Discussion and Improvement of the Refined COSMO-SAC Parameters for Solubility Predictions: Part 2

Baptiste Bouillot,* Sébastien Teychené, and Béatrice Biscans

Laboratoire de Génie Chimique, Université de Toulouse - CNRS, UMR CNRS 5503, BP 84234 Campus INP-ENSIACET, 4 allée Emile Monso, 31432 Toulouse Cedex 4, France

ABSTRACT: Solubility of drugs is a key piece of information for the pharmaceutical industry. Despite its importance, particularly at the beginning of a new drug process development, this thermodynamic property of the solid–liquid equilibria (SLE) can hardly be predicted for a given molecule in a given solvent. In our recent works, some thermodynamic models (UNIFAC and its modifications, COSMO-SAC and its refinements, NRTL-SAC) were investigated and compared for solubility prediction. The main drawbacks of these methods concern the strongest molecular interactions (dipole–dipole, hydrogen bonding), which are not properly taken into account. In the present study, we propose a new optimization of the last two COSMO-SAC refinements (2007 and 2010) for solubility prediction. To do that, a parameters optimization upon 352 solubility data of complex organic molecules was performed. Also, to improve hydrogen bonding influence, new σ-profiles are generated by applying another probability function for the solute. The results of this work are encouraging, especially for the calculation of crystallization yields given that the solubility temperature dependence is well represented. Solubility predictions in polar solvents are improved. However, this improvement is not as good as expected since it seems that the parameters balance between hydrogen bonding and electrostatic interactions is not the best. To confirm this, a short computation of anthracene solubility in toluene and heptane was performed.

INTRODUCTION

In the pharmaceutical industry, solubility of solid compounds is one of the most important thermodynamic properties. But, despite this importance, there is no predictive method, to date, that can predict solubility for all systems with accuracy.

In the past, some attempts were made to predict solubility of solid substances in liquids using existing thermodynamic models initially design for liquid–liquid equilibria (LLE) or vapor–liquid equilibria (VLE).[1−4] The strengths and the weaknesses of the classic models are shown in these studies. Among the most important weaknesses, the difficulty to represent molecular interactions in a mixture involving "simple" molecules (solvents) and "complex" molecules (solute) stands out from the others. Indeed, various interactions of different nature can occur (weak nonpolar van der Waals interactions, polar interactions, hydrogen bonding...). More specifically, the importance of hydrogen bonding is generally under- or overestimated and thus, classic thermodynamic models are unable to handle drug solubility.

In order to take more precisely into account the diversity of these interactions, especially hydrogen bonding, some models were developed specifically for solid–liquid equilibria (SLE) predictions, like NRTL-SAC[5] (a semiempirical activity coefficient model), PC-SAFT[6] (equation of state), or nonrandom hydrogen bonding (NRHB)[7] (equation of state). However, despite their great interest, these methods present some drawbacks. In general, they need a sufficient amount of experimental data. Without these data, they are not able to predict solubility accurately whatever the solute and the solvent.

Recent methods like COSMO-based models are also interesting in their approach. In particular, the COSMO-RS model[8] is a predictive model that allows the prediction of thermodynamic properties by using only data from quantum chemical calculations. COSMO-SAC,[9] another COSMO-based model, suggests a more empirical approach in addition to quantum chemistry. However, if these models are quite promising, they still show some difficulties taking into account hydrogen bonding properly. In fact, they handle the free energy character of hydrogen bonding in a quite "empirical manner".[10] In order to be more rigorous, the different ways of associating molecules in solution have to be considered and taken into account. Several authors suggested coupling the COSMO-RS approach with an equation of state[11,12] or more recently in partial solvation parameter approaches (PSPs).[13,14] PSPs combines quantum mechanics with quantitative structure–property relationships with parameter approaches (Hansen solubility parameters, for instance). These methods have not been widely studied for drug solubility so far.

In a previous study,[15] we qualitatively and quantitatively investigated the use of some thermodynamic models (UNIFAC, UNIFAC mod. Dortmund, COSMO-SAC, NRTL-SAC) for SLE. We compared them following several criteria: the mean square errors, the predicted orders of magnitude versus the experimental ones, and the solubility temperature dependence.

Despite its relative inaccuracy, the COSMO-SAC method was found to be a promising model since it does not rely on any specific database. Moreover, it takes into account molecular conformations and configurations.

In the first part of this study,[16] the last three refinements of the COSMO-SAC model[17−19] are investigated for SLE predictions. Even if solubility predictions in polar solvents were improved compared with the original COSMO-SAC model, the

solubility of large molecules in polar solvents is systematically overestimated.

In this paper, since COSMO-SAC uses an empirical manner of taking into account molecular interactions, we propose a reoptimization of refined COSMO-SAC model parameters (2007 and 2010) for solubility predictions. To do that, 352 SLE data points were considered (130 compounds in 36 solvents, see Table 6). Then, solubility predictions were performed in the same conditions as before[15,16] in single and mixed solvents. If the number of data is smaller than in the study of Hsieh et al.,[20] this study shows that COSMO-SAC can be improved for SLE prediction. However, the relationship between the electrostatic parameters and the hydrogen bonding ones seems to be unbalanced. That is why this paper ends with a discussion on the electrostatic parameters (not optimized in this study).

THEORY

Equilibrium Equation. Phase equilibria are described by the equality of chemical potentials, μ, in each phase. In the case of SLE, the thermodynamic equilibrium is called solubility. The solubility equation for a compound is written as follows:

$$\mu^{S} = \mu^{sat} \tag{1}$$

The superscript S denotes the solid phase, and the superscript sat denotes the saturated solution.

Following a classical development of eq 1, the well-known solubility equilibrium equation is obtained:[21]

$$\ln(\gamma^{sat}x^{sat}) = \frac{\Delta H_m^{fus}(T^{fus})}{R}\left(\frac{1}{T^{fus}} - \frac{1}{T}\right) - \frac{\Delta C_{pm}(T^{fus})}{R}\left[\ln\left(\frac{T^{fus}}{T}\right) - \frac{T^{fus}}{T} + 1\right] \tag{2}$$

where γ^{sat} is the activity coefficient, x^{sat} the solubility, ΔH_m^{fus} is the molar fusion enthalpy, T^{fus} the fusion temperature, and ΔC_{p_m} the difference between the molar heat capacity of the supercooled melt and the solid.

This equation gives the solubility as a function of the solid state properties of the compound, and the activity coefficient γ. The activity coefficient corresponds to the nonideal behavior of the mixture and is mostly solvent dependent.

Usually, the second part of eq 2 is neglected, and eq 3 is more commonly used:

$$\ln(\gamma_i^{sat}x_i^{sat}) = \frac{\Delta H_m^{fus}(T_m)}{R}\left(\frac{1}{T_m} - \frac{1}{T}\right) \tag{3}$$

In a previous work,[15] we did not observe significative differences between the results obtained using eq 2 or 3.

