 while those for for nondilute systems are 0.32, 0.29, 0.28, and 0.26, respectively. This shows that the decrease of RMSDs from the new models is in both dilute and nondilute concentrations. In contrast, the 2010 model has an increased accuracy for nondilute systems at the expense of that for dilute systems. Consequently, the partition coefficients calculated from the 2010 model are the least accurate of these models as shown previously.

An interesting question is whether the electrostatic interaction or the dispersion interaction is the main contributor to the improvement in the accuracy of activity coefficients in the 2013 models. The electrostatic interaction is mainly from the restoring free energy that is related to the σ-profile and segment interaction strength parameter. Figure 3 shows the σ-profiles of four selected compounds. The 2007 and 2010 models use the same σ-profiles since they both used the DMol^3 QM/COSMO database and have the same parameters for the HB probability distribution. The 2013-DMol^3 model also used the DMol^3 database but has a slightly different parameter for HB probability distribution, leading to slightly different σ-profiles. The 2013-ADF model used the ADF QM/COSMO database and a different parameter for HB probability distribution, so the σ-profiles are different, but the shape and

dx.doi.org/10.1021/ie404410v | Ind. Eng. Chem. Res. 2014, 53, 8265–8278

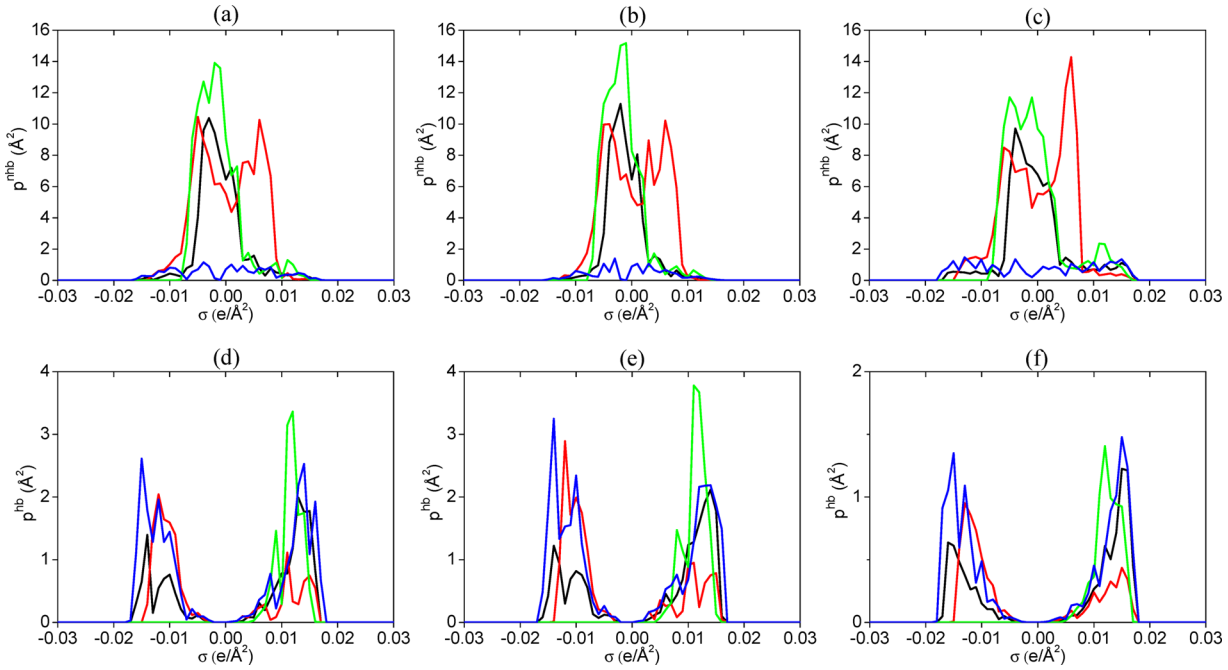


Figure 3. σ-profiles (multiplied by the surface area, A_i) for the (a–c) non-HB part and (d–f) HB part of four selected compounds: ethanol (black lines), aniline (red lines), 2-butanone (green lines), and water (blue lines). The left plots (a, d) are from the COSMO-SAC (2007) and the COSMO-SAC (2010) models, the middle plots (b, e) are from the COSMO-SAC (2013)-DMol^3 model, and the right plots (c, f) are from the COSMO-SAC (2013)-ADF model.

