Evaluation of two models for solubility of active pharmaceutical ingredients: Application on solvent selection in process design

Sara Antunes Febra
sara.febra@ist.utl.pt

Instituto Superior Técnico, University of Lisbon, Portugal

November 2013

Abstract

Having an optimal solvent selection, and ultimately designing the optimal process, can only be achieved with a suitable solubility model. The empirical NRTL-SAC model\[1, 2\] has proven to be superior to other solubility models and Pharma mod. UNIFAC(PMUNIFAC)\[3\], specially designed for application in API organic solutions, has arisen as a powerful candidate for the application in solvent selection. Unlike NRTL-SAC, PMUNIFAC has the advantage of being a predictive model, not requiring experimental data as user input, but only molecular structure, melting temperature and melting enthalpy. The implementations of NRTL-SAC and PMUNIFAC were accomplished with success in the softwares gPROMS ModelBuilder and Excel/Visual Basic for Applications (VBA). The two models were evaluated against experimental data from 8 APIs from GlaxoSmithKline (GSK). Only NRTL-SAC proved to be t for purpose (average order of magnitude of discrepancy (OMD) = 0.27 and maximum OMD = 1.63). However, when only systems with all parameters available were considered, PMUNIFAC showed good predicting potential with an average OMD = 0.40 and a maximum OMD = 2.05. It was conclusive that the bottleneck for PMUNIFAC to be applied in solubility modelling was having too many missing parameters, but its results were encouraging to support its further development. In the last stage of the study, NRTL-SAC was successfully applied in the solvent selection of two GSK crystallisation cases. Alternative solvent systems were identified and validated, adding value to the solvent selection procedures of the company.

Keywords: Solubility; Molecular modelling; NRTL-SAC; Pharma. Mod. UNIFAC; Solvent selection; Solid-liquid equilibria.

1. Introduction

There are many new medicines that are being developed today. Many of these potential new medicines will fail in clinical trials, but some may represent tomorrow’s new treatments. In total, it takes approximately 10 to 15 years to go through the drug discovery and clinical development process and bring a medicine to the market. Ultimately, the manufacturing process should give a product matching all quality standards while ensuring its economical efficiency and sustainability that starts from the optimization of yield, process volumes, mass efficiency, vessel occupancy, carbon footprint, among others. The whole procedure is also costly, which are all reasons that support the big effort by the pharmaceutical industry on the improvement of the process design strategies.

In the pharmaceutical industry, the presence of solvents is essential in all steps of manufacturing process (Figure 1). Therefore, the solubility prediction methodologies have drawn the attention of the scientific community[4], since the key parameters in the manufacture of active pharmaceutical ingredients (APIs), being related with solvent selection, are the solute solubility and the solvation promotion (form change)[5].

Figure 1: Typical process sequence in Pharma industry.

Crystallisation is the preferred method of purification in the pharmaceutical industry for both the final drug substance and the isolated intermediates in the synthesis[2], hence, the selection of a suitable solvent system is especially relevant for this step, where the quality of the isolated product is determined and the yield of the process can be the most compromised.
In practice, the selection of solvents and anti-solvents for crystallisation mostly relies on experience, analogy and experimental testing. It is time and resources consuming to test many pure solvent/mixtures systems, let alone all combinations. Hence, representatives are selected (solvent classes), leaving out many potentially better options. Moreover, it is not always possible to generate experimental data due to material constraints, as they tend to occur in the early development phases or in case of impurities. To surpass these hindrances, one interesting perspective is the use of thermodynamic models for the calculation.

Many distinct thermodynamic models have been developed throughout the years and some have predictive potential with high accuracy such as Wilson, UNIQUAC, NRTL(-RK) and modified UNIFAC (Dortmund) for vapour-liquid equilibria (VLE). On the other hand, the choice of the best thermodynamic model for solid-liquid equilibria (SLE) is not an easy task, as most of the models are not designed for SLE, but for other applications. Nevertheless, there are currently two recent models, Pharma Mod. UNIFAC (PMUNIFAC) and Non-Random Two-Liquid Segment Activity (NRTL-SAC), that were built exactly for API solubility prediction and show promising results for activity coefficients in SLE, however PMUNIFAC had not been implemented and evaluated within a pharmaceutical company yet.

