DMSO Solubility Prediction

P. Japertas[1], P. Maas[2] and A. Petrauskas[1]

[1] Pharma Algorithms
www.ap-algorithms.com info@ap-algorithms.com

[2] specs

INTRODUCTION

• Identification of compounds that are insoluble in DMSO is of great practical importance in global drug discovery, compound management, and laboratory automation.

• These compounds cannot be stored in DMSO, and their sampling is problematic.

• A joint effort of scientists from Specs [1] and Pharma Algorithms [2] resulted in developing a specialized software system (AB/DMSO) that identifies such compounds *prior* to their synthesis [3].

Experimental Data

• The analysis involved over 22,000 compounds from the Specs library.

• Only compounds of the top purity (> 90%) were considered.

• Measurement involved the following procedures:

a) A compound in DMSO is put on a shaker for 15 min.

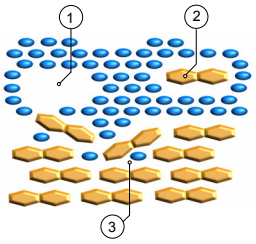
b) Visual check was performed. If not fully dissolved, the tube is placed in an ultrasonic bath at 35ºC for another 15 min.

c) If, after this, there is still solid substance left, the compound is called “insoluble” and marked as such.

• The DMSO solubility data was represented in a binary format, using a cut-off value at 20 mM (a compound is "insoluble" if its solubility is < 20 mM).

Theoretical Model

DMSO solubility involves several thermodynamic stages: (1) cavity creation, (2) solvation and (3) crystal disruption. All stages must be considered in a successful computational model.



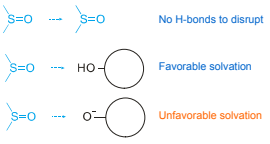
1) Cavity Creation

Cavity creation can be neglected, as DMSO has no internal H-bond network to disrupt, and thus energetically outcome is very small.

2) Solute Solvation

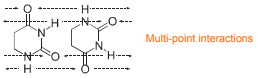
• Solvation of solute molecules should also not create any problems in most cases, as DMSO molecules are highly polar, and they form strong H-bonds with H-donors.

• But DMSO is not a H-donor by itself, so it cannot stabilize any negatively charged species. In other words, solvation of large anions and zwitterions may be problematic (compared to water).



3) Crystal Disruption

DMSO can easily disrupt any crystal structures involving single-point interactions (dipolar or H-bonding). But it may not easily disrupt “tight” crystal structures involving multi-point interactions. (For example, some purine derivatives with log *P* < 0 are poorly soluble in DMSO and water).



Factors Limiting DMSO Solubility

To summarize, we must consider the following two factors that limit solubility in DMSO:

a) poor solvation of anions and zwitterions, and

b) tight crystal packing of multi-polar skeletons.

Both of these factors must be accounted for in the successful informatic approach.

Informatic Approach

• The experimental data was represented in a binary format (S = 0 if insoluble, S = 1 if soluble).

• Binary data was analyzed using a hybrid method using logistic (binary) regression and PLS procedures.

• This method is well-suited for analysis of binary data that depends on multiple unknown factors.

• Four types of descriptors were used:

Binary Data	Log [S/(1-S)] =
Favorable Solvation	Σ Atom Increments + Σ Cationic Groups +
Unfavorable Solvation	Σ Anionic Groups + Σ Atom Chains +
Crystal Disruption	Σ Ring Scaffolds

Atom Increments

Atom increments model various additive properties that affect solvation of non-charged species. Most of these properties are used in various solubility equations (Yalkovsy's, Abraham's, and others). They can also be compared to various types of topological or e-state indices.

Mol. Size
Polarizability
Polar Surface
H-bonding

Charged Groups

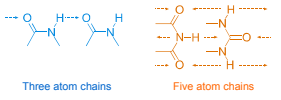
Charged groups involved counts of ionized fragments. Positive charges showed large positive increments, whereas negative groups showed large negative increments (as expected from solvation model).