COSMO-SAC Models. COSMO-SAC methods are modifications of the COSMO-RS (conductor-like screening model) model, developed by Klamt.[8,22] They are based on quantum chemical calculations for the molecule representation and on statistical thermodynamics for molecular interactions.

The first COSMO-SAC method was developed by Lin and Sandler[9] and was formerly built to predict VLE or LLE. One of the most important contributions of this model is the improvement of complex interaction descriptions (dipole–dipole or hydrogen bonds). To do so, the model takes into account the donor/acceptor behavior of the different parts or surfaces (also called "segments") of the molecule. And so, COSMO-SAC equations use interaction parameters between the different "segments" in the mixture.

The complete theory and equations of the COSMO-SAC model are presented elsewhere.[9,17−19] However, in order to explain why the successive refinements of the original method might be interesting for solubility predictions of complex molecules, the major improvements are recalled here, as well as the role of the hydrogen bond interaction parameters (optimized in this work).

In 2004, Lin et al.[17] suggested a division of the molecule surfaces of the molecules (σ-profiles) into two parts: (1) surfaces not able to form hydrogen bonds (i.e., the previous neutral segments); (2) surfaces able to form hydrogen bonds (surfaces around O, N, or F atoms and H atoms connected to these atoms).

This division allows a finer description of the molecule interactions into the ΔW term, which can take the generic form

$$\Delta W(\sigma_m^t, \sigma_n^s) = C_{ES}(\sigma_m^t + \sigma_n^s)^2 - c_{hb}(\sigma_m^t, \sigma_n^s)(\sigma_m^t - \sigma_n^s)^2 \tag{4}$$

with

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

In 2007, Wang et al.[18] suggested a new averaging method for the σ-profiles calculations and added a probability function for hydrogen bond formation (screening probability). The aim of this change is to consider that the probability of occurrence of a complex interaction is not 100%. This probability function is expressed as follows:

$$P^{hb}(\sigma) = 1 - \exp\left(-\frac{\sigma^2}{2\sigma_0^{\,2}}\right) \tag{6}$$

with σ_0 equals 0.007 e/Å^2. This parameter corresponds to the standard deviation of the probability function of a Gaussian type function for the hydrogen bonds. The higher the value of σ_0 is, the less hydrogen bonds can occur for the lowest values of $|\sigma|$ ("less polar" parts on the molecule surface).

At last, in 2010, Hsieh et al.[19] suggested the last COSMO-SAC refinement to date. This refinement can be divided into two major contributions: the electrostatic constant as a temperature dependent parameter ($C_{ES} = A_{ES} + B_{ES}/T$) and a division of the molecules surface into three parts, (1) the p_{nhb} profile, (2) the p_{OH} profile, corresponding to the surfaces around OH hydroxyl groups, and (3) the p_{OT} profile for the surfaces around O, F, or N atoms with H atoms connected to them (except OH groups).

Once again, this division allows a finer description of the interactions in the c_{hb} parameter of the ΔW term:

$$c_{hb}(\sigma_m^t, \sigma_n^s) = \begin{cases} c_{OH-OH} = 4013.78 \text{ kcal (mol Å}^4\text{)}^{-1}\, e^{-2} & \text{if s = t = OH and } \sigma_m^t\cdot\sigma_n^s < 0 \\ c_{OT-OT} = 932.31 \text{ kcal (mol Å}^4\text{)}^{-1}\, e^{-2} & \text{if s = t = OT and } \sigma_m^t\cdot\sigma_n^s < 0 \\ c_{OH-OT} = 3016.43 \text{ kcal (mol Å}^4\text{)}^{-1}\, e^{-2} & \text{if s = OH and t = OT and } \sigma_m^t\cdot\sigma_n^s < 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

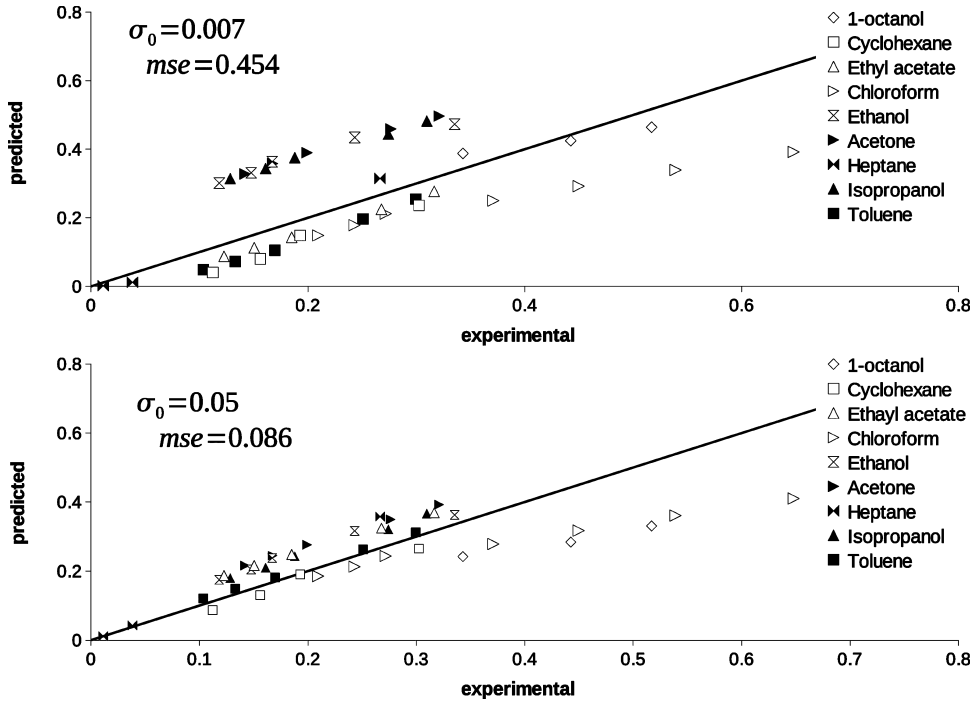

Figure 1. COSMO-SAC (2010) predicted solubility with lower hydrogen bonding possibility (for σ_0 = 0.05) versus experimental solubility for ibuprofen in different organic solvents with eq 3.

Whatever the COSMO-SAC version, the c_{hb} and σ_0 parameters were always calculated using VLE or LLE data. In the next section, an optimization for SLE predictions is suggested.

A COSMO-SAC OPTIMIZATION FOR SLE

In this section, we propose an optimization of refined COSMO-SAC (2007 and 2010) for SLE. In order to improve the model's accuracy for solubility predictions, we suggest a new optimization of the parameters using experimental solubility data (neither VLE nor LLE). The parameters considered are c_{hb} and σ_0.