Table 5. Activity Coefficient Predictions for Seven Selected Solutes at Infinite Dilution, Each of Which Exhibits a Significant Dispersion Contribution ($\ln \gamma_{i/S}^{disp}$)[a]

			2013-DMol^3					2007 model				
solute/solvent	T (K)	$\ln \gamma_{i/S}^{expt}$	$\ln \gamma_{i/S}^{res}$	$\ln \gamma_{i/S}^{disp}$	$\ln \gamma_{i/S}^{comb}$	$\ln \gamma_{i/S}^{calc}$	APE (%)	$\ln \gamma_{i/S}^{res}$	$\ln \gamma_{i/S}^{disp}$	$\ln \gamma_{i/S}^{comb}$	$\ln \gamma_{i/S}^{calc}$	APE (%)
2,2,4-trimethylpentane/quinoline	293.15	2.56	1.40	1.15	0.02	2.57	0.3	1.39	0.0	0.02	1.41	44.9
tert-butyl chloride/benzyl acetate	298.15	0.52	0.19	0.30	−0.08	0.41	21.1	0.19	0.0	−0.08	0.11	78.8
2-butanone/1,2-dichloroethane	318.15	−0.31	−0.68	0.30	0.0	−0.38	22.6	−0.68	0.0	0.0	−0.68	119.4
octane/diethylaniline	322.15	0.96	0.65	0.20	0.04	0.89	7.3	0.66	0.0	0.04	0.70	27.1
heptane/1-methylnaphthalene	393.15	0.82	0.54	0.21	0.02	0.77	6.1	0.53	0.0	0.02	0.55	32.9
benzene/water	298.15	7.83	8.12	0.83	−1.05	7.90	0.9	8.33	0.0	−1.05	7.28	7.0

[a]The restoring contribution ($\ln \gamma_{i/S}^{res}$) and the combinatorial contribution ($\ln \gamma_{i/S}^{comb}$) are also given, along with the predicted activity coefficients ($\ln \gamma_{i/S}^{calc}$), the experimental data ($\ln \gamma_{i/S}^{expt}$), and the absolute percentage error (APE) ($|1 - \ln \gamma_{i/S}^{calc}/\ln \gamma_{i/S}^{expt}|$).

range of the σ-profiles for each compound are similar from the different models. Table S4 (Supporting Information) also shows that the surface area from 2013-DMol^3 for the four selected compounds has a distribution similar to that from the 2007 and 2010 models. Although the σ-profiles are same in the 2010 and 2007 models, the 2010 model increased the accuracy of the activity coefficient predictions at nondilute concentrations by an inclusion of temperature-dependent segment interaction strengths, but did not improve the accuracy at dilute concentrations. This leads to a conclusion that the increased accuracy at dilute concentrations for the 2013 models may primarily be due to the inclusion of the mixture dispersion term.

As mentioned earlier, in the 2007 model publication,[10] a brief analysis was performed (five data points, or 10 activity coefficients) of the effect of dispersion in mixture calculations, leading to the conclusion that the effect in mixtures was negligible. This erroneous conclusion was probably due to that only intermediate concentrations were considered (mole fractions between 0.32 and 0.68), and also to the small sample size. The activity coefficient predictions for six selected solutes at infinite dilution are shown in Table 5, which compares the 2013 and 2007 models in terms of the three different contributions to the activity coefficient: the restoring ($\ln \gamma_{i/S}^{res}$), the dispersion ($\ln \gamma_{i/S}^{disp}$), and the combinatorial ($\ln \gamma_{i/S}^{comb}$) contributions (see eq 13). It can be seen from Table 5 that the error has been reduced significantly as a result of the dispersion contribution, which indicates that the dispersion contribution in mixtures, especially at dilute concentrations, cannot be ignored. Also note that the accuracy of the 2007 model is not improved by simply turning on the dispersion interaction using the dispersive parameters optimized for pure properties, as is shown in the Supporting Information.

