Contribution of solid-state properties to the aqueous solubility of drugs

Carola M. Wassvik\textsuperscript{a}, Anders G. Holmén\textsuperscript{b,\ast}, Christel A.S. Bergström\textsuperscript{a}, Ismael Zamora\textsuperscript{c,d}, Per Artursson\textsuperscript{a}

\textsuperscript{a} Department of Pharmacy, Uppsala University, P.O. Box 580, SE-75123 Uppsala, Sweden
\textsuperscript{b} Lead Generation, AstraZeneca R&D SE-43183 Mölndal, Sweden
\textsuperscript{c} Lead Molecular Design, S.L., Vallés, 96-102 (Local 27) 08190 Sant Cugat del Vallés, Spain
\textsuperscript{d} Grupo de Recerca en Informática Biomédica, Universitat Pompeu Fabra, ES-08003 Barcelona, Spain

\textbf{A R T I C L E   I N F O}

Article history:
Received 5 October 2005
Received in revised form
23 March 2006
Accepted 31 May 2006
Published on line 7 June 2006

Keywords:
Intrinsic solubility
Solubility prediction
Solid-state property
Drug solubility
Oral drug space
Solubility data set
Melting point
Enthalpy of melting
Entropy of melting
General solubility equation

\textbf{A B S T R A C T}

This study investigates the influence of the solid-state properties melting point ($T_m$), enthalpy of melting ($\Delta H_m$) and entropy of melting ($\Delta S_m$) of a drug on its intrinsic solubility ($S_0$). For this purpose, 26 chemically and structurally diverse drugs covering the oral drug space were selected and the $S_0$, $T_m$, $\Delta H_m$ and $\Delta S_m$ were determined experimentally. The influence of $T_m$, $\Delta H_m$ and $\Delta S_m$ on $S_0$ was studied using regression analysis. The overall improvement of the predictions were 0.3 log units, however, five compounds (astemizole, glyburide, fenbufen, gliclazide and griseofulvin) were improved by more than one log unit. $T_m$ and $\Delta H_m$ had a larger effect than $\Delta S_m$ on the solubility predictions. The well-known general solubility equation (GSE) and the Dannenfelser semi-empirical equation for the calculation of $\Delta S_m$ were evaluated using our data set. While predictions of drug solubility obtained using the GSE were acceptable, the use of the experimental $\Delta S_m$ values instead of the constant 56.5 J mol\(^{-1}\) K\(^{-1}\) improved the accuracy of the prediction. The Dannenfelser equation underestimated the $\Delta S_m$ for most compounds with on average 15 J mol\(^{-1}\) K\(^{-1}\). Our results show that solid-state properties should be considered for improved performance of future models for prediction of drug solubility. In addition our study provides accurate experimental data on intrinsic solubility for 26 compounds, supplying a useful external data set for validation of drug solubility models.

© 2006 Elsevier B.V. All rights reserved.

1. Introduction

During recent years there has been tremendous progress in the field of predicting the aqueous solubility of crystalline drug molecules. From a pharmaceutical perspective, the crystalline solid is often the solid-state of choice when developing a drug into a usable product and it would be of great value to be able to accurately predict the intrinsic solubility of crystalline drug molecules. This would improve the quality of the selection of compounds for synthesis and in vitro and in vivo testing, as well as improve our understanding of how structural variations change the solubility.

Several types of prediction models for intrinsic solubility have been reported in the literature. They can roughly be divided into three main categories: (I) models with a clear basis in thermodynamics (Jain and Yalkowsky, 2001; Klamt et al., 2002; Ruelle et al., 1991), (II) models based on chemical atom (Hou et al., 2004) or group contributions (Klopman...
and Zhu, 2001; Kuhne et al., 1995) and (iii) statistically derived models based on 2D or 3D chemical descriptors (Abraham and Le, 1999; Catana et al., 2005; Delaney, 2004; Huuskonen et al., 2000; Katritzky et al., 1998; Lind and Maltseva, 2003; Meylan and Howard, 2000; Raevsky et al., 2004; Tetko et al., 2001; Votano et al., 2004; Yan and Gasteiger, 2003). Several of these models provide comparable outcomes, with an average prediction error of about 0.7–0.8 log units for a chemically diverse set of drugs (Bergström et al., 2004; Jorgensen and Duffy, 2002). It is important to note that the data sets used to develop solubility models often cover a large range of solubilities (more than 10 orders of magnitude) and that they consist mainly of compounds that are structurally different from drugs.

In our previous work on the modeling of intrinsic solubility from molecular descriptors, we used data sets comprised exclusively of drug-like molecules, and found that a small number of descriptors was sufficient to obtain global models for drug-like molecules that provided similar outcomes to the models mentioned above (Bergström et al., 2002, 2004). It would be desirable to reduce prediction errors on average but it is even more important to find models that are more inclusive in the sense that the number of outliers is reduced significantly. For instance, when Yan and Gasteiger tested the performance of their neural network model on a data set from Merck KGaA that contained “more diverse compounds”, several compounds in the solubility range of mM to nM had residuals as large as 3–4 log units (a 1000- to 10,000-fold difference in solubility) (Yan and Gasteiger, 2003).

