Solubility of pharmaceuticals: A comparison between SciPharma, a PC-SAFT-based approach, and NRTL-SAC*

Baptiste Bouillot[1], Theodora Spyriouni[2,a], Sébastien Teychené[3], and Béatrice Biscans[3]

[1] École Nationale Supérieure des Mines SPIN-EMSE, CNRS: UMR5307, LFG, 42023 Saint-Etienne, France
[2] Scienomics SARL, 17 square Edouard VII, 75009 Paris, France
[3] Université de Toulouse, CNRS: UMR5503, LGC, 31432 Toulouse Cedex 4, France

Abstract. The solubility of seven pharmaceutical compounds (paracetamol, benzoic acid, 4-aminobenzoic acid, salicylic acid, ibuprofen, naproxen and temazepam) in pure and mixed solvents as a function of temperature is calculated with SciPharma, a semi-empirical approach based on PC-SAFT, and the NRTL-SAC model. To conduct a fair comparison between the approaches, the parameters of the compounds were regressed against the same solubility data, chosen to account for hydrophilic, polar and hydrophobic interactions. Only these solubility data were used by both models for predicting solubility in other pure and mixed solvents for which experimental data were available for comparison. A total of 386 pure solvent data points were used for the comparison comprising one or more temperatures per solvent. SciPharma is found to be more accurate than NRTL-SAC on the pure solvent data used especially in the description of the temperature dependence. This is due to the appropriate parameterization of the pharmaceuticals and the temperature-dependent description of the activity coefficient in PC-SAFT. The solubility in mixed solvents is predicted satisfactorily with SciPharma. NRTL-SAC tends to overestimate the solubility in aqueous solutions of alcohols or shows invariable solubility with com-position in other cases.

1 Introduction

The solubility of active pharmaceutical ingredients (API) needs to be known at different stages of the product and process development. At early stages, solubility and other important parameters for drug bioavailability such as ionization, permeability and lipophilicity are screened for the drug-likeness of new pharmaceutical molecules. At later stages of process development and optimization, the appropriate selection of

solvents and the variation of solubility with temperature are critical for the crystallization and the process design of manufacturing. Therefore, the solubility in a large number of solvents and mixtures is needed to be known with reasonable accuracy. Pharmaceutical companies measure the solubility in several solvents and mixtures and subsequently use these data to feed thermodynamic models capable of correlating solubility data of pharmaceuticals. Thus, they reduce the number of experiments needed to screen solvents.

Although there are many models for calculating phase equilibrium, few of them are designed for solid-liquid equilibrium (SLE) calculations. The models for phase equilibrium can generally be divided into two types: equation of state (EoS) models like cubic equations and higher order, and activity coefficient models like NRTL [1], and UNIFAC [2]. There are some important difficulties in the representation of the liquid and solid phase, especially when complex molecules as APIs are involved. For solubility predictions, models have to take into account the weakest intermolecular interaction (vdW) as well as hydrogen bonding in the liquid phase in the case of big and complex molecules including many chemical groups. Models that have been used to correlate SLE data include UNIFAC, the NRTL segment activity coefficient (NRTL-SAC) model [3], the PC-SAFT EoS [4], the lattice model non-random hydrogen bonding theory (NRHB) [5] and the conductor-like screening model (COSMO-RS) [6,7].

Gracin et al. [8] used UNIFAC to calculate the solubility of solid organic compounds in water and organic solvents. They concluded that the additive assumption of the approach is not sufficiently accurate since properties of a functional group depend on the rest of the molecule. Hahnenkamp et al. [9] compared UNIFAC, modified UNIFAC (Dortmund) [10] and COSMO-RS for three pharmaceutical molecules. They reported that the UNIFAC methods were more accurate than COSMO-RS and that modified UNIFAC was able to predict the solvent with the highest solubility for two of the pharmaceuticals. Bouillot et al. [11] used the modification of COSMO-SAC on the molecule parameterization, and suggested some optimization of the method [12]. The improvements were promising but the accuracy of the method was still not convincing, except for quick solubility estimations.

The NRTL-SAC model has been widely used in the pharmaceutical industry the last decades. It is based on polymer NRTL and correlates satisfactorily experimental solubility data by using four parameters per molecule. These parameters are conceptual segments that describe the effective surface interactions between a solvent and a solute. They are fitted on experimental solubility data in pure and mixed solvents. Mullins et al. [13] and Tung et al. [14] compared NRTL-SAC to COSMO-SAC. They both found that NRTL-SAC that uses experimental data to fit parameters provides more accurate results for drug solubility. According to Mullins et al. [13] the accuracy of COSMO-SAC depends on the molecular conformation but in a non-conclusive way, i.e., trying different conformers may not improve the accuracy of COSMO-RS. Mota et al. [15] used A-UNIFAC, UNIFAC and NRTL-SAC to predict solubility in pure solvents for a set of drug-like molecules. Their suggestion was to use NRTL-SAC rather than UNIFAC, unless the necessary data for NRTL-SAC are not available. Sheikholeslamzadeh and Rohani [16] also found better performance of NRTL-SAC compared to UNIFAC for solubility prediction of pharmaceuticals in pure solvents. Bouillot et al. [17] compared UNIFAC, NRTL-SAC and COSMO-SAC for five drug molecules. They found that NRTL-SAC was better than the other models although predictions were sensitive on the data used for the parameterization. They also observed that all models failed to correctly describe the solubility dependence with temperature and that an equation of state would be more appropriate for that.

Recently, the PC-SAFT EoS was applied to pharmaceutical solubility calculations [4,18]. It was shown before [18] that the appropriate parameterization of the pharmaceuticals based on some experimental data can capture adequately the

solubility in pure and mixed solvents without the need of adjustable parameters. In the present study, the parameterization scheme for pharmaceuticals has been extended. Pharmaceuticals are categorized according to their solubility in alcohols and ketones or esters. Additionally, scaling factors are derived from the solubility data that feed the model in order to scale the calculated solubility in various solvents. This empirical approach based on PC-SAFT is called SciPharma and has been implemented in the MAPS platform of Scienomics.

In this work, we show a comparison between SciPharma and the NRTL-SAC model. Both models are semi-empirical as they are fed by experimental data. What is particularly interesting in these models is the possibility to model an API with a set of parameters and then calculate the solubility in any other solvent or mixture of solvents. In this work the following seven APIs were considered: paracetamol, benzoic acid, 4-aminobenzoic acid, salicylic acid, ibuprofen, naproxen and temazepam. To conduct a fair comparison between the models we parameterized the APIs using the same solubility data on the following solvents: water, methanol, ethanol, ethyl acetate, acetone, cyclohexane. These solvents cover the spectrum of hydrophilic, polar and hydrophobic interactions needed by both models for appropriate parameterization of pharmaceuticals. The obtained parameters were then used to calculate solubility in other pure solvents and mixtures for which we were able to find experimental data for comparison. The solubility as a function of temperature is, also, examined.