The greater improvement at dilute concentrations as a result of the dispersive interaction is also understandable. At intermediate concentrations, the environment around a molecule resembles to some degree that of the reference state (i.e., the molecule in its own pure liquid). Thus, the intermolecular interactions at these concentrations may not be very different from those in the pure liquid, and there is some cancellation of effects between the two states. As the dilute limit is approached, however, the actual and reference environments

dx.doi.org/10.1021/ie404410v | Ind. Eng. Chem. Res. 2014, 53, 8265−8278

Table 6. Comparison of the Overall Deviation in Binary VLE Data[a]

		2013-DMol^3		2013-ADF		2010		2007	
	N_{sys}	AAPD-P (%)	AAD-y_1 (%)	AAPD-P (%)	AAD-y_1 (%)	AAPD-P (%)	AAD-y_1 (%)	AAPD-P (%)	AAD-y_1 (%)
type I	681	4.73	1.86	3.85	1.63	4.66	1.75	4.77	1.80
type II	344	8.77	2.88	9.03	2.96	8.37	2.95	9.47	3.15
type III	442	9.34	4.06	7.78	3.72	8.87	3.86	10.98	4.64
overall	1467	7.07	2.76	6.25	2.57	6.80	2.67	7.74	2.97

[a] N_{sys} is the number of binary VLE systems studied in this work. AAPD-P is the AAPD of predicted total pressure (P). AAD-y_1 is the AAD of predicted vapor composition of component 1 (y_1).

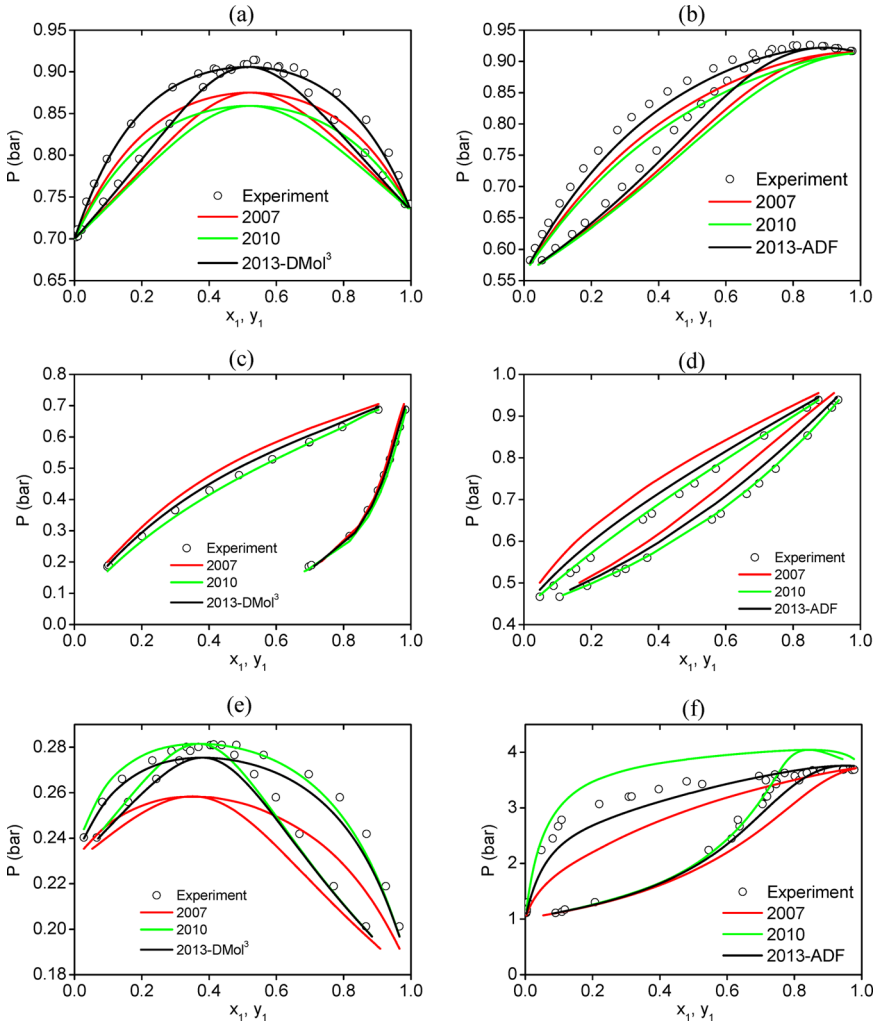


Figure 4. Comparison of vapor−liquid equilibrium predictions with experimental data for the following systems: (a) benzene + acetonitrile system (type I) at 343.15 K; (b) hexane + ethyl acetate system (type I) at 333.15 K; (c) toluene + aniline system (type II) at 373.17 K; (d) diethylamine + benzene system (type II) at 328.15 K; (e) ethanol + acetonitrile system (type III) at 313.15 K; (f) acetone + water system (type III) at 373.15 K. The open symbols are the experimental data, and the solid black, green, and red lines are the COSMO-SAC predictions with the 2013, 2010, and 2007 models, respectively.

