In silico calculation of aqueous solubility

Dr John Mitchell
University of St Andrews
Background ...
Solubility Measurement ...
Diclofenac
- First precipitation – Kinetic Solubility (Not in Equilibrium)
- Thermodynamic Solubility through “Chasing Equilibrium” - Intrinsic Solubility (In Equilibrium)

Supersaturation Factor
SSF = $S_{kin} - S_0$

“CheqSol”
- We continue “Chasing equilibrium” until a specified number of crossing points have been reached
- A crossing point represents the moment when the solution switches from a saturated solution to a subsaturated solution; no change in pH, gradient zero, no re-dissolving nor precipitating. . .
  SOLUTION IS IN EQUILIBRIUM

Modelling and Predicting Solubility …
How should we approach the prediction/estimation/calculation of the aqueous solubility of druglike molecules?

Two (apparently) fundamentally different approaches
The Two Faces of Computational Chemistry

Informatics

Theoretical Chemistry
Informatics

“The problem is too difficult to solve using physics and chemistry, so we will design a black box to link structure and solubility”
Informatics and Empirical Models

• In general, Informatics methods represent phenomena mathematically, but not in a physics-based way.
• Inputs and output model are based on an empirically parameterised equation or more elaborate mathematical model.
• Do not attempt to simulate reality.
• Usually High Throughput.
Theoretical Chemistry

“The problem is difficult, but by making suitable approximations we can solve it at reasonable cost based on our understanding of physics and chemistry”
Theoretical Chemistry

• Calculations and simulations based on real physics.
• Calculations are either quantum mechanical or use parameters derived from quantum mechanics.
• Attempt to model or simulate reality.
• Usually Low Throughput.
Our Methods …

(1) Random Forest (informatics)
Our Random Forest Model …

We want to construct a model that will predict solubility for druglike molecules …

We don’t expect our model either to use real physics and chemistry or to be easily interpretable …

We do expect it to be fast and reasonably accurate …
Random Forest

Machine Learning Method

- Looks sort of soluble.
- Looks soluble to me...
- This guy is soluble!
- I know it’s soluble!
- I say insoluble!
- As soluble as can be!
- I guess it’s insoluble.
Random Forest for Solubility Prediction

A Forest of **Regression** Trees

- Dataset is partitioned into consecutively smaller subsets (of similar solubility)
- Each partition is based upon the value of one descriptor
- The descriptor used at each split is selected so as to minimise the MSE

Random Forest for Predicting Solubility

- A Forest of Regression Trees
- Each tree grown until terminal nodes contain specified number of molecules
- No need to prune back
- High predictive accuracy
- Includes descriptor selection
- No training problems – largely immune from overfitting
- “Out-of-bag” validation – using those molecules not in the bootstrap samples.
Dataset

Literature Data
• Compiled from Huuskonen dataset and AquaSol database – pharmaceutically relevant molecules
• All molecules solid at room temperature
• n = 988 molecules
• Training = 658 molecules
• Test = 330 molecules
• MOE descriptors 2D/3D

• Aqueous solubility – the thermodynamic solubility in unbuffered water (at 25°C)
Dataset

Literature Data
• Compiled from Huuskonen dataset and AquaSol database – pharmaceutically relevant molecules
• All molecules solid at room temperature
• \( n = 988 \) molecules
• Training = 658 molecules
• Test = 330 molecules
• MOE descriptors 2D/3D

Datasets compiled from diverse literature data may have significant random and systematic errors.
Random Forest: Solubility Results

RMSE(tr)=0.27
r^2(tr)=0.98
Bias(tr)=0.005

RMSE(oob)=0.68
r^2(oob)=0.90
Bias(oob)=0.01

RMSE(te)=0.69
r^2(te)=0.89
Bias(te)=-0.04

These results are competitive with any other informatics or QSPR solubility prediction method.