2. Solubility Theory

According to literature, the general relation for SLE calculation of simple eutectic systems is described in equation 1.

\[ \ln x^L_{I^L} = -\frac{\Delta h_m}{RT} \left(1 - \frac{T}{T_{m,I}}\right) - \frac{\Delta c_p}{R} \times \left(1 - \frac{T}{T_{m,I}}\right) + \frac{\Delta c_p}{R} \ln \frac{T}{T_{m,I}} \]  

(1)

For temperatures that are not very far from the melting point, the contributions of the heat capacity are negligible, since they have tendency to cancel out due to the opposite signs. After this simplification a simple relation is obtained, equation 2.

\[ \ln x^L_{I^L} = -\frac{\Delta h_m}{RT} \left(1 - \frac{T}{T_{m,I}}\right) \]

\[ = \frac{\Delta s_m}{R} \left(1 - \frac{T}{T_{m,I}}\right) \]  

(2)

The solubility of a solid compound in the liquid phase can then be calculated for a given temperature if, apart from the enthalpy (or entropy) of fusion and the melting temperature information, the solute activity coefficient is available. By the use of an activity coefficient model, it is possible to determine the solute solubility in an iterative way.

2.1. NRTL-SAC model

NRTL-SAC is a segment contribution activity coefficient model, derived from the polymer non-random two-liquid model (NRTL). It describes the effective surface interactions for any organic, non-electrolyte and non-solvate molecule in terms of four types of conceptual segments. The hydrophobic segment X represents the molecular surface area that is adverse to hydrogen bonding. The hydrophilic segment Z simulates polar molecular surfaces with the tendency to form a hydrogen bond. The polar segments represent the molecular surface area with interactions characteristic of an electron donor or acceptor. The polar-attractive segment Y− exhibits attractive interactions with a hydrophilic molecular surface, whereas the polar-repulsive segment Y+ exhibits repulsive interactions with a hydrophilic molecular surface. These segments have unary parameters \(\tau_k\), with \(k = \text{segment letter or just } X, Y^-, Y^+\) and Z) and binary parameters \(\alpha_{nm}\) and \(\tau_{nm}\), with \(n\) and \(m\) as the segment indices) associated.

The generation of the parameters \(X, Y^-, Y^+\) and Z for solvent parameters is done by the model developers through regressions that involve available experimental VLE or liquid-liquid equilibria (LLE) data for solvent-solvent binary pairs. There is total of five different possible natures for solvents that are related with the unary parameters weights: hydrophobic, hydrophilic, polar, hydrophobic/polar and hydrophilic/hydrophobic.

The solute parameters \(X, Y^-, Y^+\) and Z can be retrieved by a model user in a capable tool with NRTL-SAC implemented, such as Aspen Solubility Modeler, by feeding a representative solubility dataset of solute-solvent(s) systems. It is also important to cover all solvents natures (up to five) that are present in the target systems to be predicted, in the initial experimental dataset.

The correlative character of the model enables a different approach for the general relation for SLE (equation 1 or its simplification, equation 2). Since, as suggested in literature, there is the possibility of creating up to three adjustable parameters A, B and C, according to equations 3 and 4.

\[ \text{equation (2)} = A + \frac{B}{T} \]  

\[ A = \frac{\Delta h_m}{RT_{m,I}} \]  

\[ B = -\frac{\Delta h_m}{R} \]  

\[ \text{equation (1)} = A + \frac{B}{T} + C \ln(T) \]  

(3)

(4)
\[
A = \frac{\Delta h_{m,I}}{R T_{m,I}} - \frac{\Delta c_{P,I}}{R} - \frac{\Delta c_{P,I} \ln(T_{m,I})}{R},
\]
\[
B = -\frac{\Delta h_{m,I}}{R} + \frac{\Delta c_{P,I} T_{m,I}}{R},
\]
\[
C = \frac{\Delta c_{P,I}}{R}.
\]

Instead of using the melting temperature and enthalpy, it is possible to regress these three additional API-specific parameters or only A and B, as the heat capacity contribution is typically negligible. This approach is specially useful for cases when there is no knowledge on the fusion temperature, the fusion enthalpy or the liquid and solid heat capacity (when significant). Plus, it enables a better fit for the experimental data.

Regarding the binary interactions, the developers proposed “reference compounds” for the conceptual segments - Hexane for X, Water for Z and Acetone for Y. The interaction between the polar molecule and water can be negative or positive, origin of the Y” and Y+”. The reference compounds are used to identify the segment-segment non-randomness factor (\(\alpha_{nm}\)) and binary interaction energy (\(\tau_{nm}\)) parameters for the conceptual segments from regression of available experimental VLE and LLE data associated with these reference compounds. This way, the binary interaction matrices \(\tau_{nm}\) and \(\alpha_{nm}\), have each sixteen predefined entries. Despite the activity coefficients being highly dependent on the temperature, NRTL-SAC attempts to provide order-of-magnitude estimates based on minimal experimental information. Thus no temperature dependency is provided in the model, apart from the contribution in the equilibrium constant parameters from equations 3 and 4.