2 Theory

2.1 Solid-liquid equilibrium

Pharmaceuticals are crystalline solids at room temperature. Therefore, the solubility in a liquid solvent can be described by the solid-liquid equilibrium equation [19]. By assuming that the solid phase consists of pure pharmaceutical, the following simplified equation is reached:

$$\ln\left(\frac{1}{x_i}\right) = \ln\gamma_i + \frac{\Delta H_m}{RT_m}\left(\frac{T_m}{T} - 1\right) \tag{1}$$

where x_i and γ_i are the mole fraction and the activity coefficient of the pharmaceutical in the liquid solvent. $\Delta \mathrm{H}_m$ and T_m are the melting enthalpy and temperature, respectively, R is the gas constant and T the temperature. For the derivation of (1) it has been assumed that the difference of the pharmaceutical heat capacity in the solid and liquid phase can be neglected.

Equation (1) shows that the solubility of pharmaceuticals is determined by the activity coefficient and the pure component properties (the melting enthalpy and temperature). The activity coefficient is a measure of the non-ideal behavior between the pharmaceutical and solvent molecules and can be calculated from an appropriate model, such as NRTL-SAC [3] or PC-SAFT [20].

2.2 NRTL-SAC model

The NRTL-SAC model [3] is a semi-predictive method based on polymer-NRTL [21]. NRTL-SAC characterizes the molecules in terms of conceptual segments. Chen and Song [3] defined four types of segments: hydrophobic (X), repulsive (Y^-) and attractive (Y^+) polar, and hydrophilic (Z). Each component is represented by a quadruplet

[XY^-Y^+Z] weighing the contribution of each segment. Like in UNIFAC [22], the activity coefficient of component I can be written as the sum of a combinatorial and a residual term:

$$\ln\gamma_I = \ln\gamma_I^C + \ln\gamma_I^R. \tag{2}$$

The combinatorial contribution γ_I^C is calculated from the Flory-Huggins approximation for the entropy of mixing:

$$\ln\gamma_I^C = \ln\frac{\Phi_I}{x_I} + 1 - r_1 \sum_J \frac{\Phi_J}{r_J} \tag{3}$$

where r_I and Φ_I are the total segment number and the segment mole fraction of component I, respectively.

The residual term $\gamma_I^R$is calculated as the sum of the local composition (lc) interaction contribution of each segment [21]

$$\ln\gamma_I^R = \ln\gamma_I^{lc} = \sum_k r_{k,I}\left[\ln\Gamma_k^{lc} - \ln\Gamma_{k,I}^{lc}\right] \tag{4}$$

where $\ln\Gamma_k^{lc}$and $\ln\Gamma_{k,I}^{lc}$ are the activity coefficients of segment k in the mixture and in the pure component I, respectively.

The segment–segment interaction parameters and the conceptual segment values of solvents are determined by the regression of experimental vapor-liquid and liquid-liquid equilibrium data [3]. The segment values for a solute are obtained by the regression of solubility data in at least four solvents: a hydrophilic one (i.e. water), a polar attractor (i.e. alcohol), a polar donor (i.e. ketone), and a hydrophobic solvent (i.e. n-alkane). For a few selected compounds, segment values can be found in the literature [15,17,23].

Once the segment values of the solute are obtained, they can be used to predict solubility in other solvents or solvent mixtures. However, the solubility predictions are very sensitive to the solubility data used for the regression of parameters. A method for the choice of solvents has been suggested by Bouillot et al. [17].

2.3 SciPharma: a PC-SAFT based approach

The SciPharma approach is based on PC-SAFT. In PC-SAFT the residual Helmholtz energy a^{res} of a fluid, is written as the sum of the Helmholtz free energy of a reference fluid and a perturbation term, by using the perturbation theory for fluids. The reference fluid for PC-SAFT is the hard chain fluid while perturbations account for dispersion forces and hydrogen bonding:

$$a^{res} = a^{hc} + a^{disp} + a^{assoc} \tag{5}$$

where a^{hc}, a^{disp} and a^{assoc} represent the hard chain, dispersion and association terms. The functional forms for these terms can be found in the original publication [20]. PC-SAFT is extended to mixtures using standard mixing rules [24].

In the context of PC-SAFT each component is described with five parameters that include the segment number m, the segment diameter σ, the segment dispersion energy ε/k_B, the association energy ε_{hb}/k_B, and the association volume κ_{hb}. The last two parameters are needed for components that form hydrogen bonds. For the case of solvents, these parameters are fitted to vapor pressure and saturated liquid density from low temperature up to close to the critical point. All of the pharmaceuticals examined here contain functional groups that hydrogen bond, either with other

Table 1. Multiplying factors for the actual solubility in water in order to calculate the target solubility in water used in the parameterization.

Aqueous solubility (mg/g)	Multiplying factor
<1.E-03	10000
1.E-03 << 1.E-02	1000
1.E-02 << 1.E-01	500
1.E-01 << 1	100
1 << 10	7.5
>10	10

pharmaceutical molecules or with solvent molecules. For simplicity, all pharmaceutical molecules are assumed to have 4 associating sites of equal strength (two electron acceptor and two electron donor). As Ruether and Sadowski [4] pointed out, κ_{hb} has a relatively little effect on the solubility calculations and it was set equal to 0.01. Thus, the number of pure component parameters for regression practically reduces to four.

The regression of parameters is the most important step for achieving good accuracy in solubility calculations without the need of binary interaction parameters. The main features of SciPharma that differentiate it from the classical approach with PC-SAFT are the parameterization of pharmaceuticals and the calculated solubility that might be scaled depending on the available data. The parameterization of pharmaceuticals was discussed in [18]. This approach has been further refined in SciPharma. Pharmaceuticals are classified in families according to their solubility in polar solvents such as alcohols (methanol, ethanol) and ketones (methyl ethyl ketone, methyl isobutyl ketone) or esters (ethyl acetate). We have, empirically, found that pharmaceuticals with high solubility in methanol, higher than approximately 30 mg/g-solvent belong in the same family. All pharmaceuticals in the present study belong in this category. For these molecules, the parameterization uses methanol as the polar solvent. The regression of pharmaceutical parameters requires the solubility in water, at least one polar solvent and at least one hydrophobic solvent. The hydrophobic solvent in the present study is cyclohexane. As discussed in [18] the solubilities in water and in the hydrophobic solvent, used for the regression (target solubilities), need to be adjusted. The target solubility in water is multiplied by a factor given in Table 1 depending on the actual solubility in water. The water solubility affects mainly the size parameters m and σ of the pharmaceutical. By increasing the target solubility we basically decrease the size of the molecule. Modelling the pharmaceuticals as small molecules with high interaction energy per segment seems to be important to capture their solubility in polar and associative solvents. The solubility in cyclohexane, like in heptane [18], is scaled up to the closest power of 10, i.e., if the actual solubility is less than 0.1 the target solubility is set to 0.1, if the actual value is between 0.1 and 1 the target is set to 1, and so on. This might be the necessary counteraction of the increase in the water solubility. The target solubility in the polar solvent remains unscaled.