The majority of the solubility models mentioned above are mainly based on molecular descriptors related to the solvation of the drug molecule, most commonly represented by the octanol/water partition coefficient (log P). It is suggested that inclusion of a better description of the solid-state properties of the molecules in these models would improve their accuracy. Of particular interest in this respect is the general solubility equation (GSE) developed by Yalkowsky and co-workers (Jain and Yalkowsky, 2001; Yalkowsky and Valvani, 1980), which is based on the logP together with the melting point (Tm). This semi-empirical equation has been validated in the literature using a large set of simple organic molecules, herbicides and pesticides, but very few drugs (Ran et al., 2001; Ran and Yalkowsky, 2001).

The aim of this work was to determine how the stability of the solid-state of crystalline drugs influences their intrinsic solubility (S0). For this purpose we selected 26 drug compounds that displayed only a rough log S0 – log P correlation, since we hypothesised that the solubility of such compounds would be significantly influenced by their solid-state properties. The data set was also selected to cover solubilities in a range considered to be relevant for drug candidates intended for oral administration (mM to nM) and to cover the oral drug space. We determined the intrinsic solubility of the data set and characterised its solid-state properties using differential scanning calorimetry (DSC). Statistical analysis of the data set showed that, apart from logP, solid-state properties such as the enthalpy of melting (ΔHm) and Tm contributed significantly to the modeling of solubility. Thus, for some compounds the difference between the observed and the predicted S0 was reduced by as much as 1.5 log units when solid-state properties were considered. It is clear that the stability of the solid-state as quantified by thermochemical properties distinguishes compounds with low solubility from those with higher solubility.

Finally, as a first step towards the development of a purely theoretical description of the contribution of solid-state properties to predictions of drug solubility, the semi-empirical equation for the calculation of ΔSm developed by Dannenfelder and Yalkowsky (Dannenfelder and Yalkowsky, 1996, 1999) was evaluated against our experimental data. It was found that in general, the ΔSm of the compounds in our data set were underestimated by on average 15 J mol⁻¹ K⁻¹.

2. Materials and methods

2.1. Data set

The data set comprised compounds that (i) were drugs or drug-like, (ii) were in their free form (no salts or solvates), (iii) preferably did not exhibit polymorphism (if they did, experimental values for the stable polymorph were used), (iv) were stable at temperatures above the Tm (to allow accurate measurement of the ΔHm) and (v) were commercially available.

The drugs were selected to be structurally and chemically diverse and to cover most of the oral drug space. This selection was made by the use of the ChemGPS methodology (see Section 2.5). In Fig. 1, the 26 drugs are represented together with 402 orally administered drugs obtained from an AstraZeneca in-house database as a reference of the oral drug space. Our set of 26 drugs covers the greater part of the oral drug space. They were also selected to cover the range of solubilities (mM to nM) considered relevant for drug candidates intended for oral administration. Finally, the data set was selected to reflect only

[Fig. 1 – Diversity analysis of the data set used in the present study (N = 26) according to ChemGPS methodology. The first three principal components, t1[1], t2[2] and t3[3], mainly represent the size, polarity and flexibility, respectively, of the molecules. As a representation of the oral drug space, a set of 402 orally administered drugs (open circles) from an AstraZeneca in-house database are projected together with our data set (filled circles).]
a rough correlation between log P and log S_0. We hypothesised that outliers from the log P – log S_0 correlation were limited in solubility by factors other than log P, such as those related to the solid-state.

The physicochemical properties and experimentally determined intrinsic solubility and solid-state property data for the 26 drugs are compiled in Table 1 and the chemical structures are given in Fig. 2.

2.2. Chemicals

Chlorpropamide was purchased from MP Biomedicals LLC, OH, USA, phenytoin from Lancaster synthesis Ltd., UK, tamoxifen from ICN Biomedicals Inc., California, USA, and all other drugs from Sigma–Aldrich Chemie GmbH, Germany. The purity of all drugs used was greater than 98% with the exception of griseofulvin (a natural product) which had a purity of 96%. Ammonium acetate (NH₄Ac) and acetonicitile (AcN) (isocratic grade for liquid chromatography) were purchased from Sigma–Aldrich Chemie GmbH, Germany and Merck KGaA, Germany, respectively. Ultra-pure water (resistivity at 25°C: 18.2 MΩ cm) filtered through a Milli-Q® system from Millipore, MA, USA, was used in all experiments.

2.3. Differential scanning calorimetry (DSC)

Thermograms were recorded using a Seiko instrument consisting of a DSC220C analysis module with automatic cooling controller and analysed with EXSTAR6000 software version 3.4A (Seiko Instruments Inc., Japan) running on an HP712/100 work station. Duplicate samples of 1–3 mg were accurately weighed in sealed and pierced aluminium pans (TA Instruments, Delaware, USA). The instruments were calibrated for T_m and ΔH_m using high purity gallium (Ga) (Sigma–Aldrich GmbH, Germany), indium (In), tin (Sn) (Acrôs Organics, NJ, USA) and zinc (Zn) (TA Instruments, DE, USA) standards. T_m and ΔH_m were calculated using the GSE and/or based on previous determinations found in the literature.