### 2.2. Pharma. mod. UNIFAC model

Pharma. mod. UNIFAC is a group contribution method, hence, it looks at a solution as a mixture of functional groups (FGs) with predetermined parameters rather than molecules. The advantage of this concept is that the number of FGs is much smaller than the number of possible molecules. The model is, in fact, a derivative of modified UNIFAC model, created with the intention of overcoming the functional groups diversity limitation that the mod. UNIFAC faced when applied to API solutions. Therefore, the system of equations of PMUNIFAC is the same as the one of mod. UNIFAC and the difference between the models resides solely in the set of available FGs and its unary (\(R_b\) and \(Q_b\)) and binary parameters values \(\Psi_{nm}\), where \(k\) and \(i\) are the FG and component indices, respectively.

Incrementation is the process of decomposing a molecule in functional groups respective to the group contribution method in use. Each group structure is associated to a main group (MG), and, to illustrate, an example of an incrementation of Heptane is present in Figure 2. The respective functional groups count \(\nu^k_i\) are in Table 1, where \(k\) and \(i\) are the FG and component indices, respectively.

There are two functional group specific parameters, the surface area \(R\) and the volume \(Q\), that are obtained from the van der Waals surface area and volumes. As for the binary parameters, \(\Psi_{nm}\), as clear from equation 5, they have three contributions, \(a_{nm}\), \(b_{nm}\) and \(c_{nm}\), defined between two main groups, \(m\) and \(n\), rather than functional groups.

\[
\Psi_{nm} = \exp\left(-\frac{a_{nm} + b_{nm} T + c_{nm} T^2}{T}\right)
\]

The author of the model started by identifying 5 solvent classes (plus Water) that cover 75% of all solvents used in pharmaceutical industry: alkane, alcohol, ketone, ester, ether and Water. Each one of these classes is associated to a main group, being main group 1 characteristic of alkane class, 5 of alcohol, 13 of ketone, 19 of ester, 34 of ether and 73 of Water. For this first approach of the model, the author mainly considered the binary interactions between each main group present in APIs and solvents MGs 1, 5, 13, 19, 34 and 73. There were only a few binary interactions considered with both the groups in large concentration. This shows that the model is not ready yet to applied in solvent mixtures. For instance, a ternary system of an API with a solvent mixture of alcohol/ketone would not be possible to work with because it would imply parameters for the interaction 5-13 with both groups in large quantity, and, currently, this interaction is only described for API/alcohol or API/ketone interactions (parameters fitted over limited concentration range).

An analysis was done for GSK solvents, in order to assess at which extent they were described by the model. The results show that considering only the groups 1, 5, 13, 19, 34 and 73, to represent the solvents that are used in API manufacture, is a fair approach to start with, since only 34 out of 162 solvents in the GSK database and 3 out of 18 in GSK manufacture are not being considered. However, the fact that the model can not describe the solvents in their totality is already representing a limitation for the application of the model in solvent selection procedures.
3. Models Evaluation
This evaluation was based on the comparison between experimental data and solubility predictions for GSK molecules. All the molecules that made part in the evaluation step are in the scope of the models:

- Organic
- Non-electrolytes
- Non-solvates
- Fully incrementable by Pharma Mod. UNIFAC

The solubility values generated by the models in this exercise were for the systems included in the experimental database. While NRTL-SAC had the experimental data from the database fed into the solute parameters, Pharma Mod. UNIFAC, as a semi-empirical model, had not. This translates into NRTL-SAC predicting interpolations and PMUNIFAC predicting “extrapolations”. The objective of this stage was to assess if either NRTL-SAC or Pharma Mod. UNIFAC had the potential to be applied in solvent selection procedures. But it is not conclusive for NRTL-SAC, since solvent selection procedures require predicting potential for extrapolations (other systems and temperatures different from the experimental database).

Two different selections of the same database were targeted for the global analysis. The first selection being the complete set of solubility data (8 GSK API molecules that included 87 points) and, the second, the points correspondent to pairs API-solvent that have all PMUNIFAC functional groups and binary parameters available (3 APIs that included 13 points). It is the second scenario that can actually show the potential of the UNIFAC model in a fair way.

