

Modeling Aqueous Solubility

Darko Butina* and Joelle M. R. Gola

Computational Chemistry and Chemoinformatics, ArQule (UK) Limited, Science Park, Cambridge, U.K.

Received October 23, 2002

This paper describes the development of an aqueous solubility model based on solubility data from the Syracuse database, calculated octanol–water partition coefficient, and 51 2D molecular descriptors. Two different statistical packages, SIMCA and Cubist, were used and the results were compared. The Cubist model, which comprises a collection of rules, each of which has an associated Multiple Linear Regression model (MLR), gave better overall results on a test set of 640 compounds with an overall squared correlation coefficient of 0.74 and an absolute average error of 0.68 log units. Both training and independent test sets had similar distributions of structures in terms of the different functionalities present—60% neutral, 14% acidic, 8% phenolic, 11% monobasic, 4% polybasic, and 3% zwitterionic molecules. Sets were designed by random selection, with 2688 (81%) and 640 (19%) molecules, respectively, forming the training and the test sets.

INTRODUCTION

Of all the molecular properties which can profoundly affect a compound's biological activity, aqueous solubility is probably one of the most fundamental and deserves attention in the early phases of drug discovery. Not surprisingly, therefore, aqueous solubility has been extensively studied, and a large number of computational methods for the estimation of this highly important property have been reported.¹

Predictive models for aqueous solubility are generally based on a diverse set of descriptors such as experimentally based descriptors, molecular properties, and collection of relevant structural features, that are correlated to activity by means of various statistical techniques including MLR and neural networks.¹ Based on experimentally derived descriptors, Jain and Yalkowsky² reported a general solubility equation requiring the water–octanol partition coefficient, logP, and melting point. The authors tested the proposed equation on 580 molecules that are not ionizable in the pH-range 2–13. This model gave an average absolute error (AAE) of 0.42 log units. Quantitative structure property relationships (QSPR) based on molecular properties have also been proposed. Huuskonen³ published a QSPR for solubility prediction based on 30 topological descriptors. The author developed a multilinear regression model that produced a squared correlation coefficient of 0.89 and a standard error (SE) of 0.67 log unit for 884 training set compounds plus $R^2 = 0.88$ and $SE = 0.71$ log units for 413 test set compounds. Huuskonen also developed a 30–12–1 artificial neural network that gave an impressive $R^2 = 0.94$ and $SE = 0.47$ log unit for the training set plus $R^2 = 0.92$ and $SE = 0.60$ log units for the test set.³ Similarly, Liu and So⁴ developed a QSPR based on seven 1D and 2D descriptors and an artificial neural network. The 7:2:1 neural network model was able to predict the solubility of 1312 compounds

with an overall correlation coefficient of 0.92 and a standard deviation of 0.72 log unit. The prediction results for a test set of 258 compounds were essentially the same as those in the training set, with $R^2 = 0.93$ and $SE = 0.71$ log units.⁴ Klopman et al.⁵ proposed two models using a group contribution approach that focused on pharmaceutical drugs, the first consisting of 45 fragments and one constant (based on 496 compounds) and the second consisting of 33 fragments (based on 483 compounds). A test set of 21 compounds was applied to the models. The authors found that the first model could be applied to 13 out of 21 compounds with a standard deviation of 0.58, and the second model could be applied to 19 out of the 21 test compounds with a standard deviation of 0.86.⁵ Klopman and Zhu⁶ have recently improved their aqueous solubility approach by using a larger number of compounds and a set of 118 structural descriptors. Using a similar approach, Kuhne et al.⁷ proposed a logS correlation, with 55 structural fragments and two melting point terms derived from a heterogeneous set of 694 organic compounds, that performed with an AAE of 0.38 log units. These examples represent only a small selection from the range of predictive solubility models available in the literature. Yaffe et al. have recently published a more comprehensive list of predictive models for aqueous solubility.¹

Our objective was to develop a generic aqueous solubility model for our internal use and with the following features:

- Based on as many molecules as possible and in particular those with potentially ionizable atoms (which are usually present in drug molecules and which are very often predicted poorly)
- Solubility calculated based upon structure features exclusively
- Easy to integrate within the suite of predictive activity/ADMET models deployed in our internal drug discovery processes

Our strategy was to combine calculated logP (internal $clogP^8$) with a 2D descriptor set comprising various counts

*Corresponding author phone: +44(0)1462 670856; e-mail: darko_Butina@hotmail.com.