2.4. Solubility determinations

The S_0 (expressed as the log S_0 in M) of crystalline compounds was determined in quadruplicate according to the shake-flask method described by Bergström et al. (2002). First of all, a rough estimation of the expected value of S_0 was made from the log P and T_m using the GSE and/or based on previous determinations found in the literature. At least three times excess of solid was weighed into 1.5 ml Eppendorf tubes, 1 ml of distilled water was added and the samples were thoroughly mixed on a vortex in order to achieve maximum wetting of the solid. For weak bases and weak acids, the pH was adjusted to at least 2 pH units above (bases) or below (acids) the pK_a with 0.01 M NaOH or 0.01 M HCl to ensure that all of the molecules were present in their neutral form. For neutral (bases with pK_a < 2 or acids with pK_a > 12) and zwitterionic compounds, pH was not adjusted. The samples were agitated on an orbital plate shaker at 300 rpm for 24 h at room temperature (21 ± 0.5°C). The pH was then measured and the presence of undissolved material was confirmed before the samples were centrifuged in an Eppendorf centrifuge model 5403 for 15 min at a relative centrifugal acceleration of 23,500 × g to separate the saturated solution from the solid. After centrifugation, 0.5 ml of the supernatant was carefully pipetted using a Pasteur glass pipette into 2 ml HPLC auto sampler glass vials and the samples were analysed by HPLC.

To validate our method of using centrifugation rather than filtration to separate the solid from the saturated solution, we compared the S_0 values obtained using the two methods for 15 out of the 26 compounds, for which the experiments were performed at AstraZeneca R&D Mölndal where we had access to vacuum filtration equipment. The solubility experiments were performed according to the protocol above with one addition. After the removal of ~0.5 ml of the supernatant, the deposited solid was re-suspended in the remaining solution and the resulting suspension was filtered through 96-well GF/B filters with an average pore size of 1.3 μm (Whatman International Ltd., Kent, UK) and analysed as explained below. The values for the centrifuged and filtered samples originated from the same sample.

For both methods, the largest experimental variation was recorded for compounds at the poorly soluble end of the scale. A regression analysis of the centrifuged and filtered samples gave a correlation with R^2 = 0.998. Thus, the two experimental approaches were in excellent agreement. The correlation for the 15 drugs is supplied as supporting information (Fig. S1).