For each model and scenario, the difference between the logarithms to base 10 of the calculated and the experimental data was plotted against the experimental data. The plots enabled a visual perception of the accuracy of each model by the scattering of the data points and the included boundaries for one and half an order of magnitude up and down the reference to easily pin down how much the calculated data points differed from the experimental data. It is desirable that the points are mainly inside the half an order of magnitude boundaries and not so much to have the points outside the order of magnitude boundaries. The statistical analysis calculations are presented next.

Mean Absolute Deviation of the logarithm of base 10 (MAD LOG_{10}) - it gives the average for the order of magnitude of discrepancy between calculated and experimental values.

\[
\text{MAD LOG}_{10} = \frac{1}{N} \sum_{i} |\log(x_{\text{calc},i}) - \log(x_{\text{exp},i})| \quad (6)
\]

Maximum Logarithm Absolute Deviation (Max AD LOG_{10}) - it gives the maximum for the order of magnitude of discrepancy between calculated and experimental values. This value is sometimes an outlier due to experimental error.

\[
\text{Max AD LOG}_{10} = \max \{|\log(x_{\text{calc},i}) - \log(x_{\text{exp},i})|\} \quad (7)
\]

In Figures 3, 4 and 5, it is possible to visualise the accuracy of each model for the two selections of data, as well as the extent of the scattering.

![Figure 3: NRTL-SAC performance for the complete database.](image)

![Figure 4: PMUNIFAC performance for the complete database.](image)

![Figure 5: PMUNIFAC performance for the selected database of API-solvent pairs with all PMUNIFAC parameters available.](image)

Tables 2 and 3 display the statistical results for the two scenarios.
Looking at Table 2, NRTL-SAC shows a very good fit. The average MAD of 0.85 from PMUNIFAC would label the model as unsuitable for the application in solvent selection. But having the knowledge that the majority of the required parameters were not available (not yet regressed), that could be a precipitated conclusion before having looked at the second scenario. Indeed, Table 3 shows that the results from PMUNIFAC significantly improve when all interactions and functional groups are well described by parameters.

Figures 3 and 4 show that the deviations observed have reduced scattering for NRTL-SAC and low accuracy with high scattering for PMUNIFAC. But as can be seen from Figure 5, PMUNIFAC becomes more accurate when only points with structural groups and group interaction parameters available are considered. In Figure 5, the point with larger deviation is correspondent to Ethylene glycol that is considered a natural outlier by the author of the model\textsuperscript{10}, that may be due to its propensity to self-association, resulting in an equilibrium between different incremenantations.

To finish the study, the probability of prediction failure (PPF) for each model over the MAD LOG\textsubscript{10} was calculated and plotted. The PPF of solubility (mg API/ mL solvent) in the MAD LOG\textsubscript{10} equation 8, is used for comparison between models. Being $N_n$ relative to molecule ID, $N_{max}$ to maximum number of molecules (8 for the complete dataset and 3 for the selection) and value to MAD value.

$$\text{PPF} \times 100\% = 100\% \times (1 - \frac{\sum_{n=1}^{N_{max}} N_n \{(MAD \ LOG\textsubscript{10})_n \leq \text{value}\}}{N_{max}})$$

Both PPFs are shown in Figures 6 and 7. They show which percentage of data points possesses larger deviations than a certain value, for instance, in Figure 6, the probability to have a PMUNIFAC fit with a MAD LOG\textsubscript{10} value larger than 1 is 30%. Comparing to Figure 6, a great improvement for PMUNIFAC performance is clear in Figure 7.

### Table 2: Fit statistics of the two models for the common set of the company molecules. Number of APIs: 8; number of points: 87; number of solvents: 26; temperature range: 0.6 - 50 °C.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAD LOG\textsubscript{10}</th>
<th>Max AD LOG\textsubscript{10}</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMUNIFAC</td>
<td>0.85</td>
<td>6.12</td>
</tr>
<tr>
<td>NRTL-SAC</td>
<td>0.27</td>
<td>1.63</td>
</tr>
</tbody>
</table>

### Table 3: Fit statistics of the two models for the selected set of API-solvent pairs with all PMUNIFAC parameters available. Number of APIs: 3; number of points: 13; number of solvents: 6 solvents; temperature range: 6.2 - 47.5 °C.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAD LOG\textsubscript{10}</th>
<th>Max AD LOG\textsubscript{10}</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMUNIFAC</td>
<td>0.40</td>
<td>2.05</td>
</tr>
<tr>
<td>NRTL-SAC</td>
<td>0.18</td>
<td>0.74</td>
</tr>
</tbody>
</table>

3.1. Evaluation conclusions

It is possible to relate A, B and C with the thermodynamic properties by equation 3, but it does not mean that the properties can be calculated from the regressed parameters. In fact, it was found that many times, even with an extensive experimental database, A and B regression results would not have physical meaning, such as negative melting temperature.