RMSE(tr)=0.27  
$r^2$(tr)=0.98  
Bias(tr)=0.005

RMSE(oob)=0.68  
$r^2$(oob)=0.90  
Bias(oob)=0.01

RMSE(te)=0.69  
$r^2$(te)=0.89  
Bias(te)=-0.04

Random Forest Models To Predict Aqueous Solubility

David S. Palmer, Noel M. O’Boyle,† Robert C. Glen, and John B. O. Mitchell*
Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge,
Lensfield Road, Cambridge CB2 1EW, United Kingdom


Why Are Some Properties More Difficult To Predict than Others? A Study of QSAR Models of Solubility, Melting Point, and Log P

Laura D. Hughes,† David S. Palmer, Florian Nigsch, and John B. O. Mitchell*
Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge,
Lensfield Road, Cambridge CB2 1EW, United Kingdom

Our Methods …

(2) Thermodynamic Cycle (A hybrid of theoretical chemistry & informatics)
Our Thermodynamic Cycle method …

We want to construct a theoretical model that will predict solubility for druglike molecules …

We expect our model to use real physics and chemistry and to give some insight …

We may need to include some empirical parameters…

We don’t expect it to be fast by informatics or QSPR standards, but it should be reasonably accurate …
For this study Toni Llinàs measured 30 solubilities using the CheqSol method and took another 30 from other high quality studies (Bergstrom & Rytting).

We use a Sirius glpKa instrument
Our goal is to ask ...

\[ \Delta G_{\text{sol}} = \Delta G_{\text{sub}} + \Delta G_{\text{solv}} + \Delta G_{\text{tr}} = -RT \ln S_o V_m \]
Can we use theoretical chemistry to calculate solubility via a thermodynamic cycle?

\[ \Delta G_{\text{sol}} = \Delta G_{\text{sub}} + \Delta G_{\text{solv}} + \Delta G_{\text{tr}} = -RT \ln S_0 V_m \]
$\Delta G_{\text{sub}}$ from lattice energy & an entropy term (DMAREL based on B3LYP/6-31G*)

$\Delta G_{\text{solv}}$ from a semi-empirical solvation model (SCRF B3LYP/6-31G* in Jaguar)

$\Delta G_{\text{tr}}$ from ClogP

(i.e., different kinds of theoretical/computational methods, albeit with consistent functional and basis set)
\[ \Delta G_{\text{sol}} = \Delta G_{\text{sub}} + \Delta G_{\text{solv}} + \Delta G_{\text{tr}} = -RT \ln S_o V_m \]
$\Delta G_{\text{sub}}$ comes mostly from lattice energy minimisation based on the experimental crystal structure.
\[ \Delta G_{\text{sol}} = \Delta G_{\text{sub}} + \Delta G_{\text{solv}} + \Delta G_{\text{tr}} = -RT \ln S_0 V_m \]
\( \Delta G_{\text{solv}} \) comes from a semi-empirical solvation model (SCRF B3LYP/6-31G* in Jaguar)

\[
\Delta G_{\text{sol}} = \Delta G_{\text{sub}} + \Delta G_{\text{solv}} + \Delta G_{\text{tr}} = -RT \ln S_0 V_m
\]
$\Delta G_{\text{solv}}$ comes from a semi-empirical solvation model (SCRF B3LYP/6-31G* in Jaguar)

This is likely to be the least accurate term in our equation.

We also tried SM5.4 with AM1 & PM3 in Spartan, with similar results.
\[ \Delta G_{\text{tr}} \text{ comes from ClogP} \]

\[ \Delta G_{\text{sol}} = \Delta G_{\text{sub}} + \Delta G_{\text{solv}} + \Delta G_{\text{tr}} = -RT \ln S_0 V_m \]
ΔG_{tr} comes from ClogP

ClogP is a fragment-based (informatics) method of estimating the octanol-water partition coefficient.
\[ \Delta G_{\text{sol}} = \Delta G_{\text{sub}} + \Delta G_{\text{solv}} + \Delta G_{\text{tr}} = -RT \ln S_o V_m \]
What Error is Acceptable?

• For typically diverse sets of druglike molecules, a “good” QSPR will have an RMSE $\approx 0.7 \log S$ units.

• An RMSE $> 1.0 \log S$ unit is probably unacceptable.

• This corresponds to an error range of 4.0 to 5.7 kJ/mol in $\Delta G_{\text{sol}}$. 
What Error is Acceptable?