In fact, because there is a difference in the complexity and size between the solvent and solute molecules, another modification about the σ-profiles is introduced here. Indeed, if the solute molecules are bigger and more complex, their ability to form hydrogen bonds should be different than that of small molecules. As previously shown,[16] in the case of complex organic molecules, interaction sites can be hidden or buried, so the probability to form hydrogen bonds should be smaller. Figure 1 shows that prediction accuracy is increased when the hydrogen bonding probability is decreased. That is why in this section two probability functions (screening probability) will be used depending on the considered molecules: P^{hb}_{solute} and $P^{hb}_{solvent}$ for solute or solvent, respectively. The first function has σ_0 for standard deviation, and the second σ_1. However, it can be suggested that the σ_1 calculated by Hsieh et al.[19] can be preserved (0.007). Indeed, the hydrogen bonding ability of the solvents (small molecules) should be the same as in VLE or LLE (same molecules involved).

To summarize, the parameters optimized in this section are (1) c_{hb} and σ_1 (COSMO-SAC 2007) and (2) c_{OH-OH}, c_{OT-OT}, c_{OH-OT}, and σ_1 (COSMO-SAC 2010).

To regress the parameters, 352 solid−liquid equilibria data points were considered (data from the literature[23−25]). In order to be representative of the complex organic molecules that are involved in drug solubility modeling, 130 complex molecules with various functional groups were chosen (aspirin, caffein, cortisone, prostaglandine, etc.). A wide range of classic organic solvents were also considered (35 different solvents). The list of the compounds is given in the appendix (Tables 6 and 7). About 24% of the considered equilibria data involved a nonpolar or aprotic solvent. The others involved a polar or a protic solvent.

Even if the quantity of data is smaller than the study of Hsieh et al.,[19] the main issue is to investigate the possibility of improving the model for SLE, discuss the relevancy of the different parameters, and give quite reliable parameters.

The regression procedure was the following: we performed solubility calculations using the thermodynamic equation for SLE (eq 3). Then, we used a mean square algorithm in order to minimize the error between calculated and experimental data in logarithm:

$$\text{lmse} = \frac{1}{n}\sum_{i}^{n}\left(\ln(x_{i\text{predicted}}) - \ln(x_{i\text{experimental}})\right)^2 \quad (8)$$

where n is the number of data points.

In order to optimize the convergence of the global optima, several initial estimates were considered. However, different sets of parameters can lead to quite similar lmse. The results are presented in Table 1.

This table shows the improvement of the optimization from equilibrium data. Original COSMO-SAC shows a lmse of 6.6878, while the optimized COSMO-SAC 2010 shows a lmse of 4.7411. The mean square errors are still significantly high because there are some equilibria that are difficult to accurately represent. Among them are (1) systems involving compounds with low solubility ($<10^{-3}$), such as sulfamethoxypyridazine, sulfadiazine, and piroxicam (these systems cause the lmse to increase (50% of the greatest errors at the end of the optimization), especially when the experimental solubility is very low (about 10^{-6} for caffeine in hexane, for example)),

Table 1. Initial and Optimized COSMO-SAC Parameters

model	LMSE
original COSMO-SAC	6.6878

	parameters			
model	C_{hb}	σ_0	σ_1	LMSE
COSMO-SAC 2007	3484.42	7×10^{-3}	7×10^{-3}	5.2298
COSMO-SAC 2007 (opt.)	1613.8	7×10^{-3}	2.75×10^{-3}	4.8474

	parameters					
model	C_{OH-OH}	C_{OT-OT}	C_{OH-OT}	σ_0	σ_1	LMSE
COSMO-SAC 2010	4013.78	932.31	3016.43	7×10^{-3}	7×10^{-3}	3.588
COSMO-SAC 2010 (opt. 1)	1501.1	8452.4	1816.2	8.97×10^{-3}	7×10^{-3}	4.7411

(2) systems involving nonpolar/aprotic solvents (40% of the greatest errors), and (3) systems involving water for COSMO-SAC 2007 optimization (70% among equilibria involving water).

It is also important to notice that the new parameters may not be the optimal ones (globally). Some other simulations led to different parameters, although not very different. The use of a much larger database could give the "best values".

In detail, interaction parameters changed for the two versions (2007 and 2010). In the model, there are two means of decreasing hydrogen bonding importance: a decrease of c_{hb} or an increase of σ_1. Our calculations led to a significant decrease of c_{hb}; c_{OH-OH} and c_{OH-OT} also decreased, whereas c_{OT-OT} more surprisingly significantly increased. Interactions of the kind OT−OT (dipole−dipole) take now a more significant part in the global quantification. This new balance shows that hydrogen bonds take a smaller importance in the model and that dipole−dipole interactions are revaluated.

For COSMO-SAC 2007, $\sigma_1 > \sigma_0$ and c_{hb} decreased. However, for COSMO-SAC 2010, this is not the case since $\sigma_1 < \sigma_0$ and c_{OH-OH} decreased at the same time (probably because the whole balance between c_{OH-OH}, c_{OT-OT}, and c_{OH-OT} was changed).

RESULTS AND DISCUSSION

Reference Molecules and Criteria. In this work, form I of paracetamol, ibuprofen, benzoic acid, salicylic acid, and 4-aminobenzoic acid were chosen as model drugs. These molecules were chosen because of their different functional groups (OH, COOH, NH, NH2). Anthracene was added to this list because of its simple structure. Their structures are presented in Figure 2.

The experimental data used in this study (thermodynamic properties and solubility) are taken from the literature[26−37,37,38,38−45] and from previous works.[15,16]

Uncertainties in the predictions due to the error in the pure solid properties were again investigated following a Monte Carlo method. These last errors were found to be about 17%, that is, inferior to the nonoptimized models (39%).

The results of this work will be analyzed following three main criteria: (1) The relative mean square errors (rmse) between the predicted and experimental solubility (eq 9),

$$\text{rmse} = \frac{1}{n}\sum_{i}^{n}\left(\frac{x_{i\text{predicted}} - x_{i\text{experimental}})}{x_{i\text{experimental}}}\right)^2 \quad (9)$$

(2) The temperature dependence, evaluated as the slope of the solubility curve dx/dT. It is a very interesting observation since it gives information on supersaturation calculations or crystallization yields. As suggested in a previous work,[16] the slope of the function $\ln(x) = k(1/T) + K$ (in which T is the temperature and k and K are two contants) is investigated. A correct k factor, combined with unique accurate solubility data, may lead to the whole solubility curve:

$$x_2 = x_1 \exp\left(\frac{k}{T_2} - \frac{k}{T_1}\right) \quad (10)$$

where x_1 and x_2 are the solubility at temperature T_1 and T_2, respectively. (3) The orders of magnitudes (in logarithm) and solubility ranking preservation. (4) The ability of the models to predict the influence of a new solvent on the solubility.