PMUNIFAC proved to be a model with potential for future applications based on the predictions of great accuracy when all inputs were gathered. Nevertheless, it is still a work in progress, since the majority of binary parameters is missing. It was also conclusive that further binary interactions should be considered for future regressions, since having all considered ones available would still not be enough for application in solvent selection procedures. Even though NRTL-SAC results are highly dependent on the initial experimental database, they endorse that NRTL-SAC is a very strong model, at least for interpolations. The only model that could be fit for purpose in the solvent selection procedures is NRTL-SAC.
4. Solubility Modelling Application

The last part of the present study describes the application of solubility modelling with NRTL-SAC in an actual GSK project as an attempt to contribute for the improvement of processes in API manufacture. The regression of the parameters was done in Aspen Solubility Modeler V7.3. For the regression only the molar weight of the solid and experimental solubility data were required. The knowledge about the melting point was not required since A and B were regressed together with X, Y, Y' and Z.

Having the parameters, predictions were generated for 130 solvents at several temperatures, pure and binary mixtures with a selected co-solvent. The results were compiled in a new spreadsheet together with the relevant database columns for each case. Keeping in mind the target and solvent selection criteria, Excel filters were applied and a group of solvent alternatives stood out. New meetings were arranged where the discussion of the alternative solutions took place. Closing the cycle, follow-ups to assess the quality of the results were achieved by both author and project scientists.

It is worth mentioning that, unfortunately, the case that that came up in the short time window of this study was not organic non solvate and non-electrolyte (model scope), but hemi-hydrate organic electrolyte. The solvent selection analysis was taken forth anyway and it was possible to assess the value of NRTL-SAC outside its scope.

Some concepts used in the solvent selection procedures are clarified in the next paragraphs.

Common solvents - The favourable solvents. They are the ones that the company commonly uses in manufacture. These are listed in internal documentation of the company.

Form screen - The objective of a form screen is to discover and characterize crystal-forms of the API that are relevant to the project objectives. The screen includes a set of crystallisation experiments in several solvents and solvent mixtures to identify and characterize crystal-forms (polymorphs, hydrates, and solvates).

Process volume (vol) - This value is one of the most relevant to get a feel about the equipment dimensions requested by the process under consideration. It equals to the mL solvent per g input material. Normally it is desirable to have vol < 30.

Red solvents - The solvents to avoid. Their denomination comes from the consideration of several weights related with environmental, health and safety issues and are listed in internal documentation of the company.

Yield - Yield of the system undergoing the crystallisation. The yield can be calculated by equation 9, where solubility ($x_{\text{sat}}$) is in mg API/mL solvent, and vol and V in L.

$$\text{yield(\%)} = \left(\frac{x_{\text{sat}} \times \text{vol}_{\text{final}} - (x_{\text{sat}} \times \text{vol})_{\text{initial}}}{(x_{\text{sat}} \times \text{vol})_{\text{initial}}} \right) \times 100\% \quad (9)$$

$$\text{vol}_{\text{initial}} = \frac{1000 \text{ L/mL}}{x_{\text{sat}} \text{ initial}} \quad (10)$$

$$\text{vol}_{\text{final}} = \text{vol}_{\text{initial}} + V_{\text{added anti-solvent}} \quad (11)$$

4.1. Solubility Modelling - Case

The process to move from was a cooling crystallisation in 10% water/IPA. The kinetics of the crystallisation were very slow (>24h for cooling and ageing) and the final yield was not ideal (<70%). The API (Molecule D) is an hemi-hydrate HBr salt formed from the reaction of Molecule C with aqueous HBr. The process is represented in Figure 8 where the grey area is related to the solid state of the API and the coloured pointers illustrate heat exchange. All weights, volumes and equivalents are relative to Molecule C.

The data available on this HBr salt compound included solubility data in 18 solvents at 2 temperatures (20 and 50 °C). A form screen was also available, it was known that higher hydrates could be generated with solvent mediums with high water activity and HBr was only available in water source, max 48 %wt HBr in water. BuOH was not suitable because it promotes a different form. DMSO/THF system result in gelling. The objective was to identify solvent/anti-solvent pair or pure solvent for cooling crystallisation (by first intent) that would give improved crystallisation kinetics (currently >24h), avoid gelling and improve yield. The solvent selection procedure comprised:

1. Regression of Molecule D NRTL-SAC parameters from experimental data;
2. Calculation of solubility values for 130 pure solvents at 5, 20, 50 and 80 °C;
3. Calculation of solubility values for mixtures of 130 solvents with 10 %wt water at 20 °C;
4. Calculation of solubility values for mixtures of 130 solvents with 20, 30, 40, 60 %wt water at 5, 20, 50, 80 °C;
5. Analysis of calculated data for high temperature and low water content;
6. Evaluation of the results from the analysis through solubility experiments.