The rest of the solubility data are used in order to screen the multiple parameter sets produced by the regression against the three solvents (water, methanol, cyclohexane). We will call the three solvents regression solvents and the rest screen solvents. The parameter set that gives the lowest error for the screen solvents is retained. For the pharmaceutical family examined in this work, a scaling constant (ScC) is calculated from the ratio of the experimental solubility in methanol versus the solubility in ethanol, expressed in mg/g-solvent. ScC is an important parameter for scaling the solubilities calculated with PC-SAFT for specific solvents. For the family examined in

this work, ScC scales down the calculated solubilities in ketones, namely methyl ethyl ketone and methyl isobutyl ketone that are commonly used solvents in the pharmaceutical industry. The scaling rules described here refer only to the pharmaceutical family examined in this work. Discussion on other pharmaceutical families will be the subject of another publication.

Apart from ScC that is characteristic for each pharmaceutical, scaling factors are also calculated for the screen solvents as the ratio of the experimental value and the calculated value with the retained parameter set for the pharmaceutical. The parameter set might not reproduce exactly the solubility in the screen solvents. Therefore, the ratios between the calculated and the experimental values are calculated to correct accordingly the calculated solubilities. These can be thought as effective k_{ij} parameters. The ratio of the experimental versus the calculated solubility in ethanol scales also the calculated solubilities of higher alcohols up to butanol. It has been observed that for higher, than butanol alcohols, scaling the calculated values does not improve the prediction. The scaling factor for ethyl acetate (experimental versus calculated value) scales also the calculated solubility in other esters like propyl acetate. Thus, the available solubility data are used to improve the predictions in other solvents of the same family. The classification for the solvent families is empirical and does not rely necessarily on the chemical families. It has been built on the compilation of available data. If no relevant data are provided, the scaling ratio is equal to one. Finally, the elevated solubility in the hydrophobic solvent, used for the regression, needs to be scaled down to the actual value. The correction also applies to other hydrophobic solvents like hexane, heptane and so on.

For calculating the solubility in aqueous mixtures, the high solubility in water imposed by the parameterization scheme needs to be brought down to its original value. For this purpose, a binary interaction parameter k_{ij} between the pharmaceutical and water is used [18]. The scaling factors calculated for the pure solvents are transferred also to the mixtures. For mixed solvents at specific composition the calculated solubility is scaled by the arithmetic average of the scaling factors for the solvents weighted by the molar composition of the mixture. For aqueous mixtures, the scaling factor for water is taken equal to the ScC of the pharmaceutical.

3 Computational details

In this paper, paracetamol, benzoic acid, 4-aminobenzoic acid, salicylic acid, ibuprofen, naproxen and temazepam were chosen as model drugs. The chemical structure of these molecules is shown in Figure 1. The choice was based on the chemical diversity they offer because of the various functional groups they contain and the numerous solubility data available in the literature for these molecules.

The melting temperature and enthalpy of these molecules are listed in Table 2. Some of these data were measured previously [17] by some of the authors of this work. Uncertainties are also provided when available, in parenthesis. The same solubility data were used for both SciPharma and NRTL-SAC. More specifically, the solubility in water, methanol, ethanol, ethyl acetate, acetone and cyclohexane at 298.15 K were used. These data are given in Table 3. The selected solvents cover the spectrum of hydrophilic, polar and hydrophobic solvents needed for adequate parameterization of the solutes for both models.

The calculations with SciPharma were conducted within the MAPS platform of Scienomics. The parameters of the pharmaceuticals calculated with SciPharma are given in Table 4. In the same table are listed the parameters of some of the solvents used in the solubility calculations for which the parameters were regressed in this work by using vapor pressure data and saturated liquid densities. The temperature range of the regression and the resulting average absolute deviation (%) are also given in Table 4. The parameters for ethyl acetate, acetone, 1,4-dioxane, n-butyl acetate,

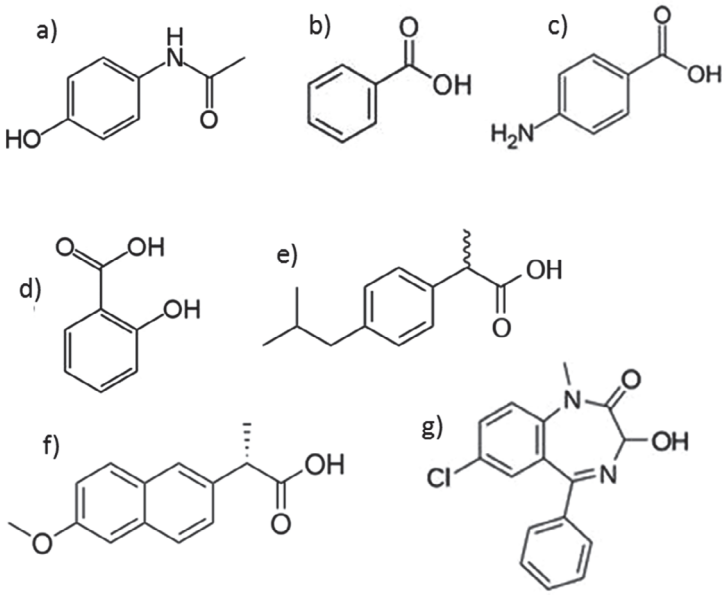

Fig. 1. Chemical structure of the pharmaceuticals examined in this work: (a) paracetamol, (b) benzoic acid, (c) p-aminobenzoic acid, (d) salicylic acid, (e) ibuprofen, (f) naproxen and (g) temazepam.

Table 2. Melting temperature and enthalpy for the pharmaceutical compounds examined in this work. Uncertainties, when available, are given in parenthesis.

Compounds	T_{m}**(K)**	ΔH_{m}**(J/mol)**
Paracetamol [17]	442.1 (0.5)	27471 (1720)
Benzoic acid [17]	395.2 (0.3)	17350 (775)
p-Aminobenzoic acid [42]	461.7	24030
Salicylic acid [17]	431.4 (0.1)	24626 (2652)
Ibuprofen [17]	347.9 (0.7)	25204 (1577)
Naproxen [42]	427.6	31500
Temazepam [40]	432.5	25581

methyl ethyl ketone, methyl isobutyl ketone and propylene glycol were taken from previous work [18]. For the rest of the solvents used in this work the parameters were taken from [20] and [25].