Drug concentrations were determined using either HPLC–UV or HPLC–MS. In general, a standard curve of five concentration levels was used for quantification. Mobile phase A consisted of 95% 10 mM NH₄Ac buffer : 5% AcN and mobile phase B of 5% 10 mM NH₄Ac buffer : 95% AcN. The choice of LC method depended on the chromatographic behaviour of the compound. Either of these two were used: (1) a reversed phase C₈-separation column, SymmetryShield™ RP₈, 5 μm, dimensions 3.9 mm × 150 mm (Waters Corporation, MA, USA), with the following gradient at a flow rate of 1 ml min⁻¹: 0–2 min, 0% B; 2–12 min, 0–100% B; 12–15 min, 100% B; 15–15.5 min, 100–0%.
<table>
<thead>
<tr>
<th>Name</th>
<th>Drug</th>
<th>CAS #</th>
<th>log S₀ ± S.D.</th>
<th>nₑS</th>
<th>Reference</th>
<th>pKₑ</th>
<th>Acid/ base</th>
<th>CLOGP</th>
<th>Tₑm ± S.D.</th>
<th>∆Hₑm ± S.D.</th>
<th>∆Sₑm ± S.D.</th>
<th>nₑDSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AST</td>
<td>Astemizole</td>
<td>6884-77-9</td>
<td>-7.18 ± 0.14</td>
<td>3</td>
<td>RES</td>
<td>4.9:8.7</td>
<td>b:b</td>
<td>6.09</td>
<td>174.4 ± 0.10</td>
<td>51.1 ± 0.77</td>
<td>114.1 ± 1.72</td>
<td>3</td>
</tr>
<tr>
<td>CHL</td>
<td>Chlorpropamide</td>
<td>94-20-2</td>
<td>-3.30 ± 0.002</td>
<td>4</td>
<td>RES</td>
<td>4.8</td>
<td>a</td>
<td>2.35</td>
<td>127.8 ± 0.07</td>
<td>25.7 ± 0.41</td>
<td>64.0 ± 0.102</td>
<td>2</td>
</tr>
<tr>
<td>CLO</td>
<td>Clozapine</td>
<td>5786-21-0</td>
<td>-4.64 ± 0.03</td>
<td>4</td>
<td>RES</td>
<td>3.6:7.9</td>
<td>b:b</td>
<td>4.93</td>
<td>183.9 ± 0.07</td>
<td>35.9 ± 0.53</td>
<td>78.4 ± 1.16</td>
<td>2</td>
</tr>
<tr>
<td>DIP</td>
<td>Diazepam</td>
<td>439-14-5</td>
<td>-3.85 ± 0.01</td>
<td>4</td>
<td>RES</td>
<td>3.4</td>
<td>b</td>
<td>3.17</td>
<td>131.6 ± 0.03</td>
<td>24.7 ± 0.10</td>
<td>61.0 ± 0.24</td>
<td>4</td>
</tr>
<tr>
<td>DIX</td>
<td>Diazoxide</td>
<td>364-98-7</td>
<td>-3.60 ± 0.01</td>
<td>4</td>
<td>RES</td>
<td>8.5</td>
<td>a</td>
<td>1.20</td>
<td>327.2 ± 0.24</td>
<td>34.1 ± 1.12</td>
<td>56.8 ± 1.89</td>
<td>4</td>
</tr>
<tr>
<td>DES</td>
<td>Diethylstilbestrol</td>
<td>56-53-1</td>
<td>-5.00 n.s.</td>
<td>n.s.</td>
<td>1</td>
<td>n</td>
<td>n</td>
<td>4.96</td>
<td>177.9 ± 0.21</td>
<td>33.4 ± 2.24</td>
<td>73.9 ± 4.96</td>
<td>3</td>
</tr>
<tr>
<td>FEN</td>
<td>Fenbufen</td>
<td>36330-85-5</td>
<td>-5.19 ± 0.06</td>
<td>4</td>
<td>RES</td>
<td>4.5</td>
<td>a</td>
<td>3.14</td>
<td>186.1 ± 0.15</td>
<td>46.2 ± 0.12</td>
<td>100.5 ± 4.64</td>
<td>3</td>
</tr>
<tr>
<td>GLC</td>
<td>Gliclazide</td>
<td>21187-98-4</td>
<td>-4.07 ± 0.033</td>
<td>4</td>
<td>RES</td>
<td>6.2</td>
<td>a</td>
<td>1.09</td>
<td>171.4 ± 0.14</td>
<td>44.2 ± 0.63</td>
<td>99.4 ± 1.45</td>
<td>2</td>
</tr>
<tr>
<td>GLY</td>
<td>Glyburide</td>
<td>10238-21-8</td>
<td>-7.05 ± 0.19</td>
<td>4</td>
<td>RES</td>
<td>5.3</td>
<td>a</td>
<td>4.24</td>
<td>173.6 ± 0.13</td>
<td>46.3 ± 0.09</td>
<td>103.7 ± 0.23</td>
<td>2</td>
</tr>
<tr>
<td>GRI</td>
<td>Griseofulvin</td>
<td>126-07-8</td>
<td>-4.83 ± 0.08</td>
<td>3</td>
<td>2</td>
<td>n</td>
<td>n</td>
<td>1.77</td>
<td>218.0 ± 0.00</td>
<td>44.7 ± 0.78</td>
<td>90.8 ± 1.59</td>
<td>3</td>
</tr>
<tr>
<td>HYD</td>
<td>Hydrochlorothiazide</td>
<td>58-93-5</td>
<td>-2.70 ± 0.03</td>
<td>3</td>
<td>2</td>
<td>7.9:9.2</td>
<td>a:a</td>
<td>-0.37</td>
<td>267.6 ± 0.41</td>
<td>33.6 ± 0.14</td>
<td>62.2 ± 0.31</td>
<td>2</td>
</tr>
<tr>
<td>IBU</td>
<td>(±)-Ibuprofen</td>
<td>15687-27-1</td>
<td>-3.38 ± 0.03</td>
<td>4</td>
<td>RES</td>
<td>4.4</td>
<td>a</td>
<td>3.68</td>
<td>73.2 ± 0.10</td>
<td>26.6 ± 0.15</td>
<td>76.8 ± 0.42</td>
<td>3</td>
</tr>
<tr>
<td>INM</td>
<td>Indomethacin</td>
<td>53-86-1</td>
<td>-5.95 ± 0.01</td>
<td>4</td>
<td>RES</td>
<td>4.1</td>
<td>a</td>
<td>4.18</td>
<td>159.8 ± 0.03</td>
<td>37.9 ± 0.18</td>
<td>87.6 ± 0.43</td>
<td>3</td>
</tr>
<tr>
<td>INP</td>
<td>(±)-Indopropen</td>
<td>31842-01-0</td>
<td>-4.72 ± 0.12</td>
<td>4</td>
<td>RES</td>
<td>4.6</td>
<td>a</td>
<td>2.74</td>
<td>211.4 ± 0.46</td>
<td>40.3 ± 2.38</td>
<td>83.2 ± 4.84</td>
<td>3</td>
</tr>
<tr>
<td>KET</td>
<td>(±)-Ketoprofen</td>
<td>22071-15-4</td>
<td>-3.52 ± 0.01</td>
<td>4</td>
<td>RES</td>
<td>4.0</td>
<td>a</td>
<td>2.76</td>
<td>94.8 ± 0.12</td>
<td>37.3 ± 0.33</td>
<td>101.2 ± 0.12</td>
<td>3</td>
</tr>
<tr>
<td>MIF</td>
<td>Mifepristone</td>
<td>84371-65-3</td>
<td>-5.75 ± 0.02</td>
<td>3</td>
<td>RES</td>
<td>5.2</td>
<td>b</td>
<td>4.46</td>
<td>193.9 ± 0.14</td>
<td>31.7 ± 0.61</td>
<td>67.9 ± 1.33</td>
<td>3</td>
</tr>
<tr>
<td>NAP</td>
<td>Naproxen</td>
<td>22204-53-1</td>
<td>-4.23 ± 0.02</td>
<td>4</td>
<td>RES</td>
<td>4.2</td>
<td>a</td>
<td>2.82</td>
<td>155.6 ± 0.12</td>
<td>34.2 ± 0.85</td>
<td>79.7 ± 2.01</td>
<td>3</td>
</tr>
<tr>
<td>PER</td>
<td>Perphenazine</td>
<td>58-39-9</td>
<td>-4.62 ± 0.02</td>
<td>4</td>
<td>RES</td>
<td>7.8</td>
<td>b</td>
<td>4.31</td>
<td>96.8 ± 0.36</td>
<td>41.8 ± 0.69</td>
<td>113.0 ± 1.91</td>
<td>3</td>
</tr>
<tr>
<td>PHA</td>
<td>Phencetin</td>
<td>62-44-2</td>
<td>-2.48 ± 0.002</td>
<td>4</td>
<td>RES</td>
<td>n</td>
<td>n</td>
<td>1.77</td>
<td>154.2 ± 0.06</td>
<td>34.1 ± 0.92</td>
<td>83.7 ± 2.27</td>
<td>3</td>
</tr>
<tr>
<td>PHY</td>
<td>Phenyltoin</td>
<td>57-41-0</td>
<td>-4.15 ± 0.04</td>
<td>4</td>
<td>RES</td>
<td>8.3</td>
<td>a</td>
<td>2.09</td>
<td>295.6 ± 0.28</td>
<td>40.1 ± 0.75</td>
<td>70.4 ± 2.97</td>
<td>2</td>
</tr>
<tr>
<td>PIR</td>
<td>Piroxicam</td>
<td>36322-90-4</td>
<td>-4.03 ± 0.01</td>
<td>4</td>
<td>RES</td>
<td>4.5:3.6</td>
<td>a:b</td>
<td>1.89</td>
<td>200.3 ± 0.45</td>
<td>36.3 ± 0.15</td>
<td>76.7 ± 0.25</td>
<td>3</td>
</tr>
<tr>
<td>PRO</td>
<td>Probenecid</td>
<td>57-66-9</td>
<td>-4.90 ± 0.10</td>
<td>4</td>
<td>2</td>
<td>3.4</td>
<td>a</td>
<td>3.37</td>
<td>198.9 ± 0.06</td>
<td>40.9 ± 0.20</td>
<td>86.7 ± 0.43</td>
<td>2</td>
</tr>
<tr>
<td>SUL</td>
<td>Sulindac</td>
<td>38194-50-2</td>
<td>-4.78 ± 0.03</td>
<td>4</td>
<td>RES</td>
<td>4.7</td>
<td>a</td>
<td>3.16</td>
<td>187.0 ± 0.06</td>
<td>33.4 ± 0.36</td>
<td>72.5 ± 0.78</td>
<td>3</td>
</tr>
<tr>
<td>TAM</td>
<td>Tamoxifen</td>
<td>10540-29-1</td>
<td>-8.49 ± 0.22</td>
<td>4</td>
<td>RES</td>
<td>8.7</td>
<td>b</td>
<td>6.82</td>
<td>97.8 ± 0.29</td>
<td>34.0 ± 0.40</td>
<td>91.6 ± 1.14</td>
<td>3</td>
</tr>
<tr>
<td>TES</td>
<td>Testosterone</td>
<td>58-22-0</td>
<td>-4.20 ± 0.08</td>
<td>4</td>
<td>2</td>
<td>n</td>
<td>n</td>
<td>3.22</td>
<td>153.3 ± 0.36</td>
<td>28.2 ± 0.61</td>
<td>66.1 ± 1.48</td>
<td>3</td>
</tr>
<tr>
<td>TRI</td>
<td>Trimethoprim</td>
<td>738-70-5</td>
<td>-2.87 ± 0.003</td>
<td>4</td>
<td>RES</td>
<td>7.5</td>
<td>b</td>
<td>0.98</td>
<td>199.7 ± 0.10</td>
<td>49.8 ± 0.39</td>
<td>105.4 ± 0.85</td>
<td>3</td>
</tr>
</tbody>
</table>