The results for the predictions fit with the NRTL-SAC parameters from the regression are condensed in Table 4 and illustrated in Figure 9.

Table 4: Fit statistics for Molecule D.

<table>
<thead>
<tr>
<th>Fit statistics - deviations</th>
<th>MAD LOG$_{10}$</th>
<th>Max AD LOG$_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.54</td>
<td>1.37</td>
<td></td>
</tr>
</tbody>
</table>


Figure 8: Process Definition Diagram of Molecule D from Molecule C.

The fit is not perfect, maybe due to the larger amount of points, possible experimental error and the molecule being an electrolyte (NRT-SAC parameters do not consider an electrolyte segment). Nevertheless, the fitted parameters still show some predicting potential, as observable from Figure 9.

Figure 9: Predicted and experimental solubility LOG values.

An API is, commonly, a large organic molecule which hampers its solubility. Hence, the bottleneck is to identify a suitable solvent for the crystallisation. Because one restriction for the solvent system is not to have high water activity, the focus was reduced to the 0 and 20 %wt water mixtures. The desired solvent system must provide high solubility and the model was sub-estimating some values, so the focus moved to the higher temperatures (50 and 80 °C). At 80 °C the solubility values are the highest but on the other hand, some solvents are not suitable anymore if they have a lower boiling point.

The criteria was applied to the calculated data as summarized in Table 5. The solvent should have solubility larger than 30 mg/mL since this translates into process volumes smaller than 30 vol. The solvent should not belong to the red category in first intent. Alcohols, bases and acids are to be avoided, since they may promote undesirable reactions. There will always be water in the process (the HBr is added in aqueous solution), so the solvent should be water miscible. Finally, the solvent should be preferably listed in the common solvents list from the company.

Table 6: Experimental versus predicted solubility values for the Acetone and wet Acetone systems.

<table>
<thead>
<tr>
<th>Temp. °C</th>
<th>%wt water (W)</th>
<th>W activity (molar)</th>
<th>vol</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.0</td>
<td>0</td>
<td>25.0</td>
</tr>
<tr>
<td>20</td>
<td>5.2</td>
<td>0.48</td>
<td>5.8</td>
</tr>
<tr>
<td>20</td>
<td>10.1</td>
<td>0.62</td>
<td>1.3</td>
</tr>
<tr>
<td>20</td>
<td>19.5</td>
<td>0.70</td>
<td>2.0</td>
</tr>
<tr>
<td>50</td>
<td>5.2</td>
<td>0.50</td>
<td>4.8</td>
</tr>
<tr>
<td>Criteria</td>
<td>50 °C Solubility &gt; 30 mg/mL</td>
<td>80 °C Solubility &gt; 30 mg/mL</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0 %wt water</td>
<td>20 %wt water</td>
<td>0 %wt water</td>
</tr>
<tr>
<td>Solubility</td>
<td>12</td>
<td>102</td>
<td>21</td>
</tr>
<tr>
<td>Red solvents</td>
<td>2</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>Alcohols/Bases/Acids</td>
<td>9</td>
<td>23</td>
<td>15</td>
</tr>
<tr>
<td>Water non-miscible</td>
<td>6</td>
<td>85</td>
<td>9</td>
</tr>
<tr>
<td>Common solvent</td>
<td>3</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>After restrictions</td>
<td>3</td>
<td>9</td>
<td>4</td>
</tr>
</tbody>
</table>

|                          | (0 common)   | (3 common: Acetone; 2-MeTHF; 1,4-dioxane) |

It was a case of success since 1 alternative came out of 1 proposed solvent mixture (wet acetone).

Despite the maximum value for solubility being around 10 %wt water content, 5 %wt is a better option, because oiling phenomena is promoted when the solute has such an affinity with the solvent (solubility too high), keeping the API from crystallising out. Plus, at 5 %wt water content a good process volume is already allowed ($\approx 6$ vol), and the water activity is at a reasonable value (not very high or low) to expect the right form - hemi-hydrate. In order to determine if the form of the API in the selected solvent system was the correct one, the form screen was consulted and it was found that the hemi-hydrate was the form obtained in the Acetone/5 %wt water system.