The quadruplets $[XY^{-}Y^{+}Z]$ of NRTL-SAC were regressed with the help of the lsqnonlin function of Matlab and are listed in Table 5 along with the regression error. The error is calculated with an equation similar to (6) except that mole fractions were used instead of their logarithms.

4 Results and discussion

4.1 Pure solvents

The solubility of the pharmaceuticals in various solvents at different temperatures was calculated with both SciPharma and NRTL-SAC. The results are compared against

Table 3. The solubility data at 298.15 K used for the regression of pharmaceutical parameters with both models. Data are given in mole fraction units.

Regression solvents	Paracetamol	Benzoic acid	p-Aminobenzoic acid	Salicylic acid	Ibuprofen	Naproxen	Temazepam
Water	1.77E-03	5.00E-04	7.75E-04	2.88E-04	1.36E-05	5.13E-06	6.10E-06
Methanol	6.58E-02	0.1632	4.73E-02	1.28E-01	*2.14E-01	1.46E-02	5.47E-03
Ethanol	6.01E-02	0.1789	4.47E-02	1.39E-01	1.98E-01	2.01E-02	3.00E-03
Ethyl acetate	5.48E-03	0.165	5.32E-02	1.36E-01	2.25E-01	3.55E-02	1.45E-02
Acetone	3.69E-02	0.1925	4.84E-02	1.79E-01	2.33E-01	6.92E-02	2.07E-02
Cyclohexane	2.00E-05	0.012	1.20E-05	4.30E-04	1.12E-01	1.20E-04	1.58E-03

* Solubility at 303.15 K.

Table 4. PC-SAFT parameters for the pharmaceutical compounds and solvents regressed in this work. The temperature range and the average absolute deviation (%) of the regression are also listed for the solvents.

Compounds	m	σ(Å)	ε/k (K)	Nassoc	ε_{hb}/k (K)	k_{hb}	T(K)	ΔP^{sat} (%)	$\Delta\rho^{liq}$ (%)
Paracetamol	1.5210	4.3806	388.53	4	1783.97	0.01			
Benzoic acid	2.3221	3.5626	247.82	4	946.93	0.01			
p-Aminobenzoic acid	3.5807	4.0700	400.65	4	1930.94	0.01			
Salicylic acid	1.5364	4.7402	349.88	4	1349.85	0.01			
Ibuprofen	3.6175	3.4231	256.78	4	1148.89	0.01			
Naproxen	3.8212	3.8088	354.09	4	1408.52	0.01			
Temazepam	4.7180	3.6069	356.57	4	1265.82	0.01			
Solvents									
Acetic acid	1.5026	3.7014	286.12	1	5248.62	0.0067	293–563	1.21	0.084
Acetic anhydride	4.2175	3.1329	250.93	–	–	–	220–562	0.37	1.75
Acetophenone	3.8408	3.4574	219.26	2	1952.05	0.9632	295–680	1.33	0.7
Anisole	3.3319	3.5611	290.17	2	498.46	0.0252	250–610	0.92	0.43
Benzyl alcohol	5.7211	2.9154	226.99	2	983.98	1.2718	265–555	0.81	1.48
2-Butanol	2.9580	3.5012	240.75	2	2123.63	0.0141	170–500	1.57	1.39
Carbon tetrachloride	2.3182	3.8116	292.55	–	–	–	200–490	0.16	0.42
Chloroform	2.3767	3.5406	278.63	2	811.02	0.0033	220–505	0.79	0.55
Dichloromethane	1.9808	3.4966	292.93	2	1032.41	0.0034	200–485	0.64	1.55
Diethyl ether	2.8662	3.5531	223.75	2	1097.99	9.8E-04	170–440	0.06	0.34
Dimethyl sulfoxide	3.6364	2.9678	209.18	2	2259.18	1.0484	293–630	1.26	0.59
Dimethylformamide	3.5169	3.1019	211.69	2	1564.27	1.4000	240–580	0.75	0.86
Ethylene glycol	1.9088	3.5914	325.23	4	2080.03	0.0235	270–610	0.75	2.1
Methyl acetate	3.1994	3.1723	233.35	2	1056.84	1.7E-04	200–480	0.38	0.8
2-Methyl-1-propanol	2.4236	3.7659	264.16	2	2811.02	0.0033	200–500	1.13	0.35
N-Methyl-2-pyrrolidone	3.6473	3.3241	242.18	2	1658.75	1.0979	260–675	0.37	0.28
Tetrahydrofuran	3.1437	3.2217	196.28	2	733.27	1.0979	180–520	0.69	0.18

Table 5. The NRTL-SAC segments for the pharmaceutical compounds. The error of the regression is listed in the last column (*$RMSE$ calculated from mole fractions instead of their logarithms).

Compounds	X	$Y-$	$Y+$	Z	$RMSE(*)$
Paracetamol	0.369	0.88	0.392	0.659	0.066
Benzoic acid	0.53	0	0.396	0.493	0.003
p-Aminobenzoic acid	0.122	1.115	2.235	0.388	0.108
Salicylic acid	0.844	2.216	0.664	0	0.003
Ibuprofen	0.889	0.55	0.38	0.063	0.025
Naproxen	0.653	0.046	1.25	0.653	0.008
Temazepam	0.415	0	1.58	0	0.565

Table 6. The $RMSE$ of the two models for the solubility of the compounds in pure solvents. A total of 386 data points were considered comprising pure solvents at one or more temperatures. The data used for parameterization were excluded.

Compounds	Data points	$RMSE$ (SciPharma)	$RMSE$ (NRTL-SAC)
Paracetamol	69	0.181	0.219
Benzoic acid	87	0.160	0.136
p-Aminobenzoic acid	30	0.260	0.319
Salicylic acid	63	0.158	0.306
Ibuprofen	95	0.158	0.129
Naproxen	29	0.362	0.452
Temazepam	13	0.470	0.498
Average		0.196	0.227

experimental data for paracetamol [26–28], benzoic acid [29–31], p-aminobenzoic acid [32], salicylic acid [32–35], ibuprofen [36,37], naproxen [38,39], and temazepam [40]. A compilation of these data can be found in [41]. Comparison between calculations and experimental data for all molecules in pure solvents is shown in Figure 2 for SciPharma (top) and NRTL-SAC (bottom). The average root-mean-squared error ($RMSE$) is calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log X_i^{\exp} - \log X_i^{calc}\right)^2} \tag{6}$$

where X_i is the solubility of the compound i in mole fraction and N is the number of data points. The $RMSE$ of the models is given in Table 6. The data points used for the parameterization of the compounds are excluded from Figure 2 and from the calculation of $RMSE$ in Table 6. The data points are pure solvents at one or more temperatures for which there were available data. All solvents for which there was at least one solubility available at some temperature are included in Figure 2. The solubilities span many orders of magnitude. Almost all results are within one order of magnitude of the experimental values, as depicted by the parallel lines that set this boundary. The data are, also, given in tabular form as Supplementary material.