are most dissimilar, and the cancellation of effects does not occur. As a result, changes to the modeling of all types of interactions, dispersion included, are more pronounced at dilute concentrations. This can be understood numerically by examining eq 13. At nearly pure concentrations, $\Delta G_{i/S}^{*\mathrm{disp}}$ is almost identical to $\Delta G_{i/i}^{*\mathrm{disp}}$, and the two effectively cancel out. As the solvent concentration is increased, the difference between $\Delta G_{i/S}^{*\mathrm{disp}}$ and $\Delta G_{i/i}^{*\mathrm{disp}}$ increases. At the dilute limit, the difference between the two is greatest, and how they are modeled has a significant effect on the predicted activity coefficient values. The same argument applies, of course, to the other types of interactions in the model.

In aqueous mixtures, the magnitude of the electrostatic interactions is relatively large, due primarily to the HB effects. Therefore, the relative contribution of the dispersion interactions in these mixtures is small. Thus, including dispersion should have less effect on aqueous mixtures as seen in the last example in Table 5. Note, however, that Table 4 shows that the RMSD of aqueous systems decreases similarly to, or even slightly greater than, that of nonaqueous systems. This observation can be attributed to the improvement in the

dx.doi.org/10.1021/ie404410v | Ind. Eng. Chem. Res. 2014, 53, 8265−8278

Table 7. Comparison of the Accuracy of Predictions for Pure Compound Thermodynamic Properties[a]

property	classification	N	RMSD or AAD (AAPD) 2007	2013-DMol^3	2013-ADF	2013-DMol^3-pure	2013-ADF-pure
vapor pressure (P^{vap})	non-HB	583	0.47	0.52	0.47	0.43	0.42
	HB	877	0.63	0.67	0.62	0.53	0.54
	overall	1460	0.57	0.62	0.57	0.49	0.50
heat of vaporization (ΔH)	non-HB	565	1.17	1.16	1.15	1.17	1.15
	HB	865	1.78	1.82	1.63	1.78	1.74
	overall	1430	1.57	1.59	1.46	1.57	1.53
NBTs	non-HB	583	15.8 (3.7%)	17.2 (4.1%)	15.6 (3.7%)	14.3 (3.4%)	13.6 (3.2%)
	HB	877	20.5 (4.5%)	22.0 (4.9%)	18.7 (4.1%)	17.5 (3.8%)	17.0 (3.7%)
	overall	1460	18.6 (4.2%)	20.1 (4.6%)	17.4 (4.0%)	16.2 (3.6%)	15.6 (3.5%)

[a]The accuracy of the vapor pressure is evaluated using the RMSD of ln P^{vap}, the heat of vaporization is evaluated using the RMSD of ΔH (kcal/mol), and the NBTs (kelvin) are evaluated using AAD with AAPD in parentheses. The model parameters for 2013-DMol^3-pure and 2013-ADF-pure models here are biased for pure compounds in the optimization (see text for details). N is the number of compounds evaluated.

electrostatic interaction contribution since the data for most aqueous systems used here are at nondilute concentrations.

5.3. Prediction of Phase Equilibria. While the parameters in the 2013 models were obtained from fitting to octanol−water partition coefficients, activity coefficients, and pure compound vapor pressures as well as heats of vaporization, the accuracy of the 2013 models for phase equilibria such as VLE are also evaluated. The experimental data for the vapor-phase and liquid-phase compositions and the total pressures are from the DECHEMA[37] database. When calculating the total pressures and vapor-phase compositions, the experimental pure component vapor pressures from the DIPPR[26] database were used, not the predictions from the COSMO-SAC model, so that errors in the activity coefficients are not confounded with those from the vapor pressures. A large set of systems (1467 binary mixtures over 20 000 data points with temperatures ranging from 207.92 to 623.15 K and pressures from 0.11 kPa to 6.89 MPa) was used for the evaluation. The accuracy is evaluated using the average absolute percentage deviation (AAPD) in pressure (AAPD-P)

$$\text{AAPD-}P\ (\%) = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i^{calc} - P_i^{expt}|}{P_i^{expt}}\cdot 100 \tag{25}$$

and the average absolute deviation (AAD) in vapor-phase composition (AAD-y_1)

$$\text{AAD-}y_1\ (\%) = \frac{1}{N}\sum_{i=1}^{N}|y_{1,i}^{calc} - y_{1,i}^{expt}|\cdot 100 \tag{26}$$

where N is the number of data points within a binary mixture. The accuracy of the VLE predictions for each binary system can be found in the Supporting Information.