Having a strong candidate for solvent and cooling crystallisation being excluded (low yields), an anti-solvent had to be selected. From the initial experimental data, three candidates for anti-solvent were identified: t-Butylmethyl ether (TBME), Ethyl Acetate and Butyl Acetate.

In order to check the miscibility in the ternary systems, ternary maps were generated with the “Conceptual design” tool from Aspen Plus V7.3. All three candidates for anti-solvent were miscible with wet Acetone at process conditions. Although, the miscibility behaviour might change with the addition of the API.

Several solubility experiments were done for the three different ternary solvent systems so that the miscibility could be checked, the yield predicted and the API form determined. The study included volumetric ratios of anti-solvent:solvent of 1:1, 2:1, 3:1 and pure anti-solvent, all at ambient temperature.

The yields and final process volumes were calculated based on the initial values (in 5 % wet Acetone): solubility of 159.9 mg/mL and process volume of 6.3 vol. High yields and low process volumes were obtained for the three anti-solvents in volumetric ratio 2:1, especially for TBME.

The form was checked through IR spectroscopy. IR confirmed that TBME gives the same form as the reference (i.e., the hemi-hydrate). Butyl Acetate and Ethyl Acetate give a slightly different spectra, however, because the reference is a dry powder whilst the samples are slurries, there is a possibility they are actually giving the hemi-hydrate as well. The study was not carried further since an alternative process had been successfully identified - anti-solvent crystallisation at 20 °C of 5 %wt Acetone as solvent and TBME as anti-solvent in the volumetric ratio of 1:2, respectively.

Because there are always yield losses in steps previous to the crystallisation, such as vessel transfer, process yield $> 95\%$ is a realistic value.

The new proposal for the process diagram is presented in Figure 10. It has 2 steps less than initial (heating is no longer required), the system does not contain reactive species and the yield has improved significantly. The process volume increased from $\approx 11$ vol to $\approx 20$ vol, which is still a good value and that can be reduced compromising the yield. It is also expected a great decrease on the crystallisation time, but no crystallisation experiments were carried out during the study duration.

The project team members accepted the proposed alternative process and further experiments would be done for optimisation purposes before the scale-up.

**5. Conclusions**

It was found that NRTL-SAC, unlike PMUNIFAC at this time, can be successfully used for solvent selection. NRTL-SAC can be applied on electrolytes and solvates, however higher quality of dataset is required and lower success rate is expected from re-
Table 7: Experimental solubility for the ternary systems Solvent (S)/Anti-solvent (AS) at 20 °C with Solvent = 5 %wt wet Acetone.

<table>
<thead>
<tr>
<th>AS</th>
<th>Ratio AS/S (volumetric)</th>
<th>Solubility (mg/mL)</th>
<th>Yield (%)</th>
<th>Final Process Vol. (vol)</th>
<th>Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>None Pure Solvent</td>
<td>159.9</td>
<td>159.9</td>
<td></td>
<td>Initial vol = 6.3</td>
<td>Hemi-hydrate</td>
</tr>
<tr>
<td>Ethyl Acetate</td>
<td>1:1</td>
<td>20.32</td>
<td>87.3</td>
<td>12.5</td>
<td>Other</td>
</tr>
<tr>
<td>Ethyl Acetate</td>
<td>2:1</td>
<td>5.54</td>
<td>96.5</td>
<td>18.8</td>
<td></td>
</tr>
<tr>
<td>Ethyl Acetate</td>
<td>3:1</td>
<td>2.83</td>
<td>98.2</td>
<td>25.0</td>
<td></td>
</tr>
<tr>
<td>Ethyl Acetate</td>
<td>Pure AS</td>
<td>0.5</td>
<td>99.7</td>
<td>∞</td>
<td></td>
</tr>
<tr>
<td>Butyl Acetate</td>
<td>1:1</td>
<td>20.63</td>
<td>87.1</td>
<td>12.5</td>
<td>Other</td>
</tr>
<tr>
<td>Butyl Acetate</td>
<td>2:1</td>
<td>5.92</td>
<td>96.3</td>
<td>18.8</td>
<td></td>
</tr>
<tr>
<td>Butyl Acetate</td>
<td>3:1</td>
<td>2.74</td>
<td>98.3</td>
<td>25.0</td>
<td></td>
</tr>
<tr>
<td>Butyl Acetate</td>
<td>Pure AS</td>
<td>0.1</td>
<td>99.9</td>
<td>∞</td>
<td></td>
</tr>
<tr>
<td>TBME</td>
<td>1:1</td>
<td>10.85</td>
<td>93.2</td>
<td>12.5</td>
<td>Hemi-hydrate</td>
</tr>
<tr>
<td>TBME</td>
<td>2:1</td>
<td>1.07</td>
<td>99.3</td>
<td>18.8</td>
<td></td>
</tr>
<tr>
<td>TBME</td>
<td>3:1</td>
<td>0.19</td>
<td>99.9</td>
<td>25.0</td>
<td></td>
</tr>
<tr>
<td>TBME</td>
<td>Pure AS</td>
<td>0.0</td>
<td>100.0</td>
<td>∞</td>
<td></td>
</tr>
</tbody>
</table>