In Figure 2 some outliers stick out from the rest of the data points. For SciPharma, the points lying on the boundary line and above are paracetamol [27] and salicylic acid [32] in chloroform (the experimental values are low compared to the prediction). Also, close to the boundary line with error close to 0.7 or above are the data for temazepam [40] in dichloromethane and naproxen in 1,4-dioxane [39] and dimethylformamide [38] ($RMSE = 0.66$). For these data points the predictions of SciPharma are low compared

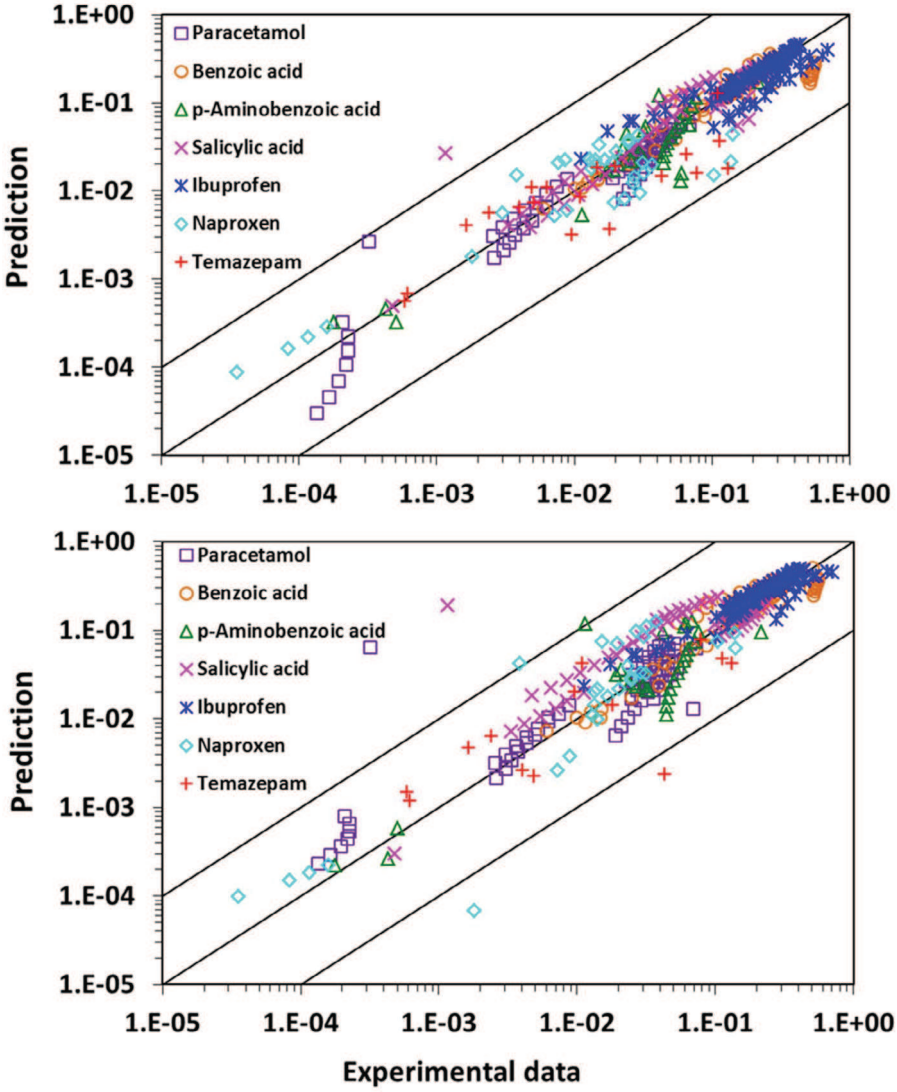

Fig. 2. Predicted versus experimental solubility (mole fraction) of the pharmaceutical compounds in various pure solvents at one or more temperatures. Results with SciPharma (top) and NRTL-SAC (bottom).

to experiment. Solvents like dichloromethane, 1,4-dioxane, or dimethylformamide are polar solvents that depending on the polar groups of the pharmaceutical may result in high solubility. To account implicitly for polar interactions and enhance the interactions of polar solvents with the pharmaceuticals which, usually, contain several polar groups, the solvents were modelled with association sites as can be seen in Table 4. Nevertheless, the results are not always satisfactory.

For NRTL-SAC, the points lying outside the boundary of one order of magnitude are paracetamol [27], p-aminobenzoic acid [32] and salicylic acid [32] in chloroform (low experimental values compared to the prediction), naproxen [38] in heptane (much lower solubility is predicted) and in ethylene glycol, and temazepam in anisole [40].

As can be seen from the average *RSME* in Table 6, calculated over 386 data points, SciPharma is about 13% more accurate than NRTL-SAC. For some solvents for which there were no available parameters with NRTL-SAC, namely 1-hexanol, 1-heptanol, 2-methyl-1-propanol, methyl acetate, propylene glycol, diethyl ether, benzyl alcohol, acetophenone, acetic anhydride, results are reported only for SciPharma along with the experimental values. These data points were excluded from the error calculation. From the error calculation were also excluded the data used for the parameterization and the solubilities of paracetamol and salicylic acid in chloroform for which the deviation of NRTL-SAC exceeds 4 orders of magnitude. All data are given in the Supplementary material.

One reason for the better accuracy achieved by SciPharma is that temperature-dependent results of solubility are included and this is described better by PC-SAFT since the activity coefficient is temperature-dependent while in NRTL-SAC the temperature dependence is not built in. The temperature dependence is depicted in Figure 3 where solubility data for selected compounds are plotted for some solvents in a range of temperatures. In each plot the same color/symbol points correspond to the same solvent at different temperatures. The correct qualitative description of solubility as a function of temperature would result in points lying on a line parallel to the diagonal line. The deviation of each model from the correct description of temperature dependence is calculated as the difference of the slope of the calculated versus experimental solubility curve from 1 (slope of the diagonal):

$$\text{Dev} = 1 - \frac{\boldsymbol{dX}^{pred}}{\boldsymbol{dX}^{exp}}. \tag{7}$$

The slopes were calculated with linear regression. The average deviation for each pharmaceutical is given in Table 7 for both models. The solvents over which the deviation was calculated are also listed. On the average, the deviation for SciPharma is 30% smaller compared to NRTL-SAC. The data can be found in the Supplementary material.

4.2 Mixed solvents

The predictive ability of the models was, subsequently, tested on solubility calculations in mixed solvents. As the solubility in pure solvents spans many orders of magnitude (Fig. 2) one expects that for selected mixed solvents the solubility will vary substantially as a function of composition. In the pharmaceutical industry a mixed solvent is commonly used to tune solubility. From this point of view a thermodynamic model capable to predict the variation of solubility is highly desirable.