In all the figures comparing predicted and experimental data, the models errors from solid state property uncertainties are represented by error bars. The experimental errors are not represented since they are of the same order of magnitude as the symbol size.

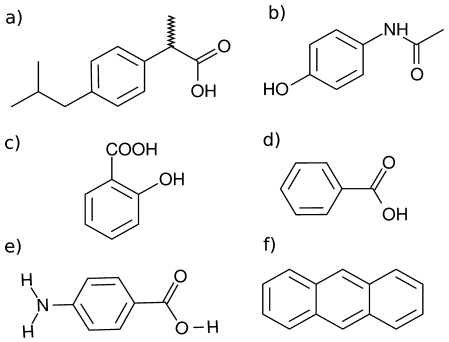

Figure 2. Structures of (a) ibuprofen, (b) paracetamol, (c) salicylic acid, (d) benzoic acid, (e) 4-aminobenzoic acid, and (f) anthracene.

Solubility in Pure Solvents. First, Table 2 shows that the optimized models give better results than previously.

Table 2. Mean Quadratic Errors (mse) of Various COSMO-SAC Predictions (Original, 2007, 2010, and Optimizations)

model	rmse[a]	mean standard deviation (%)
original COSMO-SAC	23.5 (24)	29
COSMO-SAC 2007	11.4 (10.4)	39
COSMO-SAC 2010	155 (5.4)	64
COSMO-SAC 2007 (opt.)	7.15 (7.20)	17
COSMO-SAC 2010 (opt.)	15.70 (15.70)	17

[a]Numbers in parentheses represent the rmse calculated without 4-aminobenzoic acid.

The relative mean square errors are lower and are much less sensitive to strong interactions (no noticeable influence of 4-aminobenzoic acid on the rmse, unlike the original models).

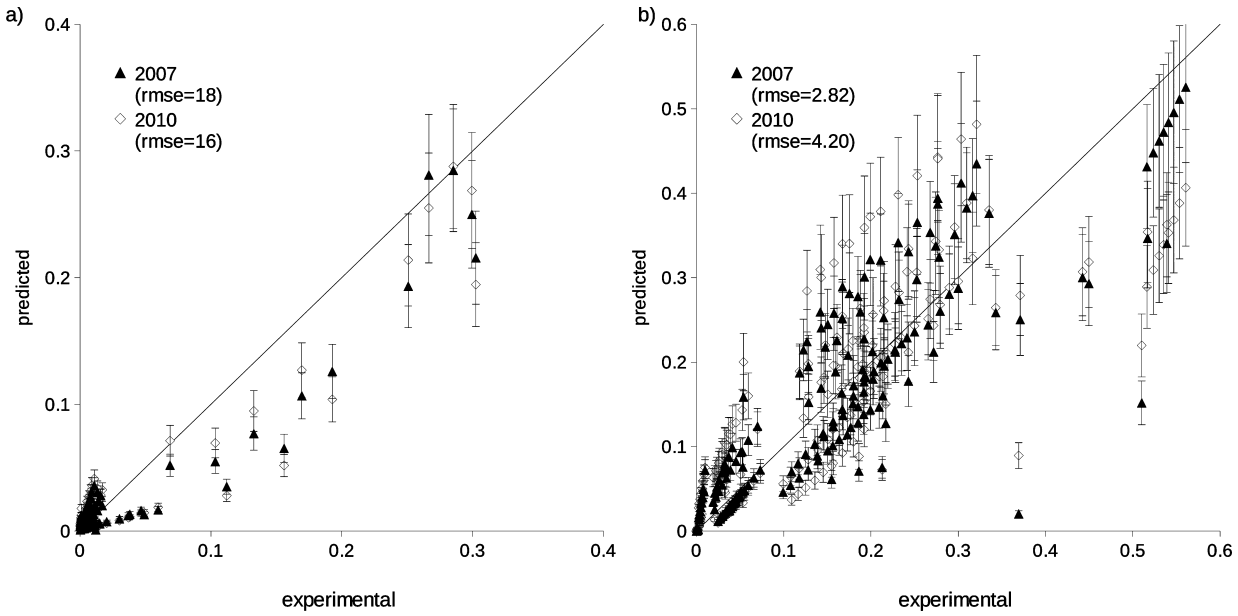

Figure 3. Optimized COSMO-SAC (2007 and 2010) predictions versus experimental solubility: (a) no hydrogen bonding or polar/polar interactions between solute and solvent and (b) possible interactions.

Table 3. Solubility Orders of Magnitude Experimental and Predicted Using Optimized COSMO-SAC (2007 and 2010), at 303.15 K[a]

experimental		2007		2010	
solvent	order of magnitude	solvent	order of magnitude	solvent	order of magnitude
		Ibuprofen			
heptane	[−3;−2]	heptane	[−4;−3.5]	heptane	[−4;−3.5]
cyclohexane	[−2;−1.5]	cyclohexane	[−3;−2.5]	cyclohexane	[−3]
ethanol	[−1.5;−1]	toluene	[−2;−1.5]	toluene	[−1.5]
toluene		chloroform	[−1.5;−1]	ethyl acetate	[−1.5;−1]
ethyl acetate		octanol		octanol	
isopropanol		ethanol		chloroform	[−1]
acetone		isopropanol		ethanol	
octanol	[−1;−0.5]	ethyl acetate	[−1]	isopropanol	[<1]
chloroform		acetone	[−1;−0.5]	acetone	
		Benzoic Acid			
hexane	[−4.5;−4]	cyclohexane	[−5.5;−5]	cyclohexane	[−5.5;−5]
cyclohexane		heptane		heptane	
heptane		hexane		hexane	
acetonitrile	[−3;−2.5]	carbon tetrachloride	[−4.5;−4]	carbon tetrachloride	[−4;−3.5]
carbon tetrachloride		benzene	[−3;−2.5]	benzene	[−2.5;−2]
benzene		DMSO	[−2;−1.5]	octanol	[−2;−1.5]
acetone	[−1.5;−1]	acetonitrile		acetonitrile	
octanol		octanol		butanol	[−1.5;−1]
isopropanol		butanol	[−1.5;−1]	isopropanol	
butanol		isopropanol		DMSO	
dioxane		dioxane		dioxane	
DMSO	[−1;−0.5]	acetone		*N*-methylpyrrolidone	
N-methyl pyrrolidone		N-methylpyrrolidone	[−1;−0.5]	acetone	[−1;−0.5]

[a]For ibuprofen and benzoic acid in solubility logarithm.

Also, optimized COSMO-SAC 2007 seems to give the most accurate results in the studied mixtures.

Figure 3 shows that optimized models give good accuracy when hydrogen bonding can occur (rmse between 2.8 and 4.2). This observation tends to validate the optimization, even if the predicted results do not exactly match the experimental ones. In the other case (no hydrogen bonds or dipole−dipole interactions between solute and solvent), rmse are higher (between 16 and 19).