- **AST**: Astemizole
- **CHL**: Chlorpropamide
- **CLO**: Clozapine
- **DIP**: Diazepam
- **DIX**: Diazoxide
- **DES**: Diethylstilbestrol
- **FEN**: Fenbufen
- **GLC**: Gliclazide
- **GLY**: Glyburide
- **GRI**: Griseofulvin
- **HYD**: Hydrochlorothiazide
- **IBU**: (±)-Ibuprofen
- **INM**: Indomethacin
- **INP**: (±)-Indopropen
- **KET**: (±)-Ketoprofen
- **MIF**: Mifepristone
- **NAP**: Naproxen
- **PER**: Perphenazine
- **PHA**: Phencetin
- **PHY**: Phenyltoin
- **PIR**: Piroxicam
- **PRO**: Probenecid
- **SUL**: Sulindac
- **TAM**: Tamoxifen
- **TES**: Testosterone
- **TRI**: Trimethoprim

---

---

---

---
Fig. 2 – Chemical structures of the 26 drugs in this study.

B; 15.5–25 min, 0% B (re-equilibration) and (2) a reversed phase C$_{18}$-separation column, X Terra® MS C$_{18}$, 3.5 μm, dimensions 2.1 mm × 100 mm (Waters Corporation, MA, USA), with the following gradient at a flow rate of 0.3 ml min$^{-1}$: 0–2 min, 0% B; 2–12 min, 0–100% B; 12–15 min, 100% B; 15–16 min, 100–0% B; 16–40 min, 0% B (re-equilibration). Compounds were detected using UV absorbance at three wavelengths: 210, 230 and 260 nm. The HPLC system consisted of a Midas type 830 auto sampler (Spark Holland BV, The Netherlands), a MERCK solvent degasser unit L-7612 (Coricon, Sweden), a BICHOFF binary pump model 2250 and a Multiwavelength Detector DAD-3L, controlled by McDAcq32 Control software version 2.0 (BISCHOFF Analysenteknik und geräte GmbH, Germany). For the HPLC-MS analysis, we used an Agilant 1100 Series LC system (binary pump, degasser and diode-array detector), a CTC PAL auto injector and a Micromass LCT time-of-flight
mass spectrometer operating in positive ion mode. Data for the M+H ions were used for quantification with the Micromass Quanlynx software version 4.0.