Standard mixing rules were used for PC-SAFT [24]. Also, for aqueous mixtures, a k_{ij} parameter between the pharmaceutical and water was used to bring down to its actual value the elevated solubility in water due to the parameterization scheme. The scaling factors calculated for pure solvents are also used for mixtures. These factors can be thought as effective binary interaction parameters k_{ij}. As we discussed above, the calculated solubility with PC-SAFT at any composition is scaled with the molar average of the scaling factors of the two solvents. For water the scaling constant (ScC) characteristic of each pharmaceutical is used. Finally, no binary parameter was used for the interactions between the solvents.

Some typical examples are shown in Figure 4 to Figure 6. Figure 4 shows comparison between experimental solubility data of paracetamol [43] and naproxen [44] in ethanol/water with both models. To correct the solubility in water with SciPharma, binary interaction parameters k_{ij} are calculated from the actual solubility of paracetamol and naproxen in water as 0.04803 and 0.071, respectively. Ethanol is one

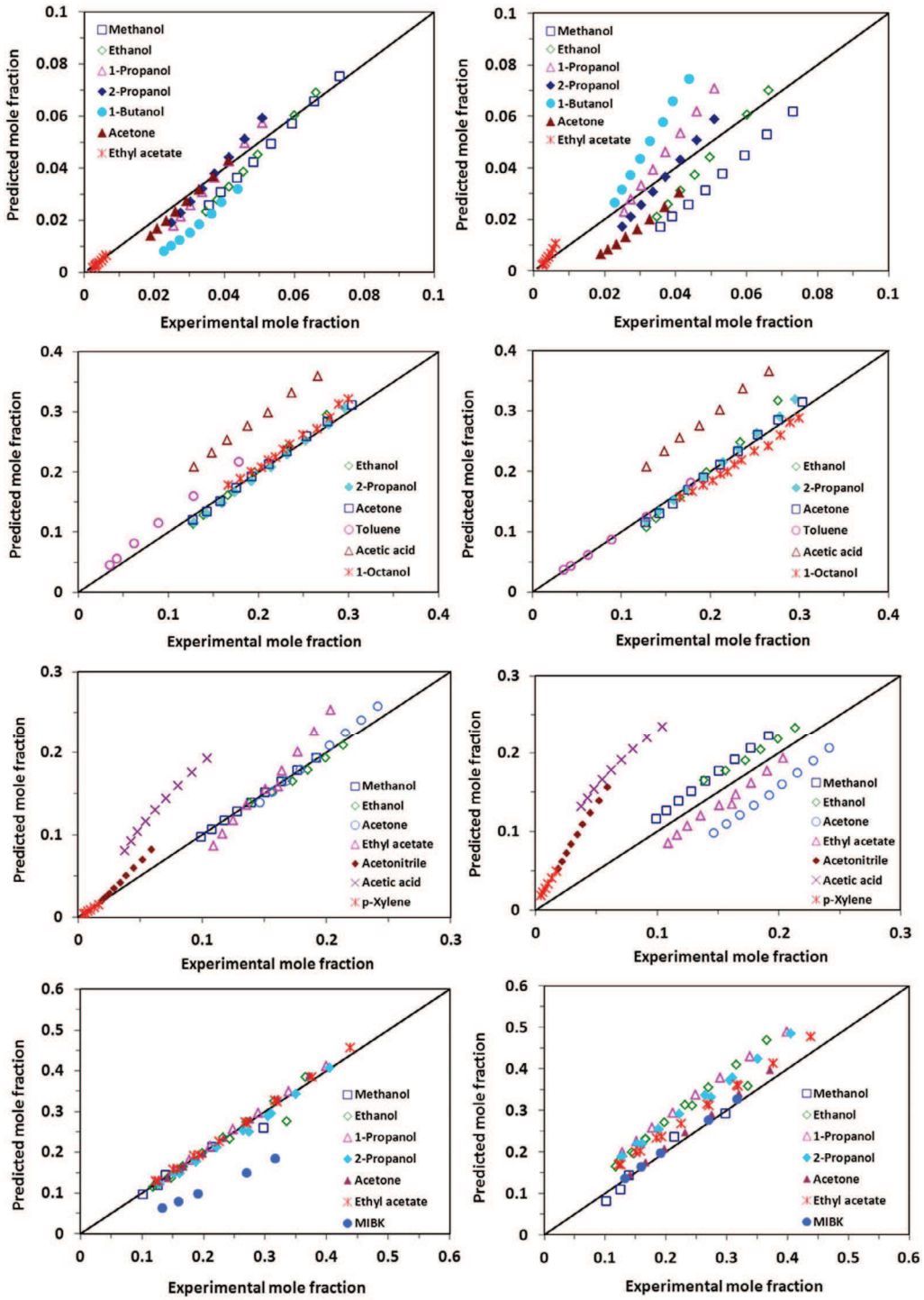

Fig. 3. Predicted versus experimental solubility (mole fraction) for selected drugs in pure solvents at a range of temperatures to illustrate the temperature dependence of solubility. From top to bottom are shown: paracetamol, benzoic acid, salicylic acid, ibuprofen with SciPharma (left) and NRTL-SAC (right).

Table 7. The deviation of the two models from the correct description of temperature dependence as calculated from (7). The solvents considered for the deviation estimation are listed.

Compounds	Solvents	Dev (SciPharma)	Dev (NRTL-SAC)
Paracetamol	**MeOH, EtOH,1-PrOH, 2-PrOH, 1-BuOH, Acetone, EtOAc**	0.38	0.70
Benzoic acid	**EtOH, 2-PrOH, 1-C8OH, Acetone, Acetic acid, Toluene**	0.13	0.15
Salicylic acid	**MeOH, EtOH, Acetone, Acetic acid, MeCN, EtOAc, Xylene**	0.32	0.57
Ibuprofen	**MeOH, EtOH, 1-PrOH, 2-PrOH, Acetone, EtOAc, MIBK**	0.11	0.07
Naproxen	**Cyclohexane, Chloroform, 1-C8OH**	0.46	0.56
Average		0.28	0.41