Large positive increments: >N⁺<, =N⁺<, -O⁺<, etc.

Large negative increments: CO_2^-, SO_3^-, PO_3^-, etc.

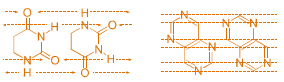
Atom Chains

“Atom chains” model multipoint interactions of non-cyclic skeletons. Chains of three atoms can model two-point interactions (first example), whereas chains of five atoms can model three-point interactions (second example).



Ring Scaffolds

“Ring scaffolds” model multipoint interactions in cyclic skeletons. First example – multiple H-bonds, second example – multiple dipoles.



Novelty of Approach

• The proposed approach can automatically identify both non-specific (solvation) and specific (crystallization) effects thanks to a combination of small and large structural descriptors.

• It provides a unique possibility of balancing between the accuracy and generality (reliability) of predictions by using atom chains of variable length.

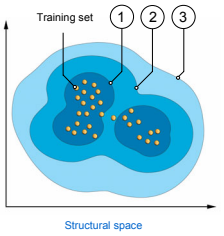
Accuracy vs. Generality

The longer atom chains are used, the more specific structural effects can be captured. On the other hand, this also leads to the decreased generality of considerations, as we cannot capture specific crystallization effects of compounds that are not similar to the training set.

Therefore, three different algorithms were obtained:

1) The high-reliability algorithm employs five-atom chains (along with other descriptors). It is used for compounds that are similar to the training set.

2) The moderate-reliability algorithm employs three-atom chains (along with other descriptors). It is used for compounds that are not very similar to the training set.



3) The low-reliability algorithm does not use atom chains and ring scaffolds. It only uses increments of simple atoms and charged groups that do not model crystallization effects. This algorithm is only suitable for making very crude predictions, and it can be used for compounds that are highly dissimilar from the training set.

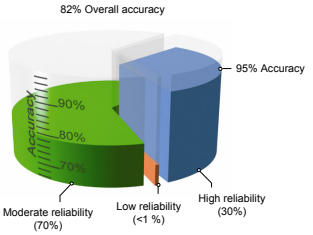
Validation Results

• Validation set involved 30% (N = 6,679) of compounds that were not used in development.

• Using this data set, the overall accuracy of predictions was 82%.

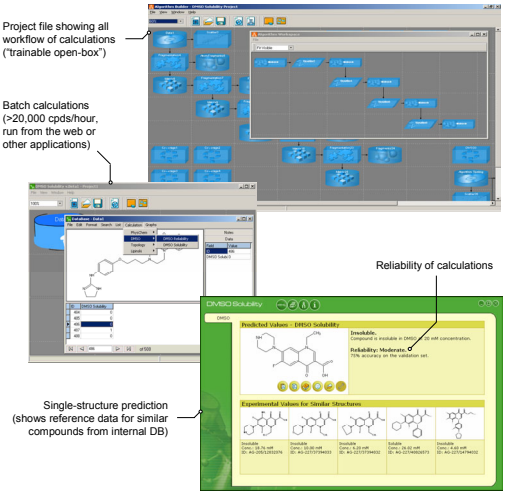
• The high-reliability predictions produced 95% accuracy. They were shown to be applicable for *c.a.* 30% of compounds from the World Drug Index (WDI).

• The moderate-reliability predictions produced 75% accuracy. They were shown to be applicable for *c.a.* 70% of compounds from WDI.



Product Interface

Three types of interfaces for DMSO solubility predictions are available:



All calculations are provided with reliability estimations. Compounds that are similar to the training set are processed by the most accurate algorithm, whereas compounds that are dissimilar are processed by the least accurate (but most general) algorithm.

References

[1] Specs, Delftechpark 30, 2628 XH Delft, The Netherlands, www.specs.net

[2] Pharma Algorithms, A. Mickeviciaus g. 29, Vilnius, Lithuania, www.ap-algorithms.com

[3] Evaluation version is available from: www.ap-algorithms.com/dmso.html