To facilitate the analysis, the mixtures were categorized into three types.[11] Type I systems are those that do not have any HB interactions, such as nitromethane + benzene and acetone + hexane. Type II systems include a compound that has HB self-association, for example, water + hexane and ethylamine + benzene. Type III systems include both self- and cross-HB associating systems such as in the water + acetone and nitromethane + ethylamine systems. Table 6 summarizes the accuracy for the binary VLE predictions of the different models for different types of systems.

As shown in Table 6, compared to those from the 2007 model, the overall AAPD-P and AAD-y_1 from the 2010 model are reduced about 10%.[11] The overall AAPD-P and AAD-y_1 from the 2013-DMol^3 model are also reduced by almost 10% compared to those from the 2007 model, though they are only slightly lower than, but comparable to, those of the 2010 model. This is because 2013-DMol^3 has an accuracy for activity coefficient predictions similar to that of the 2010 model at nondilute concentrations typical of most measured VLE liquid compositions. There is almost no accuracy improvement for type I systems from 2013-DMol^3 compared to those from the 2007 model, but there was an slight improvement over the 2010 model, which is probably due to the temperature-dependent electrostatic interaction used in the 2010 model.[11] The temperature-dependent interaction strength parameter was not used in 2013 models because it had been introduced empirically. Interestingly, without any temperature-dependent correction, 2013-ADF reduces AAPD-P for type I systems by up to 20% compared to those from 2007 and 2010 models. It also reduces AAPD-P for type III systems by 30 and 10% compared to those from 2007 and 2010 models, respectively. This is because 2013-ADF has more accurate activity coefficient predictions at both nondilute and dilute concentrations than those of the 2007 and 2010 models. The increased accuracy is a result of the combination of dispersive interaction and the change in the σ-profiles as shown in Figure 3. The overall AAPD-P and AAD-y_1 from 2013-ADF are reduced 20 and 10%, respectively, compared to those from the 2007 model, and 8 and 4%, respectively, compared to those from the 2010 model.

Figure 4 illustrates the performances of the different models for type I, II, and III systems. Figure 4a,b shows example predictions for type I systems for some cases where the 2010 model is worse than the 2007 model, and the 2013 models are better than both. For type II systems, the 2010 model is very slightly better than the 2013 models, and the 2007 model is the least accurate as illustrated in Figure 4c,d. The predictions for type III systems are most challenging because they require an accurate description of the subtle differences in the HB interactions among all the species pairs. Figure 4e,f demonstrates that the 2013 and 2010 models are more accurate than the 2007 model, and that the 2013 models are the most accurate at dilute concentrations.

5.4. Prediction of Thermodynamic Properties for Pure Compounds. Although the main purpose in this work was to increase the predictive accuracy for mixture thermodynamic property predictions, it is interesting to evaluate the accuracy of the pure compound thermodynamic properties using the same set of parameters. [Note that the 2007 model calculates the pure fluid thermodynamic properties by including dispersive

dx.doi.org/10.1021/ie404410v | Ind. Eng. Chem. Res. 2014, 53, 8265−8278

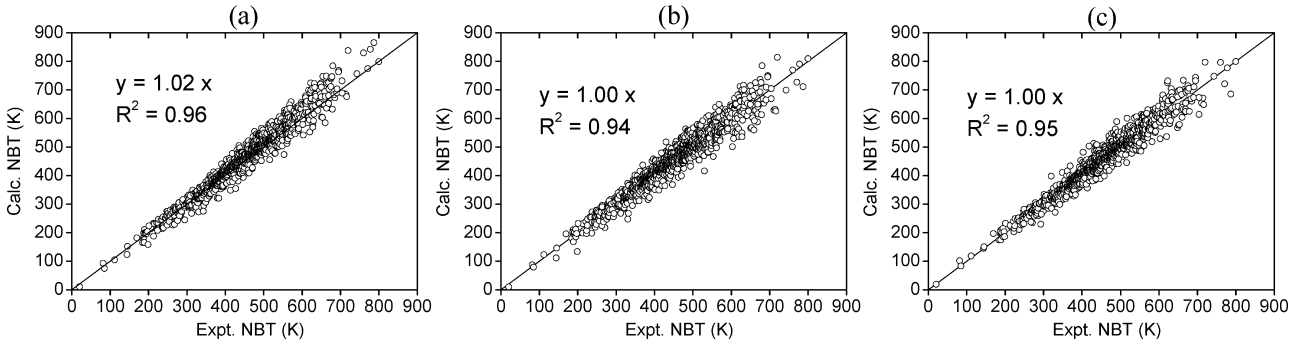


Figure 5. NBT predictions for 1460 compounds from (a) the 2007 model, (b) the 2013-DMol^3 model, and (c) the 2013-ADF model. The linear regressions between the predicted and experimental data assuming a zero intercept are also shown on the figures.