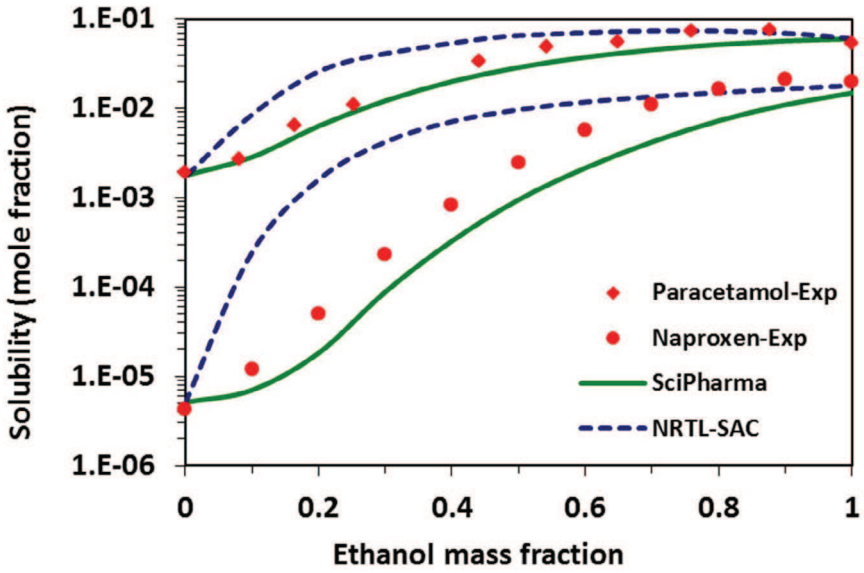

Fig. 4. Experimental and predicted solubility of paracetamol and naproxen in the ethanol water binary mixture at 298.15 K.

of the solvents used for the parameterization, thus, the solubility for this binary is known at the ends of the phase diagram. The solubility of naproxen spans three orders of magnitude through the entire composition range. Both models seem to predict qualitatively the solubility variation with composition. NRTL-SAC overestimates the solubility, especially for naproxen, up to 0.5 ethanol mass fraction.

Figure 5 depicts the solubility of paracetamol in 2-propanol/water. This mixture was selected because the solubility in 2-propanol is not known by the models. SciPharma treats 2-propanol with the same scaling factors for ethanol, since they are in

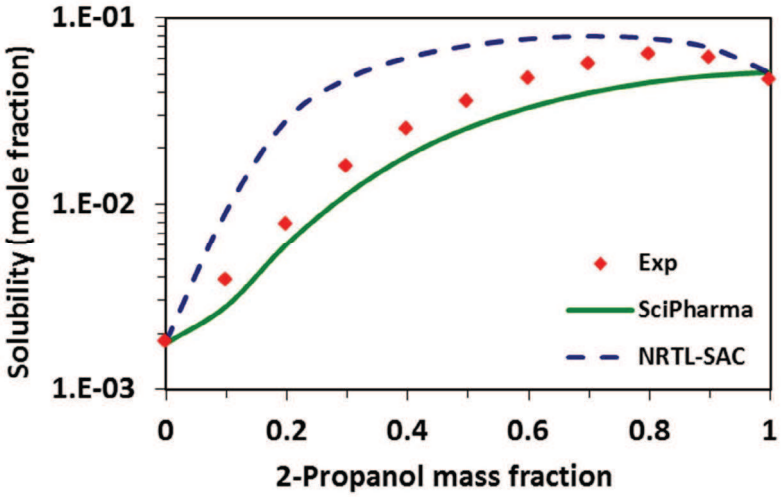

Fig. 5. Experimental and predicted solubility of paracetamol in the 2-propanol/water mixture at 298.15 K.

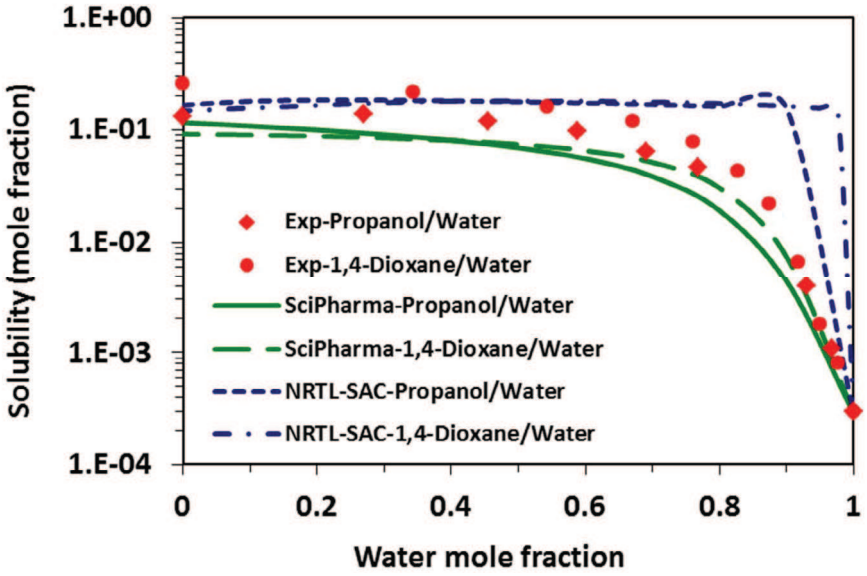

Fig. 6. Experimental and predicted solubility of salicylic acid in the propanol/water and 1,4-dioxane/water mixtures at 298.15 K.

the same family of solvents. The predictions are in good agreement with the experimental data [45]. NRTL-SAC tends to overestimate the solubility especially from low to intermediate alcohol mass fractions. However, it manages to predict a maximum of solubility at high propanol concentration (although the maximum is extended to lower concentrations).

Finally, in Figure 6 are shown the experimental [46,47] and calculated solubilities of salicylic acid in aqueous mixtures of propanol and 1,4-dioxane. Solubility in pure propanol and 1,4-dioxane solvents is not known to the models. For SciPharma, similarly to 2-propanol, propanol has a scaling ratio inherited from the known

solubility in ethanol. No scaling factor is used for 1,4-dioxane. The solubility of salicylic acid in water is corrected with a k_{ij} parameter equal to 0.038. SciPharma shows good agreement with the experimental data. It, even manages to predict the slightly higher solubility in the 1,4-dioxane aqueous solution. NRTL-SAC gives a flat curve showing no variation of solubility with composition up to about 0.9 water mole fraction with propanol or higher in the case of 1,4-dioxane.

The results in the mixtures show that SciPharma can be used with confidence to calculate the solubility in mixtures with no other adjustable parameter than the scaling factors derived from the known solubilities in specific solvents. By extending these factors to other solvents of the same family (propanol, 2-propanol) seems sufficient for quantitative results. Also, the appropriate parameterization of the pharmaceuticals with the proposed scheme is a key aspect that permits reliable predictions even without scaling factors (1,4-dioxane).

5 Conclusions

In this work, SciPharma, a PC-SAFT-based approach, and NRTL-SAC model were used to study the solubility of seven pharmaceutical compounds in pure and mixed solvents as a function of temperature. For the parameterization of the compounds, the same solubility data were used for both models. The six solvents involved were chosen to cover the whole range of hydrophilic, polar and hydrophobic molecules needed for adequate parameterization with both models. These six solubility data were the only data used by both models for predicting the pharmaceutical solubility in other pure and mixed solvents.

The two models were compared against available experimental data. The solubility in all pure solvents that could be found in literature at a single temperature and in some solvents in a range of temperatures was captured with an overall better accuracy of 13% by SciPharma in terms of RMSE. Results from solvents only as a function of temperature show that SciPharma is more accurate in the description of the temperature dependence of solubility. The reason for this is that the temperature-dependence of the activity coefficient is built-in for PC-SAFT.