More specifically, quantitative results in aprotic and apolar solvents are not improved compared with the former models. It is not surprising since the electrostatic parameters were not

changed. This observation and the poor results observed for anthracene show the method's limits when there are no complex interactions such as hydrogen bonding or dipole–dipole.

Table 3 shows the solubility ranking and predicted solubility orders of magnitude compared with the experimental ones. The improvement is not as good as expected. Despite the optimization, solubility in polar and aprotic solvent still presents some issues (see benzoic acid in acetone or acetonitrile). However, the 2010 optimization shows an actual improvement compared with the nonoptimized model.

The most interesting results concern the solubility temperature dependence. Indeed, it is difficult to judge quantitatively the models. The relative mean square errors may give some clues to evaluate globally the results. But, there are still some weaknesses, such as uncertainties introduced by the thermodynamic properties of the solids, or the quality of the data used for the optimization. That is why, in this work, we looked more precisely at the ability of the models to predict the solubility temperature dependency.

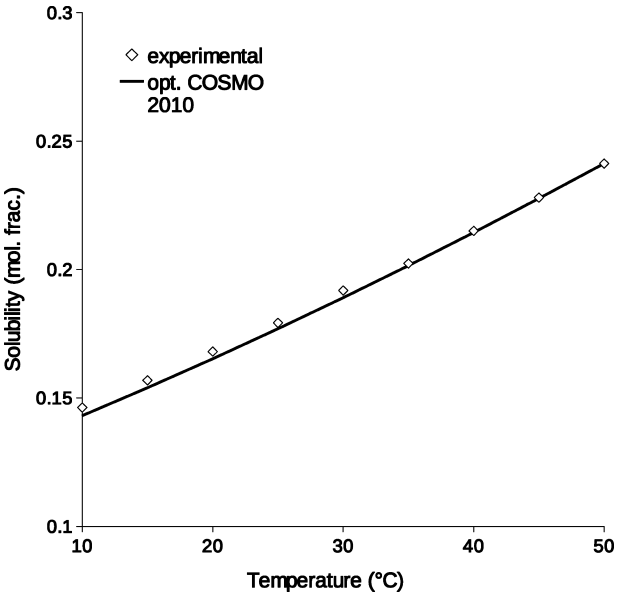

Figure 4. Mole fraction solubility calculated using COSMO-SAC 2007 of salicylic acid in acetone using one experimental data point at 50 °C compared with experimental solubility as a function of the temperature.

Solubility Temperature Dependence. The predicted solubility temperature dependence is in good agreement with the experimental data (Figure 4 and Table 5). The k factors (eq 10) are in better agreement with experimental values than with the different original COSMO-SAC models (see Table 4).

Moreover, there is a great improvement in polar solvents (Table 5, except paracetamol in ethanol), especially using the 2010 optimized version. So, it seems possible to obtain the solubility of a complex organic molecule at any temperature, knowing one solubility data point at a given temperature, with a precision lower than 30% (Figure 4). The only problem occurs in the case of anthracene, for which there is no hydrogen bonding phenomenon but only electrostatic interactions between instantaneous dipoles. This last observation strenghtens the idea that the electrostatic parameter could be inappropriate for solubility predictions in its actual form.

Solubility in Mixed Solvents. To check the accuracy of COSMO-SAC for solubility prediction in mixed solvents, two cases were investigated: (1) the solubility increases monotonically (ibuprofen in ethanol with n-heptane or propylene glycol) and (2) a second solvent in the mixture implies a synergic effect on solubility (maximum in the solubilty curve, for example, paracetamol in acetone/water).

Figure 5 shows that the optimized version of COSMO-SAC 2007 (but actually not the 2010) improves the prediction compared with the original versions (see heptane/ethanol). As before,[16] the better the solubility prediction in the pure solvents, the better the results in mixed solvents when no synergy effects are observed.

In the second case (Figure 6), the results do not show at first an improvement of the COSMO-SAC optimizations. The 2010 version is even less accurate in the representation of the solubility maximum. Another optimization upon 85 equilibrium data (with mostly polar solvents) led to better results. The same conclusion can be found with other mixtures,

Table 4. Experimental and Predicted Temperature Dependence: k Coefficient (Experimental Relative Error %), and Regression Coefficient

		experimental		opt. COSMO-SAC 2007		opt. COSMO-SAC 2010	
product	solvent	k	reg. coeff.	k	reg. coeff.	k	reg. coeff.
ibuprofen	toluene	−3700	0.998	−5085 (38)	0.997	−4600 (24)	0.997
	acetone	−2870	0.999	−1825 (35)	0.999	−1560 (45)	0.999
	ethanol	−3460	0.981	−2440 (29)	0.999	−2440 (29)	0.999
paracetamol	acetone	−1870	0.999	−2120 (13)	0.999	−1760 (6)	0.999
	ethanol	−1467	0.998	−2690 (83)	0.999	−2750 (88)	0.999
	propanol	−1700	0.999	−2960 (74)	0.999	−3055 (80)	0.999
salicylic acid	acetone	−1140	0.999	−1652 (45)	0.999	−1195 (5)	0.999
	acetonitrile	−2600	0.999	−2575 (87)	0.999	−2095 (20)	0.999
	methanol	−1500	0.999	−2530 (68)	0.999	−2315 (54)	0.999
benzoic acid	cyclohexane	−6230	0.995	−4845 (22)	0.976	−5230 (16)	0.977
	acetone	−1700	0.999	−1205 (29)	0.999	−970 (43)	0.999
	isopropanol	−1690	0.999	−1675 (1)	0.999	−1640 (3)	0.999
	octanol	−2160	0.994	−2000 (7)	0.994	−2000 (7)	0.993
anthracene	toluene	−3040	0.992	−3480 (15)	0.999	−3455 (14)	0.999
	MEK	−4580	0.952	−3180 (31)	0.999	−3025 (34)	0.999
	isopropanol	−4880	0.938	−3670 (25)	0.999	−3645 (25)	0.999

Table 5. Experimental and Predicted (eq 10) Solubilities at Temperature T_1 and T_2 for Different Mixtures and Relative Errors

product	solvent	T_1 at T_1 (°C)	T_2 at T_2 (°C)	exptl (frac. mol)	exptl (frac. mol)
benzoic acid	isopropanol	49.57	4.91	0.297	0.129
salicylic ac.	acetone	50	10	0.241	0.146
ibuprofen	acetone	35	10	0.321	0.142
paracetamol	ethanol	20	0	0.055	0.038
paracetamol	acetone	20	0	0.041	0.021
anthracene	toluene	50	20	0.0159	0.0061

original COSMO-SAC		original 2007		original 2010		opt. 2007		opt. 2010	
pred. at T_2 (frac. mol.)	rel. error (%)	pred. at T_2 (frac. mol.)	rel. error (%)	pred. at T_2 (frac. mol.)	rel. error (%)	pred. at T_2 (frac. mol.)	rel. error (%)	pred. at T_2 (frac. mol.)	rel. error (%)
0.167	29	0.152	18	0.179	39	0.129	0	0.131	1.7
0.181	24	0.160	10	0.134	9	0.117	20	0.143	2
0.244	72	0.222	56	0.211	48	0.190	34	0.205	45
0.037	3	0.031	20	0.028	25	0.028	26	0.028	28
0.025	21	0.023	11	0.019	8	0.019	8.5	0.22	4
0.0052	13	0.0053	13	0.0053	13	0.0053	13	0.0053	13