interactions, though not for mixtures, and therefore cannot be considered as using the same model and parameters as for mixtures. Also, the 2010 model does not accurately predict the pure fluid thermodynamic properties.] In our database 1460 unique compounds containing H, C, N, O, S, P, F, Cl, Br, or I atoms were used to evaluate the predictions of vapor pressures (P^{vap}), heats of vaporization (ΔH), and normal boiling point temperatures (NBTs). The experimental data for the ΔH and NBTs are from the DIPPR,[26] AspenTech,[27] or NIST[28] databases. The NBTs range from the lowest of 20 K to the highest of 800 K. [The experimental vapor pressure at the NBTs used is 101 325 Pa, or 11.526 in natural logarithm units]. Among the 1460 compounds, there are 1430 compounds for which the experimental data for heats of vaporization at the NBTs are available. The complete list of the compounds and the predicted data can be found in the Supporting Information.

Table 7 summarizes the accuracy of the predictions of P^{vap}, ΔH, and NBTs from the different COSMO-SAC models. The prediction accuracy of $\ln(P^{vap})$ and ΔH is evaluated using the RMSD, and that of the NBTs is evaluated using AAD with AAPD shown in parentheses, which is consistent with the evaluation method used elsewhere.[7,10,39−41] All 1460 compounds were categorized into subclasses based on their HB characteristics. The predictions show that the errors for HB compounds are generally larger. The overall accuracies for the predictions from 2007, 2013-DMol^3, and 2013-ADF are quite similar. The 2013-DMol^3 predictions are slightly less accurate than those of the 2007 model, and the 2013-ADF predictions are slightly more accurate than both. This is reasonable considering that there is only a slight change in the models for the pure compounds.

In addition, two sets of new model parameters biased for the predictions of pure compound thermodynamic properties (denoted as 2013-DMol^3-pure and 2013-ADF-pure) were also obtained. The parameters are listed in Table S3 in the Supporting Information. The accuracies of 2013-DMol^3-pure and 2013-ADF-pure are also listed in Table 7 for comparison, which shows that the accuracy is improved for P^{vap} and NBTs for a large variety of compounds.

Figure 5 shows the comparisons of the predicted and experimental NBT values. The results from the 2007 model generally overestimate the NBTs for compounds that have high NBTs. Although there is a slightly larger systematic error from 2013-DMol^3, the correlation between predicted and experimental NBT values is reasonably linear over the whole temperature range. The 2013-ADF model has the best accuracy and also has the best correlation between experiment and prediction.

5.5. Prediction Trends or Limitations of the Modified COSMO-SAC Model. Finally, it is useful to discuss several trends and point out some limitations for our modified COSMO-SAC model. When we compare the RMSD values for the octanol−water partition coefficients for the compounds classified by the elements and functional groups (Table S5 in the Supporting Information), we find that the 2013 models generally improve the prediction accuracy for the different classes of compounds. The RMSD values from the 2013 models for hydrocarbon and Cl-containing compounds that contain large numbers of samples decrease by 20−30% compared to those from the 2010 and 2007 models. However, the prediction of partition coefficients for nitrogen-containing solutes from all COSMO-SAC models is of significantly lower accuracy and the 2013 models do not show much improvement. The issues for nitrogen-containing compound predictions have also been shown in the literature,[42] and are probably due to the strong hydrogen bonding to the nitrogen functional groups. A smaller improvement is also shown for other strong hydrogen-bonding compounds such as alcohols, acids, ketones, and multifunctional species.

Similarly, when applied to the prediction of VLE behavior, the new models are generally better for those systems where the interaction is dominated by a dispersive effect such as the hexane + tetrachloromethane system (Figure S6 in the Supporting Information), while for the systems containing largely polar groups such as amines, nitriles, alcohols, and ketones, the overall improvement is limited. The difficulties in predictions for these systems using COSMO-based models have been documented in the literature.[6−8,20] Therefore, for such systems caution should be exercised when using the new models.