The solubility in mixed solvents is better predicted with SciPharma. NRTL-SAC seems to overestimate the solubility in aqueous mixtures of alcohols. In some cases, NRTL-SAC showed almost invariable solubility with composition. SciPharma manages to predict quantitatively the solubility in mixtures with no other adjustable parameter than scaling factors derived from the solubility in pure solvents used for the parameterization.

References

1. H. Renon, J.M. Prausnitz, AIChE J. **14**, 135 (1968)
2. T.C. Frank, J.R. Downey, S.K. Gupta, Chem. Eng. Prog. **95**, 41 (1999)
3. C.-C. Chen, Y. Song, Ind. Eng. Chem. Res. **43**, 8354 (2004)
4. F. Ruether, G. Sadowski, J. Pharm. Sci. **98**, 4205 (2009)
5. I. Tsivintzelis, I.G. Economou, G.M. Kontogeorgis, AIChE J. **55**, 756 (2009)
6. A. Klamt, J. Phys. Chem. **99**, 2224 (1995)
7. A. Klamt, F. Eckert, M. Hornig, E.M. Beck, T. Burger, J. Comput. Chem. **23**, 275 (2002)
8. S. Gracin, T. Brinck, A. Rasmuson, Ind. Eng. Chem. Res. **41**, 5114 (2002)
9. I. Hahnenkamp, G. Graubner, J. Gmehling, Int. J. Pharm. **388**, 73 (2010)
10. U. Weidlich, J. Gmehling, Ind. Eng. Chem. Res. **26**, 1372 (1987)
11. B. Bouillot, S. Teychené, B. Biscans, Ind. Eng. Chem. Res. **52**, 9276 (2013)

12. B. Bouillot, S. Teychené, B. Biscans, Ind. Eng. Chem. Res. **52**, 9285 (2013)
13. E. Mullins, Y.A. Liu, A. Ghaderi, S. Fast, Ind. Eng. Chem. Res. **47**, 1707 (2008)
14. H. Tung, J. Tabora, N. Variankaval, D. Bakken, C. Chen, J. Pharm. Sci. **97**, 1813 (2008)
15. F.L. Mota, A.P. Carneiro, S.P. Pinho, E.A. Macedo, Eur. J. Pharm. Sci. **37**, 499 (2009)
16. E. Sheikholeslamzadeh, S. Rohani, Ind. Eng. Chem. Res. **51**, 464 (2012)
17. B. Bouillot, S. Teychené, B. Biscans, Fluid Phase Equilib. **309**, 36 (2011)
18. T. Spyriouni, X. Krokidis, I.G. Economou, Fluid Phase Equilib. **302**, 331 (2011)
19. J.M. Prausnitz, R.N. Lichtenthaler, E. Gomes de Azevedo, *Molecular Thermodynamics of Fluid-Phase Equilibria* (Prentice Hall, 1999)
20. J. Gross, G. Sadowski, Ind. Eng. Chem. Res. **40**, 1244 (2001)
21. C.-C. Chen, Fluid Phase Equilib. **83**, 301 (1993)
22. A. Fredenslund, R.L. Jones, J.M. Prausnitz, AIChE J. **21**, 1086 (1975)
23. C.-C. Chen, P.A. Crafts, Ind. Eng. Chem. Res. **45**, 4816 (2006)
24. F. Tumakaka, J. Gross, G. Sadowski, Fluid Phase Equilib. **228**, 89 (2005)
25. J. Gross, G. Sadowski, Ind. Eng. Chem. Res. **41**, 5510 (2002)
26. R.A. Granberg, A.C. Rasmuson, J. Chem. Eng. Data **44**, 1391 (1999)
27. J. Barra, F. Lescure, E. Doelker, P. Bustamante, J. Pharm. Pharmacol. **49**, 644 (1997)
28. G.L. Perlovich, T.V. Volkova, A. Bauer-Brandl, J. Pharm. Sci. **95**, 2158 (2006)
29. A. Beerbower, P.L. Wu, A. Martin, J. Pharm. Sci. **73**, 179 (1984)
30. F.A. Restaino, A.N. Martin, J. Pharm. Sci. **53**, 636 (1964)
31. B. Long, J. Li, R. Zhang, L. Wan, Fluid Phase Equilib. **297**, 113 (2010)
32. J. Barra, M.A. Pena, P. Bustamante, Eur. J. Pharm. Sci. **10**, 153 (2000)
33. A. Shalmashi, A. Eliassi, J. Chem. Eng. Data **53**, 199 (2008)
34. F.L. Nordstrom, A.C. Rasmuson, J. Chem. Eng. Data **51**, 1668 (2006)
35. H. Matsuda, K. Kaburagi, S. Matsumoto, K. Kurihara, K. Tochigi, K. Tomono, J. Chem. Eng. Data **54**, 480 (2009)
36. S. Gracin, A.C. Rasmuson, J. Chem. Eng. Data **47**, 1379 (2002)
37. D.M. Stovall, C. Givens, S. Keown, K.R. Hoover, E. Rodriguez, W.E.Jr. Acree, M.H. Abraham, Phys. Chem. Liq. **43**, 261 (2005)
38. P. Bustamante, M.A. Pena, J. Barra, J. Pharm. Pharmacol. **50**, 975 (1998)
39. C.R. Daniels, A.K. Charlton, R.M. Wold, E. Pustejovsky, A.N. Furman, A.C. Bilbrey, J.N. Love, J.A. Garza, W.E. Jr. Acree, M.H. Abraham, Phys. Chem. Liq. **42**, 481 (2004)
40. P.J. Richardson, D.F. McCafferty, A.D. Woolfson, Int. J. Pharm. **78**, 189 (1992)
41. A. Jouyban, *Handbook of Solubility Data for Pharmaceuticals* (CRC Press, Boca Raton, 2009)
42. J. Marrero, J. Abildskov, *Solubility and Related Properties of Large Complex Chemicals, Part 2* (DECHEMA Frankfurt, 2005)
43. P. Bustamante, S. Romero, A. Reillo, Pharm. Pharmacol. Commun. **1**, 505 (1995)
44. D.P. Pacheco, F. Martinez, Phys. Chem. Liq. **45**, 581 (2007)
45. H. Hojjati, S. Rohani, Org. Process Res. Dev. **10**, 1101 (2006)
46. M.A.A. Fakhree, S. Ahmadian, V. Panahi-Azar, W.E.Jr. Acree, A. Jouyban, J. Chem. Eng. Data **57**, 3303 (2012)
47. M.A. Pena, P. Bustamante, B. Escalera, A. Reillo, J. Pharm. Biomed. Anal. **36**, 571 (2004)