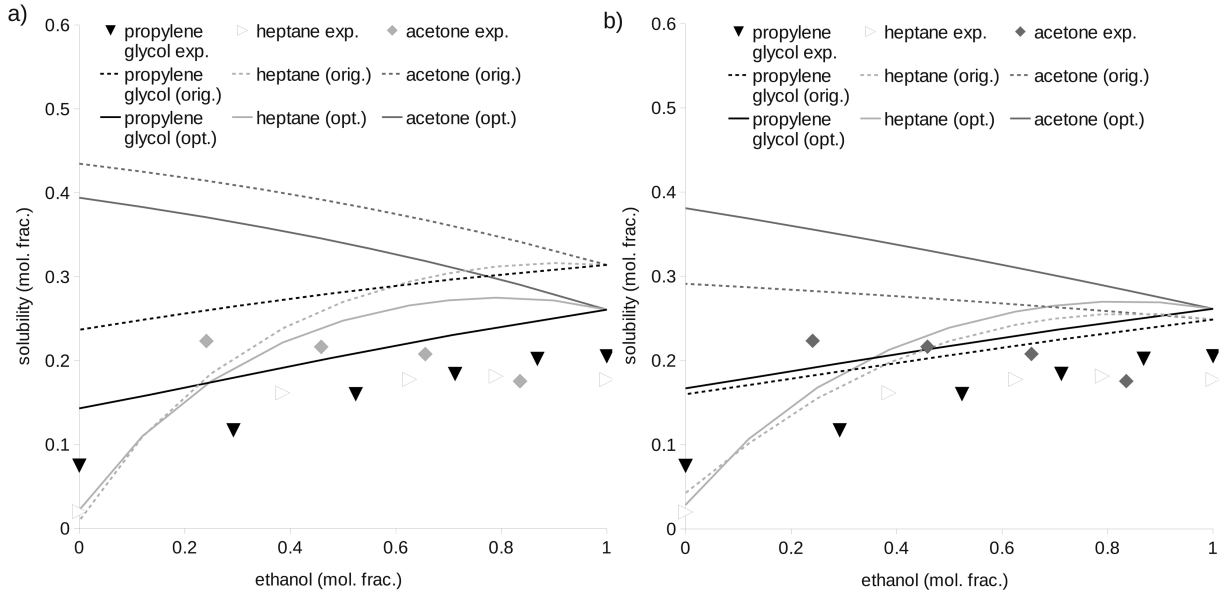

Figure 5. Experimental and predicted solubility with (a) optimized COSMO-SAC 2007 and (b) optimized COSMO-SAC 2010 of ibuprofen in mixed solvents as a function of the solvent composition.

like paracetamol in dioxane/water or salicylic acid in ethyl acetate/water. The improvement of the hydrogen bonding description may allow the models to be interesting for the prediction of the solubility maxima but not with the optimized parameter from this study. Thus, this last conclusion is not very encouraging for solvent mixtures, or maybe it would need specific parameters (especially with the presence of water, highly polar).

Electrostatic Parameter. The previous sections have shown the weaknesses of the optimized models when weak interactions (van der Waals between two instantaneous dipoles) are predominant. These interactions are taken into account in the electrostatic parameter C_{ES}. In this work, we propose to investigate the relevancy of this parameter in its actual form for solubility predictions.

Two solutions can be considered: optimizing A_{ES} and B_{ES} using equilibria where there is no hydrogen bonding possibility or optimizing all the parameters at the same time. In this paper, the first case is studied.

A_{ES} and B_{ES} were regressed with anthracene solubility data in toluene at 25 and 50 °C. In this case, the c_{hb} parameter has no influence. At 25 °C, C_{ES} must be -19500 kcal (mol Å^4)$^{-1}\cdot e^{-2}$ ($x = 6.11 \times 10^{-3}$). At 50 °C, C_{ES} must be -27000 kcal (mol Å^4)$^{-1}\cdot e^{-2}$ ($x = 1.59 \times 10^{-3}$). It leads to $A_{ES} = -61860$ kcal (mol Å^4)$^{-1}\cdot e^{-2}$ and $B_{ES} = -3.64 \times 10^{-9}$ kcal (mol Å^4)$^{-1}\cdot$(e^2 K^2)$^{-1}$.

In order to test these values, solubility predictions of anthracene in heptane were made (no hydrogen bonding). At 25 °C, we found 1.67×10^{-10} instead of 1.20×10^{-3}. These results show that the COSMO-SAC 2010 method is not suitable to represent cases where the c_{hb} parameter has no influence, at least in the simple case that was investigated.

It seems that the electrostatic (C_{ES}) and hydrogen bonding (c_{hb}) parameters might be dependent. Unless anthracene in toluene shows a very specific behavior, it can be suggested that there is a balance between them. Indeed, the electrostatic parameter improvement is complicated (the improvement for anthracene in toluene causes a decline for anthracene in heptane). Moreover, a quick attempt to optimize the A_{ES} and

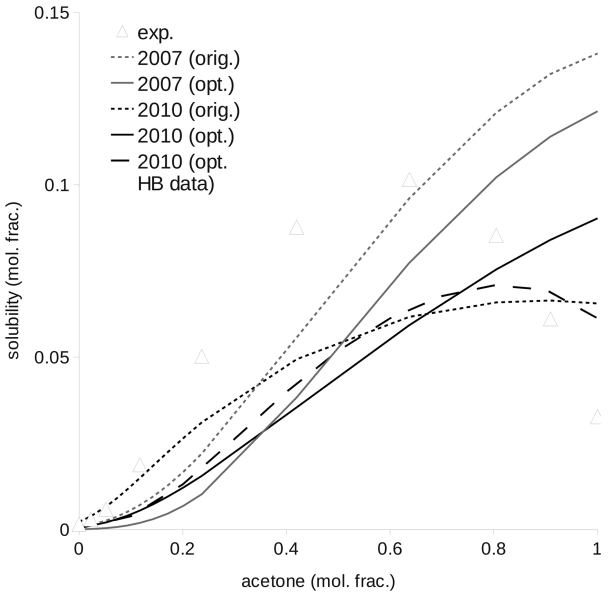

Figure 6. Experimental and predicted with optimized COSMO-SAC (2007 and 2010) solubility of paracetamol in acetone/water as a function of the initial acetone mole fraction.