2.5. Diversity analysis and molecular descriptors

The ChemGPS (Oprea and Gottfries, 2001) diversity analysis was based on molecular descriptors generated using Selma, an AstraZeneca in-house software package that calculates a number of 2D descriptors. In total, this analysis used 90 molecular descriptors extracted from large suites of descriptors such as the Kier and Hall, Molconn-Z and the BCUT (Burden-Chemical Abstracts, University of Texas) parameters related to molecular size, ring structure, flexibility as well as molecular connectivity, electronic environment, charge and lipophilicity (Olsson and Sherbuhkin, 2006). The first three principal components of the principal component analysis (PCA), t[1], t[2] and t[3] (the x-, y- and z-axes in Fig. 1), were primarily related to size, polarity and flexibility, respectively.

The calculated log P (CLOGP) value was used as a simple measure of solvation forces. It was calculated using Selma according to the method by Leo and Weininger as implemented in Daylight toolkit V 4.7.1.

2.6. Statistical analysis

Multivariate data analysis techniques, PCA and projection to latent structures by means of partial least squares (PLS) were used as implemented in the Simca-P version 10.0 software (Umetrics AB, Umeå, Sweden). PLS was used to model the additional effect of ΔH_m and ΔS_m on solubility and PCA was used for the diversity analysis and for the analysis of what properties contributed the most to the improvement. Data were mean centred and scaled to unit variance and leave-many-out (seven groups) cross-validation was used in the PLS. In total, this analysis used 90 molecular descriptors extracted from large suites of descriptors such as the Kier and Hall, Molconn-Z and the BCUT (Burden-Chemical Abstracts, University of Texas) parameters related to molecular size, ring structure, flexibility as well as molecular connectivity, electronic environment, charge and lipophilicity (Olsson and Sherbuhkin, 2006). The first three principal components of the principal component analysis (PCA), t[1], t[2] and t[3] (the x-, y- and z-axes in Fig. 1), were primarily related to size, polarity and flexibility, respectively.

The calculated log P (CLOGP) value was used as a simple measure of solvation forces. It was calculated using Selma according to the method by Leo and Weininger as implemented in Daylight toolkit V 4.7.1.

2.7. Entropy calculations

ΔS_m values were calculated using the semi-empirical equation developed by Dannenfelser and Yalkowsky (1996, 1999):

$$\Delta S_m = 50 - R \ln \sigma + R \ln \Phi$$  \hspace{1cm} (2)

where σ is the rotational symmetry number, Φ the molecular flexibility number and R is the universal gas constant 8.31 J mol\(^{-1}\) K\(^{-1}\). A program capable of automatically calculating these descriptors was kindly provided by Dr. Jan Westergren at AstraZeneca R&D Mölndal, Sweden. This program is based on the rules outlined by Dannenfelser and Yalkowsky (1996, 1999).

3. Results

3.1. Contribution of lipophilicity to intrinsic solubility

The general relationship between lipophilicity (expressed as CLOGP) and solubility (log S_0) has been analysed using regression analysis for a data set of 270 drug-like compounds (Bergström et al., 2004). The regression of the 270 compounds is represented in Fig. 3 and the resulting regression equation is:

$$\log S_0 = -(1.91 \pm 0.07) - (0.617 \pm 0.02)\text{CLOGP}, \quad N = 270.$$  \hspace{1cm} (3)

Regression parameters are indicated as ±1 S.D. and the statistical parameters given are N is the number of compounds, R\(^2\) the coefficient of determination, F the test statistic from the F-test and RMSE is the root mean square error. The 26 drugs used in this study are also represented in Fig. 3 for comparison. There was no overlap between the 270 compounds used in the regression analysis and the 26 compounds used in this study. Eq. (3) was then used for calculating the solubility and these values were compared with the experimental values for the 26 drugs in Fig. 4a. Most of the compounds were situated below the line of unity, indicating that they generally had a lower experimental solubility than predicted from the log S_0 – CLOGP correlation. For some compounds (glyburide and tamoxifen), the residual was as large as 2.5 log units. On average, the compounds were over-estimated by 0.79 log units. These results suggest that features other than those related to solvation, as described by CLOGP, contribute to the solubility of the compounds in this data set.

3.2. Contribution of connectivity, electronic environment, charge and lipophilicity

Fig. 3 – Regression analysis of the intrinsic solubility (log S_0) and the calculated octanol/water partition coefficient (CLOGP) of 270 drug-like compounds (open circles). The 26 drugs used in this study (filled circles) are also projected in the figure but were not included in the regression analysis.
3.2. Contribution of solid-state properties to intrinsic solubility

We therefore investigated if a better relationship could be obtained with a semi-empirical equation for solubility, GSE (Eq. (4)) (Jain and Yalkowsky, 2001), that in addition to log $P$ also includes $T_m$ as a parameter of the solid-state

$$\log S_0 = 0.5 - 0.01(T_m - 25) - \log P$$

(4)

Inclusion of $T_m$ in the analysis improved $R^2$ from 0.67 of the CLOGP-predicted log $S_0$ (log $S_{\text{CLOGP}}$; Fig. 4a) to 0.73, and reduced the S.E. from $-0.79$ to $-0.49$ and the RMSE from 1.12 to 0.90 log units (Fig. 4b). Interestingly, many of the compounds that were significantly over-predicted in Fig. 4a moved closer to the line of unity in Fig. 4b. For example, the residual for astemizole was $-1.51$ when predicted by CLOGP and $-0.10$ when predicted by GSE, an improvement of 1.41 log units. Thus, GSE performed relatively well for this diverse set of drugs with fairly complex chemical structures despite the associated assumptions of ideal solubility and constant entropy of melting (Jain and Yalkowsky, 2001). It is also apparent that the inclusion of the solid-state parameter $T_m$ in solubility predictions improves the accuracy of the predictions for data sets comprised exclusively of drugs.