• A useless model would have an RMSE close to the SD of the test set logS values: ~ 1.4 logS units;

• The best possible model would have an RMSE close to the SD resulting from the experimental error in the underlying data: ~ 0.5 logS units?
Results from Theoretical Calculations

- Direct calculation was a nice idea, but didn’t quite work – errors larger than QSPR

- “Why not add a correction factor to account for the difference between the theoretical methods?”

- This was originally intended to calibrate the different theoretical approaches, but

...
Within a week this had become a hybrid method, essentially a QSPR with the theoretical energies as descriptors.

\[ \log S = -0.0308(\pm 0.0050)\Delta G^*(sub) - 0.126(\pm 0.011)\Delta G^*(tr) + 3.31(\pm 1.18)b_{rotR} \]
Results from Hybrid Model

\[ \log S = -0.0308(\pm 0.0050)\Delta G^*_{\text{sub}} - 0.126(\pm 0.011)\Delta G^*_{\text{tr}} + 3.31(\pm 1.18)b_{\_\text{rotR}} \]
We find that $\Delta G_{solv}$ (gas to octanol) is poorly correlated with logS and fails to appear in the regression equation.

$B_{rotR}$ is the proportion of bonds that are rotatable and describes the propensity of flexible molecules to be more soluble.

We can also write an almost equivalent equation in the form …
\[
\log S = 0.0266(\pm 0.0044)U_{\text{latt}} - 0.696(\pm 0.063)\text{ClogP} + 3.31(\pm 1.20)\text{b_rotR} + 1.39(\pm 0.55)
\]

This regression equation gives \( r^2 = 0.77 \) and \( \text{RMSE} = 0.71 \)
How Well Did We Do?

• For a training-test split of 34:26, we obtain an RMSE of 0.71 logS units for the test set.
• This is comparable with the performance of “pure” QSPR models.
• This corresponds to an error of about 4.0 kJ/mol in $\Delta G_{\text{sol}}$. 
The main factors influencing the solvation of a crystalline compound. The + symbol refers to factors that favour the movement of the compound from the crystal lattice (bottom) into the solvent (top), the – symbol refers to factors that favour partition of compound into the lattice.
Solubility by TD Cycle: Conclusions

● We have a hybrid part-theoretical, part-empirical method.

● An interesting idea, but relatively low throughput - and an experimental (or possibly predicted?) crystal structure is needed.

● Similarly accurate to pure QSPR for a druglike set.

● Instructive to compare with literature of theoretical solubility studies.
Predicting Intrinsic Aqueous Solubility by a Thermodynamic Cycle

David S. Palmer,† Antonio Llinàs,† Iñaki Morao,‡ Graeme M. Day,§
Jonathan M. Goodman,† Robert C. Glen,† and John B. O. Mitchell*†

The Pfizer Institute for Pharmaceutical Materials Science and Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom, and Pfizer Global Research and Development, Sandwich Laboratories, Sandwich, Kent CT13 9NJ, United Kingdom

MOLECULAR PHARMACEUTICS VOL. 5, NO. 2, 266–279 2008
What Limits Us?
Does Experimental Error Limit Model Accuracy?

• We can get an RMSE of about 0.7 logS units from our models;

• We *estimate* the experimental error of literature data at about 0.5 logS units;

• Can we get much better models by using CheqSol data, where we think the errors are significantly smaller?
Does Experimental Error Limit Model Accuracy?

• We have asserted that we expect CheqSol to give “more accurate” solubilities than literature mining;

• We know what we are measuring (intrinsic solubility), we are using the same technique throughout, and multiple repetitions show that the random error is less than 0.05 logS units;

• Building a model based on our CheqSol data should give better results than building one from literature data for the same molecules.
Test to Confirm Superiority of CheqSol Data

• Use 85 molecules for which we have both Literature and CheqSol data

• Build Random Forest models based on 20 different 60:25 training:test splits for each data source
Figure 1  Correlation diagram for solubility data taken from the literature and as measured by the CheqSol method. $r^2=0.614$, RMSE=0.730 and bias=0.155.
An inconvenient truth: Models built on CheqSol data are no more accurate than those built on collated literature data.
J. Chem. Inf. Model’s

“Solubility Challenge” ...
Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements?