Figure 10: New Proposal for the Process Definition Diagram of Molecule D from Molecule C.

Solvent selection through solubility modelling has proven to add value in process design, after been applied with success on two GSK processes. The tool is best applied on early phase projects where not much data is yet available and the amount of material is limited, hence targeted experiments are conducted. Solvent selection cannot be based on modelling alone but requires experimental work.

6. Future Work
The next steps should focus on the improvement of the applicability and the accuracy of solubility predictions for certain solute-solvent pairs, when compared to NRTL-SAC. The main advantage of PMUNIFAC is that no experimental solubility data for solute is required, which widens model applicability within GSK significantly.

PMUNIFAC is currently not fit for purpose in GSK solvent selection due to the early stage of its development. Since it is highly limited in terms of parameters, only for a very small number of solvent-solute combinations all parameters are available. However, the results obtained for PMUNIFAC are encouraging to support its further development, since it showed to give comparable or even better solubility prediction for certain solute-solvent pairs, when compared to NRTL-SAC. Nevertheless, NRTL-SAC has disadvantages, such as the requirement of a representative set of experimental data for each solute (parameters regression) and the high dependency of the accuracy of prediction on the selection of dataset going into regression. Therefore, the identification of the "ideal" dataset (which is expected to give the highest accuracy on prediction) is ultimately, depending on the choice of the modeller.

PMUNIFAC is currently not fit for purpose in GSK solvent selection due to the early stage of its development. Since it is highly limited in terms of parameters, only for a very small number of solvent-solute combinations all parameters are available. However, the results obtained for PMUNIFAC are encouraging to support its further development, since it showed to give comparable or even better solubility prediction for certain solute-solvent pairs, when compared to NRTL-SAC. The main advantage of PMUNIFAC is that no experimental solubility data for solute is required, which widens model applicability within GSK significantly.

Solvent selection through solubility modelling has proven to add value in process design, after been applied with success on two GSK processes. The tool is best applied on early phase projects where not much data is yet available and the amount of material is limited, hence targeted experiments are conducted. Solvent selection cannot be based on modelling alone but requires experimental work.
models in order to allow the design and optimisation of unit operations to rely on solubility models more efficiently.

It would be desirable to improve the accuracy of predictive models such as Pharma Mod. UNIFAC to allow solubility modelling to be applied on solutes where no experimental data is available. For this, the existing Pharma Mod. UNIFAC parameter matrix is recommended to be revised, expanded and completed, which imply an increase of the data points that are fed into the parameter regression in order to generate further binary group interaction parameters. Since both the group contribution based models and the fully empirical models rely on datasets with a high quantity and quality of data points, it is strongly recommended that more efforts are made within the industry to organise existing solubility data. A central storage of solubility data would allow solubility predictions at an early stage of the solute development, enabling a more efficient process design at a later stage in development for a specific project. It would also enable model evaluation with a more relevant basis.

In order to widen the applicability of solubility models further to electrolytes and solvates, which account for a significant part of solutes that appear in the manufacturing processes of APIs, it is recommended to carry out an evaluation of existing models for electrolytes such as LIQUAC, LIFAC\cite{Mohs2013} and eNRTL-SAC\cite{Chen2005} in a similar manner as it has been done in this work. It is also recommended to explore models which take a different approach that may offer a viable alternative, such as the SAFT-gamma\cite{Lymperiadis2007} model, which is based on an equation-of-state approach.

Acknowledgements

The author would like to thank Dr.-Ing. Thoralf Hartwig, Prof. Dr. Benilde Saramago for all the support and company GlaxoSmithKline Pharmaceuticals for having provided with the working facilities and for the knowledge kindly shared.

References


[6] Prof. Dr. J. Gmehling. 7th common Meeting of UNIFAC Consortium and DDBST GmbH, Oldenburg, Germany, September 2009.