B_{ES} parameters (not shown here) did not lead to significant improvement. Maybe the form of the ΔW term is not adapted to this kind of simulations.

In the end, because the hydrogen bonds or dipole−dipole interactions are predominant in SLE, this problem is negligible when considering complex molecules in polar solvents.

CONCLUSION

In this work, the most recent refinements of COSMO-SAC have been optimized for solubility predictions. The optimizations, based on experimental solubility data extracted from the literature, have led to new interaction parameters. The obtained hydrogen bonding parameters are lower than the original, and the probability to form hydrogen bonds between two segments are also lower. That is not surprising since the molecules involved are much more complex than the ones considered in VLE or LLE. For the molecules commonly considered in drug solubility, the interaction sites could be hidden or buried, and the probability to form hydrogen bonds should indeed be lower.

With the optimized COSMO-SAC models, the results are very interesting and more accurate than previously (Tables 1 and 2). The orders of magnitude are closer to the experimental, and the prediction of the temperature dependency is also improved.

Table 2 also shows that the uncertainties of the optimized models are smaller than that of the original models. The main difference found is the modeling of systems where weak interactions are predominant: instantaneous dipole interactions are not well quantified. It seems that the electrostatic parameter is considerably dependent on the hydrogen bonding parameters in the model.

Also, solubility predictions in mixed solvents are not in better agreement with the experimental values. More accurate results can be achieved by performing an optimization using only polar solvents, especially when water is involved.

In order to improve the results of this study, a larger database could be used, considering more complex molecules. This could lead to a most obvious global optimum.

Another trail could be the adjustment of the σ_0 parameter. Even if the COSMO-SAC model may disregard the free-energy character of hydrogen bonding by being too empirical, the size and flexibility of molecules could be included in the σ_0 parameter and be characteristic for a given molecule. This could improve the hydrogen bonding representation but would need a new optimization for each solute compound.

Last, the form of the molecular interactions should probably be changed in order to improve weak van der Waals or London dispersive interactions.

Nevertheless, the refinement and SLE optimization are relevant and might be useful for quick solubility estimations, or crystallization yield calculations.

APPENDIX: EXPERIMENTAL DATA USED FOR THE MODEL OPTIMIZATION

The mixtures used for parameter regression are presented in Table 6 and the list of solvents in Table 7.

Table 6. List of Compounds for COSMO-SAC Optimization (Actual Data from the Literature[23−25])

product	number of solvents
3-hydroxybenzoic acid	1
3-methylbenzoic acid	3
4-aminobenzoic acid	17
4-hexyl resorcinol	2
4-hydroxybenzoic acid	5
acetanilide	10
anthracene	8
antipyrine	3
aspirin	14
benzoic acid	20
butabarbital	4
butylaminobenzoate	10
butylparaben	9
caffein	3
citric acid	10
cortisone	2
ephedrin	7
estriol	5
ethyl-4-hydroxybenzoate	3
ethyl-p-aminobenzoate	8
fentanyl	2
flurbiprofen	15
fumaric acid	6
ganciclovir	3
hexachlorobenzene	4
ibuprofen	17
lidocaine	5
mefenamic acid	3
metharbital	1
methylaminobenzoate	11
nandrolone propionate	1
nicotinamide	4
paracetamol	19
phenacetin	13
piroxicam	16
propylaminobenzoate	8
propylhydroxybenzoate	8

Table 6. contiuned

product	number of solvents
prostaglandine	1
p-toluic acid	2
salicylic acid	19
sulfamerazine	2
sulfamethoxypyridazine	19
sulfadiazine	13
sulfapyridine	7
theobromine	3
theophylline	6

Table 7. List of Solvents for COSMO-SAC Optimization

1-heptanol
1-hexanol
1-octano
1-pentanol
1-propanol
acetic acid
acetone
acetonitril
benzene
carbontetrachloride
chloroform
cyclohexane
dichloromethane
diethyl ether
1,2-dimethoxy ethane
dioxane
DMF
DMS
ethanol
ethoxyethanol
ethyl acetate
ethylene glycol
glycerol
heptane
hexane
isopropropanol
isopropyl miristate
methanol
MEK
n-methyl pyrrolidone
n-butanol
o-xylene
propylene glycol
THF
toluene
water

AUTHOR INFORMATION

Corresponding Author

*E-mail: bouillot@emse.fr.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Gmehling, J. G.; Anderson, T. F.; Prausnitz, J. M. Solid-Liquid Equilibria Using Unifac. *Ind. Eng. Chem. Fundam.* **1978**, *17*, 269−273.

(2) Hahnenkamp, I.; Graubner, G.; Gmehling, J. Measurement and prediction of solubilities of active pharmaceutical ingredients. *Int. J. Pharm.* **2010**, *388*, 73−81.

(3) Tung, H.; Tabora, J.; Variankaval, N.; Bakken, D.; Chen, C. Prediction of Pharmaceuticcal Solubility Via NRTL-SAC and COSMO-SAC. *J. Pharm. Sci.* **2008**, *97*, 1813−1820.

(4) Mota, F. L.; Carneiro, A. P.; Pinho, S. P.; Macedo, E. A. Temperature and solvent effect in solubility of some pharmaceutical compounds: Measurements and modelling. *Eur. J. Pharm. Sci.* **2009**, *37*, 499−507.

(5) Chen, C.-C.; Song, Y. Solubility Modeling with a Nonrandom Two-Liquid Segment Activity Coefficient Model. *Ind. Eng. Chem. Res.* **2004**, *43*, 8354−8362.

(6) Gross, J.; Sadowski, G. Perturbed-Chain SAFT: An Equation of State Based on a Perturbation Theory for Chain Molecules. *Ind. Eng. Chem. Res.* **2001**, *40*, 1244−1260.

(7) Tsivintzelis, I.; Economou, I.; Kontogeorgis, G. Modeling the Solid−Liquid Equilibrium in Pharmaceutical-Solvent Mixtures: Systems with Complex Hydrogen Bonding Behavior. *AIChE* **2009**, *55*, 756−770.

(8) 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.

(9) Lin, S.; Sandler, S. A priori phase equilibrium prediction from a segment contribution solvation model. *Ind. Eng. Chem. Res.* **2002**, *41*, 899−913.

(10) Panayiotou, C. Partial Solvation Parameters and Mixture Thermodynamics. *J. Phys. Chem. B* **2012**, *116*, 7302−7321.

(11) Panayiotou, C. Equation-of-state models and quantum mechanics calculations. *Ind. Eng. Chem. Res.* **2003**, *42*, 1495−1507.