Antonio Llinàs,* Robert C. Glen, and Jonathan M. Goodman*
Given accurately measured solubilities of 100 molecules, can you predict the solubilities of 32 similar ones?
For this study, Toni Llinàs measured 132 solubilities using the CheqSol method.

We use a Sirius glpKa instrument.
Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements?

Antonio Llinás,* Robert C. Glen, and Jonathan M. Goodman*

<table>
<thead>
<tr>
<th>Structure</th>
<th>Name</th>
<th>MW Neutral Form</th>
<th>pKa</th>
<th>Kinetic Solubility µM</th>
<th>Intrinsinc Solubility µM</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Structure" /></td>
<td>Acetaminophen</td>
<td>151.17</td>
<td>9.52 ± 0.01</td>
<td>161700 ± 7000 µM</td>
<td>86300 ± 7000 µM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24400 ± 1060 µg/ml</td>
<td>13000 ± 1060 µg/ml</td>
</tr>
<tr>
<td><img src="image2" alt="Structure" /></td>
<td>Acetazolamide</td>
<td>222.25</td>
<td>8.75 ± 0.02</td>
<td>6100 ± 3840 µM</td>
<td>3670 ± 80 µM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.31 ± 0.04</td>
<td>1360 ± 850 µg/ml</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>816 ± 18 µg/ml</td>
<td></td>
</tr>
<tr>
<td><img src="image3" alt="Structure" /></td>
<td>Alprenolol</td>
<td>249.36</td>
<td>9.47 ± 0.01</td>
<td>5080 ± 50 µM</td>
<td>2320 ± 40 µM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1266 ± 12 µg/ml</td>
<td>580 ± 10 µg/ml</td>
</tr>
<tr>
<td><img src="image4" alt="Structure" /></td>
<td>Amantadine</td>
<td>151.25</td>
<td>10.48 ± 0.01</td>
<td>17300 ± 3960 µM</td>
<td>14000 ± 1180 µM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2620 ± 600 µg/ml</td>
<td>2120 ± 180 µg/ml</td>
</tr>
</tbody>
</table>
LETTER

Findings of the Challenge To Predict Aqueous Solubility

Anton J. Hopfinger,*,†,‡ Emilio Xavier Esposito,‡
A. Llinàs,‖§ R. C. Glen,§ and J. M. Goodman§
Caveat: the official results are used in the following slides, but most of the interpretation is my own.
A prediction was considered correct if it was within 0.5 log units.
Not a very generous margin of error!
The 99 entries obtained between 5/32 and 20/32 correct results.
A “null prediction” based on predicting everything to have the mean training set solubility would have got 9/32 correct.
Using an $R^2$ threshold of 0.500, only 18/99 entries were good.
Taking both correct predictions and $R^2$ together ...
... there were 3 Pareto optimal entries which I think of as “winners”, combining $R^2$ with correct predictions.
Percentage of entrants to correctly predicted logS

Compounds ordered from smallest logS to largest logS
Some molecules proved much harder to predict than others – the most insoluble were amongst the most difficult.
Conclusions from Solubility Challenge

- My opinion is that the overall standard was rather poor;

- It’s obvious that some entries were much better than others;

- But entries were anonymous;

- So we can’t judge between either specific researchers or between their methods;

- We can only rely on the diplomatic “official” summary …
Conclusions from Solubility Challenge

• We can only rely on the diplomatic “official” summary …

… “a variety of methods and combinations of methods all perform about equally well.”
Overall Conclusions ...
The state of the art is that …

… solubility has proved a difficult property to calculate.

It involves different phases (solid & solution) and different substances (solute and solvent), and both enthalpy & entropy are important.

The theoretical approaches are generally based around thermodynamic cycles and involve some empirical element.
The state of the art is that ...
The state of the art is that …

… there has been solid progress but no solution.
Thanks

- Pfizer & PIPMS
- Dr Dave Palmer
- Pfizer (Dr Iñaki Morao, Dr Nick Terrett & Dr Hua Gao)

- Gates Cambridge Trust
- Laura Hughes

- Unilever
- Dr Florian Nigsch