Because the experimental values of $\Delta S_m$ in our data set differed substantially from the constant of 56.5 J mol$^{-1}$ K$^{-1}$ (Walden’s rule) assumed in the GSE, we replaced the constant with our experimental $\Delta S_m$ values. Thus, we combined Eq. (14) together with Eq. (26) from Jain and Yalkowsky’s derivation of the GSE (Jain and Yalkowsky, 2001) and obtained:

$$\log S_0 = 0.5 - \frac{\Delta S_m}{570.55}(T_m - 25) - \log P$$

(5)

In Fig. 4c, the observed versus predicted solubility values using Eq. (5) are represented. The overall correlation increased ($R^2 = 0.78$) and the RMSE decreased (0.73 log units) as compared to the original GSE. The sign of S.E. changed from negative to positive when the measured $\Delta S_m$ was included, indicating that compounds, on average, are predicted to be less soluble (Eq. (5)) rather than predicted to be more soluble (Eq. (4)) than the experiments show. Overall, the solubility predictions improved significantly when the constant value used in the GSE was substituted by the experimental values of the $\Delta S_m$. 

Fig. 4 – Experimental intrinsic solubility ($\log S_0$) vs. predicted log $S_0$ for the data set used in this study ($N = 26$). (a) log $S_0$ predicted by CLOGP according to Eq. (3); (b) log $S_0$ predicted by the GSE in Eq. (4); (c) log $S_0$ predicted by the modified GSE in Eq. (5); (d) log $S_0$ predicted by CLOGP, $\Delta H_m$ and $\Delta S_m$, where the contribution from CLOGP is the same as in (a) and is held constant while the contributions from $\Delta H_m$ and $\Delta S_m$ are first modelled by PLS and then added to the CLOGP contribution. The diagonal line represents the line of unity.
with properties, described the relationship of the combined solid-state prop-
erties. This analysis resulted in a one-latent-variable model with
\( Q^2 \) of 0.20 and \( R^2 \) = 0.67 for the 26 compounds of this data set.

3.3. The effect of \( \Delta H_m \) and \( \Delta S_m \) on \( S_0 \)

The solubility of some compounds was more accurately pre-
dicted when \( \Delta H_m \) and \( \Delta S_m \) were included together with
CLOGP, while others were more poorly predicted (Fig. 4).

In an attempt to further investigate the degree to which
solid-state properties contribute to the intrinsic solubility of
drugs, the residuals from log \( S_{CLOGP} \) in Fig. 4a were corre-
lated with the experimental solid-state properties \( T_m \), \( \Delta H_m \)
and \( \Delta S_m \). The residuals (Res\( S_{CLOGP} \)) were calculated as the
\( S_0 - \log S_{CLOGP} \). As seen in Fig. 5a-c, \( R^2 \) was highest for
\( \Delta H_m \) (0.26), followed by \( \Delta S_m \) (\( R^2 \) = 0.09) and \( T_m \) (\( R^2 \) = 0.09).\(^1\) This
investigation of the residuals support the results obtained
with the GSE which indicates that solid-state properties do
influence the solubility of drugs.

With the purpose of keeping the contribution of CLOGP
(from Fig. 3) constant while studying the additional effects of
\( \Delta H_m \) and \( \Delta S_m \) on intrinsic solubility, the following two-step
analysis was carried out. First, a PLS analysis was performed
on the residuals (Res\( S_{CLOGP} \)) with \( \Delta H_m \) and \( \Delta S_m \) as variables.\(^2\)
This analysis resulted in a one-latent-variable model with
\( R^2 \) = 0.20 and \( Q^2 \) = 0.16. Although this was a weak PLS model, it
described the relationship of the combined solid-state prop-
erties, \( \Delta H_m \) and \( \Delta S_m \), with the residuals (analogous to the
analysis of the residuals in Fig. 5). The analysis resulted in the
following equation for the predicted residuals (Res\( S_{CLOGP-P} \)):

\[
\text{Res}_{CLOGP-P} = -0.034 \Delta H_m - 0.09 \Delta S_m + 1.18
\]

Second, the Res\( S_{CLOGP-P} \) was added to the CLOGP-predicted sol-
ubility (log\( S_{CLOGP} \)) to create a predicted value of intrinsic solu-

---

\(^1\) The correlation with CLOGP for the respective variables are: \( T_m \)
\( (R^2 = 0.24), \Delta H_m \) \( (R^2 = 0.0001) \) and \( \Delta S_m \) \( (R^2 = 0.12) \).
\(^2\) The properties \( \Delta H_m \) and \( \Delta S_m \) display an internal correlation
with \( R^2 = 0.67 \) for the 26 compounds of this data set.
ine, trimethoprim, phenacetin and diethylstilbestrol, however, were all exceptions; the reliability of the solubility prediction was greatly reduced by the inclusion of $\Delta H_m$ and $\Delta S_m$, despite the relatively high values of all three solid-state properties for these drugs.

### 3.4 Calculated versus experimental $\Delta S_m$

We also calculated values of $\Delta S_m$ for these drugs according to the method of Dannenfelser and Yalkowsky (Eq. (2)) (Dannenfelser and Yalkowsky, 1996, 1999), and correlated the calculated values with the experimental ones (Fig. 7). This equation does not give reliable predictions of $\Delta S_m$ as shown by the $R^2$ of 0.48. All compounds except chloropropamide, glyburide, probenecid, piroxicam and diethylstilbestrol were underestimated by, on average, 15 moles K$^{-1}$. Experimental and predicted values of $\Delta S_m$ for all compounds are shown in Table S1, in the supporting information.