6. CONCLUSIONS

A more accurate COSMO-SAC model (denoted as COSMO-SAC (2013)) has been presented here to predict thermodynamic properties of pure fluids and mixtures. This model separates the charge averaging from the electrostatic interactions and includes the dispersion interaction for mixtures, which are different from the previous COSMO-SAC models. A new parametrization has been developed based on data for thousands of pure compounds and mixtures simultaneously with the σ-profiles generated separately from the DMol^3 and ADF QM packages. The accuracy of this new COSMO-SAC (2013) model has proved to be superior to that of the previous COSMO-SAC model in the predictions of thermodynamic properties for mixtures such as partition coefficients, activity coefficients, and phase equilibria while

dx.doi.org/10.1021/ie404410v | Ind. Eng. Chem. Res. 2014, 53, 8265−8278

maintaining a similar or even higher accuracy in the predictions of the pure component vapor pressures, heats of vaporization, and normal boiling point temperatures. This is based on an analysis of a total of almost 1800 unique compounds and 3000 mixtures with more than 30 000 experimental data points over a wide range of temperatures.

The COSMO-SAC (2013) model presented here increases the prediction accuracy for both pure fluid and mixture properties with only one small set of universal parameters in the model for all molecules. It can be used to predict thermodynamic properties for all pure compounds and mixtures that contain the C, H, N, O, S, P, F, Cl, Br, and/or I atoms within seconds or less on a personal computer for a species for which the σ-profiles are available in our database or calculated by the user in a QM platform including DMol^3 and ADF.

■ ASSOCIATED CONTENT

Supporting Information

Detailed derivation for the dispersion in mixtures, the quantum mechanics (QM) COSMO calculation settings using DMol^3 and ADF, the liquid density fitting procedure and additional sets of parameters for the prediction of pure compound thermodynamic properties. The list of compounds or systems and the predicted values for vapor pressures, heats of vaporization, normal boiling point temperatures, activity coefficients, and partition coefficients are also given. This material is available free of charge via the Internet at http://pubs.acs.org. Our graphic user interface (GUI) based program for the personal computer is available from the authors on request.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: sandler@udel.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the Catalysis Center for Energy Innovation (CCEI), an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Award No. DE-SC0001004, for their support. Parameter optimization calculations were performed with use of the computational resources provided by the Center for Functional Nanomaterials, Brookhave National Laboratory, supported by the U.S. Department of Energy, Office of Basic Energy Sciences, under Contract No. DE-AC02-98CH10886. We also acknowledge Prof. Dominic Di Toro for the idea of using box plots and his student, Yuzhen Liang, for gathering most of the experimental data for the partitioning coefficients.

■ REFERENCES

(1) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group-contribution Estimation of Activity-coefficients in Nonideal Liquid-mixtures. *AIChE J.* **1975**, *21*, 1086−1099.

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

(3) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995**, *24*, 279−287.

(4) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford Science Publications: Oxford, U.K., 1987.

(5) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999−3094.

(6) 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.

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

(8) 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.

(9) Lin, S. T.; Chang, J.; Wang, S.; Goddard, W. A.; Sandler, S. I. Prediction of vapor pressures and enthalpies of vaporization using a COSMO solvation model. *J. Phys. Chem. A* **2004**, *108*, 7429−7439.

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

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

(12) Sandler, S. I. *Chemical, Biochemical, and Engineering Thermodynamics*; John Wiley & Sons: New York, 2006.

(13) Staverman, A. J. The Entropy of High Polymer Solutions—Generalization of Formulate. *Recl. Trav. Chim. Pays-Bas—J. R. Neth. Chem. Soc.* **1950**, *69*, 163−174.

(14) Guggenheim, E. A. *Mixtures: The Theory of the Equilibrium Properties of Some Simple Classes of Mixtures, Solutions and Alloys*; Clarendon Press: Oxford, U.K., 1952.

(15) 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*, 1610−1624.

(16) Mu, T. C.; Rarey, J.; Gmehling, J. Performance of COSMO-RS with sigma profiles from different model chemistries. *Ind. Eng. Chem. Res.* **2007**, *46*, 6612−6629.

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

(18) Smith, J. M.; Van Ness, H. C.; Abbott, M. M. *Introduction to Chemical Engineering Thermodynamics*; McGraw-Hill: New York, 2005.