(12) Panayiotou, C. The QCHB model of fluids and their mixtures. *J. Chem. Thermodyn.* **2003**, *35*, 349−381.

(13) Panayiotou, C. Redefining solubility parameters: the partial solvation parameters. *Phys. Chem. Chem. Phys.* **2012**, *14*, 3882−3908.

(14) Panayiotou, C. Partial solvation parameters and LSER molecular descriptors. *J. Chem. Thermodyn.* **2012**, *51*, 172−189.

(15) Bouillot, B.; Teychené, S.; Biscans, B. An evaluation of thermodynamic models for the prediction of drug and drug-like molecule solubility in organic solvents. *Fluid Phase Equilib.* **2011**, *309*, 36−52.

(16) Bouillot, B.; Teychené, S.; Biscans, B. An Evaluation of COSMO-SAC Model and Its Evolutions for the Prediction of Drug-Like Molecule Solubility: Part 1. *Ind. Eng. Chem. Res.* **2013**, DOI: 10.1021/ie3015318.

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

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

(19) Hsieh, C.; Sandler, S.; Lin, S. Improvement of COSMO-SAC for Vapour-liquid and liquid-liquid equilibrium predictions. *Fluid Phase Equilib.* **2010**, *297*, 90−97.

(20) Hsieh, C.; Wang, S.; Lin, S.; Sandler, S. A Predictive Model for the Solubility and Octanol−Water Partition Coefficient of Pharmaceuticals. *J. Chem. Eng. Data* **2011**, *56*, 936−945.

(21) Prausnitz, J.; Lichtenthaler, R.; de Azevedo, E. *Molecular Thermodynamics of Fluid-Phase Equilibria*; Prentice Hall PTR: Englewood Cliffs, NJ, 1998.

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

(23) Marrero, J.; Abildskov, J. *Solubility and Related Properties of Large Complex Chemicals - Part 1: Organic Solutes ranging from C_4 to C_{40}*; DECHEMA: Frankfurt am Main, Germany, 2003.

(24) Abildskov, J. *Solubility and Related Properties of Large Complex Chemicals - Part 2: Organic Solutes ranging from C_2 toC_{41}*; DECHEMA: Frankfurt am Main, Germany, 2005.

(25) Jouyban, A. *Handbook of Solubility Data for Pharmaceuticals*; CRC Press: Boca Raton, FL, 2010.

(26) Garzon, L. C.; Martinez, F. Temperature dependence of solubility for ibuprofen in some organic and aqueous solvents. *J. Solution Chem.* **2004**, *33*, 1379−1395.
(27) Gracin, S.; Rasmuson, A. Solubility of phenylacetic acid, p-hydroxyphenylacetic acid, p-aminophenylacetic acid, p-hydroxybenzoic acid, and ibuprofen in pure solvents. *J. Chem. Eng. Data* **2002**, *47*, 1379−1383.
(28) Granberg, R.; Rasmuson, A. Solubility of paracetamol in pure solvents. *J. Chem. Eng. Data* **1999**, *44*, 1391−1395.
(29) De Fina, K.; Sharp, T.; Roy, L.; Acree, W. Solubility of 8-hydroxybenzoic acid in select organic solvents at 298.15 K. *J. Chem. Eng. Data* **1999**, *44*, 1262−1264.
(30) Fung, H.; Higuchi, T. Molecular interactions and solubility of polar nonelectrolytes in nonpolar solvents. *J. Pharm. Sci.* **1971**, *60*, 1782−1788.
(31) Nordstrom, F. L.; Rasmuson, A. C. Solubility and melting properties of salicylic acid. *J. Chem. Eng. Data* **2006**, *51*, 1668−1671.
(32) Paruta, A.; Sciarrone, B.; Lordi, N. Solubility of Salicylic Acid as a Function of Dielectric Constant. *J. Pharm. Sci.* **1964**, *53*, 1349−1353.
(33) Shalmashi, A.; Eliassi, A. Solubility of salicylic acid in water, ethanol, carbon tetrachloride, ethyl acetate, and xylene. *J. Chem. Eng. Data* **2008**, *53*, 199−200.
(34) Acree, W.; Bertrand, G. Thermochemical Investigations of Nearly Ideal Binary Solvents. 7. Monomer and Dimer Models for Solubility of Benzoic-Acid in Simple Binary and Ternary Solvents. *J. Pharm. Sci.* **1981**, *70*, 1033−1036.
(35) Beerbower, A.; Wu, P.; Martin, A. Expanded Solubility Parameter Approach 0.1. Naphthalene And Benzoic-Acid In Individual Solvents. *J. Pharm. Sci.* **1984**, *73*, 179−188.
(36) Li, D.; Liu, D.; Wang, F. Solubility of 4-methylbenzoic acid between 288 and 370 K. *J. Chem. Eng. Data* **2001**, *46*, 234−236.
(37) Lin, H.; Nash, R. An Experimental Method for Determining the Hildebrand Solubility Parameter of Organic Nonelectrolytes. *J. Pharm. Sci.* **1993**, *82*, 1018−1025.
(38) Yalkowski, S.; Valvani, S.; Roseman, T. Solubility and partitioning VI: Octanol solubility and octanol-water partition coefficients. *J. Pharm. Sci.* **1983**, *72*, 866−870.
(39) Hancock, C.; Pawlowski, J.; Idoux, J. *J. Org. Chem.* **1966**, *31*, 3801.
(40) Daniels, C.; Charlton, A.; Wold, R.; Moreno, R.; Acree, W.; Abraham, M. Mathematical correlation of 4-aminobenzoic acid solubilities in organic solvents with the Abraham solvation parameter model. *Phys. Chem. Liq.* **2004**, *42*, 633−641.
(41) Cepeda, E.; Gomez, B. Solubility of Anthracene and Anthaquinone in Some Pure and Mixed Solvents. *J. Chem. Eng. Data* **1989**, *34*, 273−275.
(42) Cepeda, E.; Diaz, M. Solubility of Anthracene and Anthaquinone in Acetonitrile, methyl ethyl ketone, isopropyl alcohol and their mixtures. *Fluid Phase Equilib.* **1996**, *121*, 267−272.
(43) Al-Sharrah, G.; Ali, S.; Fahim, M. Solubility of anthracene in two binary solvents containing toluene. *Fluid Phase Equilib.* **2002**, *193*, 191−201.
(44) Powell, J.; Acree, W. Solubility of Anthracene in Binary Alcohol + Dibutyl Ether Solvent Mixtures. *J. Chem. Eng. Data* **1995**, *40*, 914−916.
(45) Zvaigzne, A.; Acree, W. Thermochemical Investigations of Molecular Complexation. Estimation of Anthracene-Ethyl Acetate and Anthracene-Diethyl Adipate Association Parameters from Measured Solubility Data. *Phys. Chem. Liq.* **1991**, *24*, 31−42.