### 4. Discussion

Analysis of our dataset revealed that, in general, the intrinsic solubility is clearly related to the stability of the solid-state of a drug ($T_m$, $\Delta H_m$ and $\Delta S_m$), in particular for drugs with large deviations from the general $\log S_0 - \log P$ relationship. The GSE developed by Yalkowsky and co-workers relies on the assumption that $\Delta S_m$ is a constant (56.5 moles K$^{-1}$). In reality, $\Delta S_m$ of drug molecules differs considerably from this value. In our set of structurally diverse drugs, the values obtained ranged from 56.8 moles K$^{-1}$ (diazoxide) to 114.1 moles K$^{-1}$ (astemizole), with a median of 81.5 moles K$^{-1}$ (Table 1). Hence, all compounds had a $\Delta S_m$ higher than the constant used in GSE. The GSE over-predicted $S_0$ for this set of compounds, an effect that was reduced by using the experimental values of the $\Delta S_m$.

Recently, a set of 21 compounds introduced by Ran et al. has been used by several authors as a test set for predictions...
of intrinsic solubility (Delaney, 2004; Hou et al., 2004; Ran et al., 2001). This data set consists of only 10 drugs; the remaining compounds are herbicides or pesticides, of which some are liquids at room temperature. Thus, this data set is far from ideal as a reference data set for prediction of drug solubility. We propose that the 26 crystalline drugs presented in this work are better suited for this purpose. Our proposal is supported by the fact that the data set covers a relevant solubility range for drug discovery and that the compounds were selected to be diverse with regards to chemical structure, solid-state properties and the oral drug space.

Of main interest to us is to find a purely computational model that is superior to the currently available models in terms of robustness. There are at least two issues that need to be addressed in order to reach this goal. First, the data set should be expanded to include more crystalline drugs with low solubility (nM to μM range). Second, we probably need to include descriptors of the solid-state in order to really see the difference in model robustness. All the purely computational models referred to in this paper are based on single molecule descriptors. Although it is a daunting task to try to model crystals of drugs, there has recently been some encouraging data reported on this topic. Perlovich et al. (2004a,b) showed that it is possible to calculate the lattice energy, with satisfactory results for known crystal structures (ibuprofen and naproxen) using force fields developed for this particular purpose. The experimental data needed to arrive at the lattice energy is, in this case, the sublimation enthalpy (ΔH_{sub}), which is a more precise measure of the cohesive forces in the crystal than is ΔH_m. It would be highly desirable to have ΔH_{sub} values for a large number of drugs in order to parameterise models for this property. As far as we know, models for the prediction of ΔH_{sub} have been published for organic molecules, but not for drug-like molecules (Charlton et al., 1995). However, there have recently been a few attempts to predict T_m on the basis of descriptors derived for an isolated molecule (Bergström et al., 2003; Karthikeyan et al., 2005). The overall performance was fairly good, but it is clear that the difficulty in predicting properties of the solid-state inevitably makes such models unreliable if precise data are needed for a particular compound. In the case of ΔS_m, we evaluated the Dannenfelser equation against our experimental data. We found that it does not provide predictions that are reliable enough to replace ΔS_m in future solubility models.

In the absence of experimental logP data, we used one of many versions of calculated logP, CLOGP, throughout this paper. Since this is an estimate of logP, we can only speculate on the contribution from this term to the error in our solubility predictions. Thus, there is also room for improvement with regard to the description of the solvation of drug molecules in order to achieve the desired accuracy for future computational tools for solubility prediction.
The drug discovery process would benefit considerably from the development of rules or descriptors that could single out the molecules that are limited in solubility by the stability of their crystalline state from the ones that are limited in solubility by lipophilicity. In an attempt to achieve this, we constructed classification models (PCA) for the top ten improved compounds and the 10 most deteriorated compounds, using 2D molecular descriptors (data not shown). However, no reliable models were developed, mainly due to the small size and high structural diversity of the two classes. This approach will be explored further in our laboratory with an increased number of experimental data points.

In conclusion, we have demonstrated the importance of the contribution of the solid-state properties of a drug to its solubility for 26 structurally diverse drugs that cover the oral drug space. The intrinsic solubility was accurately measured for these drugs and we propose that this data set can be used as an external test set for validation of drug solubility models. Furthermore, the well-established GSE and the Dannenfelser equation were evaluated for the first time on a data set comprised only of drug molecules and a modification of the GSE using the measured ΔS_m values was explored. Finally, for future improved accuracy and robustness of solubility models, we suggest that considerable effort be put into the development of an improved theoretical description of the crystalline state and the energy of solvation of drug molecules.

Acknowledgements

This work was supported by AstraZeneca R&D Mölndal, Sweden; the Swedish Research Council Grant no. 9478; the Swedish Fund for Research without Animal Experiments and the Knut and Alice Wallenberg Foundation. Generous travelling grants were awarded by the Swedish Pharmaceutical Society and the Swedish Chemical Society. We thankfully acknowledge Dr. Lennart Lindfors (AstraZeneca R&D Mölndal) for stimulating scientific discussions as well as Karin Sundström, Nina Ginman and Rieke Draheim for experimental assistance.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ejps.2006.05.013.

REFERENCES