(19) Lin, S. T.; Sandler, S. I. Prediction of octanol-water partition coefficients using a group contribution solvation model. *Ind. Eng. Chem. Res.* **1999**, *38*, 4081−4091.

(20) Mullins, E.; Oldland, R.; Liu, Y. A.; Wang, S.; Sandler, S. I.; Chen, C. C.; Zwolak, M.; Seavey, K. C. Sigma-profile database for using COSMO-based thermodynamic methods. *Ind. Eng. Chem. Res.* **2006**, *45*, 4389−4415.

(21) *Cerius*^{2}; Accelrys, Inc.: San Diego, CA, 1999.

(22) *DMol*^{3}; Accelrys, Inc.: San Diego, CA, 1999.

(23) Te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Guerra, C. F.; Van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. Chemistry with ADF. *J. Comput. Chem.* **2001**, *22*, 931−967.

(24) Guerra, C. F.; Snijders, J. G.; Te Velde, G.; Baerends, E. J. Towards an order-N DFT method. *Theor. Chem. Acc.* **1998**, *99*, 391−403.

(25) *ADF2012*; SCM Theoretical Chemistry: Amsterdam, The Netherlands, 2012.

(26) *Design Institute for Physical Properties, Sponsored by AIChE DIPPR Project 801—Full Version*; Design Institute for Physical Property Data/AIChE.

(27) Aspen Technology, Inc.: Cambridge, MA.

(28) *NIST ThermoData Engine, NIST Standard Reference Database 103b*, version 7.0; Thermodynamics Research Center (TRC); National Institute of Standards and Technology (NIST): Boulder, CO, USA, 2012.

(29) Holland, J. H. Genetic Algorithms. *Sci. Am.* **1992**, *267*, 66−72.

(30) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes—The Art of Scientific Computing*; Cambridge University Press: Cambridge, U.K., 1986.
(31) Sangster, J. Octanol-water Partition-coefficients of Simple Organic-compounds. *J. Phys. Chem. Ref. Data* **1989**, *18*, 1111−1229.
(32) Lin, S. T.; Sandler, S. I. Multipole corrections to account for structure and proximity effects in group contribution methods: Octanol-water partition coefficients. *J. Phys. Chem. A* **2000**, *104*, 7099−7105.
(33) Hansch, C.; Leo, A.; Hoekman, D. H. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
(34) Ruelle, P. Universal model based on the mobile order and disorder theory for predicting lipophilicity and partition coefficients in all mutually immiscible two-phase liquid systems. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 681−700.
(35) Abraham, M. H.; Chadha, H. S.; Whiting, G. S.; Mitchell, R. C. Hydrogen-bonding. 32. An Analysis of Water-octanol and Water-alkane Partitioning and the Delta-log-P Parameter of Seiler. *J. Pharm. Sci.* **1994**, *83*, 1085−1100.
(36) Sprunger, L. M.; Achi, S. S.; Acree, W. E., Jr.; Abraham, M. H.; Leo, A. J.; Hoekman, D. Correlation and prediction of solute transfer to chloroalkanes from both water and the gas phase. *Fluid Phase Equilib.* **2009**, *281*, 144−162.
(37) *DECHEMA Chemistry Data Series Vol. I: Vapor-Liquid Equilibrium Data Collection*; DECHEMA: Frankfurt am Main, Germany, 1979−2010.
(38) *DECHEMA Chemistry Data Series Vol. IX: Activity Coefficients at Infinite Dilution*; DECHEMA: Frankfurt am Main, Germany, 1986.
(39) Stein, S. E.; Brown, R. L. Estimation of Normal Boiling Points from Group Contributions. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 581−587.
(40) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Normal boiling points for organic compounds: Correlation and prediction by a quantitative structure-property relationship. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 28−41.
(41) Wang, S.; Lin, S. T.; Chang, J.; Goddard, W. A.; Sandler, S. I. Application of the COSMO-SAC-BP solvation model to predictions of normal boiling temperatures for environmentally significant substances. *Ind. Eng. Chem. Res.* **2006**, *45*, 5426−5434.
(42) Mullins, E.; Liu, Y. A.; Ghaderi, A.; Fast, S. D. Sigma profile database for predicting solid solubility in pure and mixed solvent mixtures for organic pharmacological compounds with COSMO-based thermodynamic methods. *Ind. Eng. Chem. Res.* **2008**, *47*, 1707−1725.