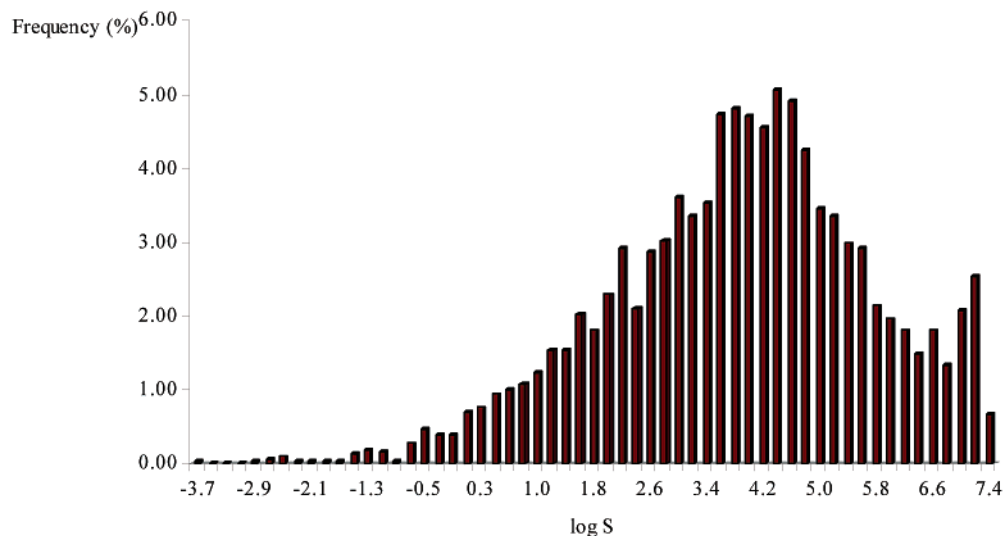


Figure 1. Distribution of logS values for the whole set.

including atom-based fragments (e.g. aromatic halogen), functional groups (e.g. sulfonamide), potential intramolecular hydrogen bonding moiety, and more general descriptors such as basic nitrogens, hydrogen bond donors, and hydrogen bond acceptors.

Two statistical approaches, Partial Least Squares (PLS)⁹ and Cubist,¹² were used in attempting to build the model. While PLS is a well established and widely used approach initially developed by Herman Wold⁹ and subsequently further enhanced and distributed by Svante Wold,¹⁴ the methodology underlying Cubist, as developed by Quinlan, is perhaps less widely used and thus deserves a few more details of explanation.

The underlying algorithm used to build the collection of the rules in Cubist is proprietary to Ross Quinlan and his company RuleQuest Research. In his 1992 paper,¹¹ Quinlan described the combination of a collection of rules and MLR to build "continuous" models, this probably prefiguring the development of what today is Cubist. The basic principle of Cubist in this context is to build a rule-based decision tree, where each rule has an associated MLR model describing the structure activity relationship (SAR) for all molecules belonging to compounds characterized by that rule. In other words, Cubist effectively classifies a set of compounds according to structural parameters and evaluates a separate SAR model for each subset, rather than fitting a single model to the entire set.

Our best results were obtained with Cubist, with R^2 values of 0.80 and 0.74 for the training and test sets, respectively (compared to values of 0.71 and 0.69 obtained with PLS), and smaller AAEs for both training and test sets (0.61 and 0.68 respectively).

DATA SET AND DESCRIPTORS

The experimental aqueous solubility data used was obtained from the Syracuse Research Corporation (SRC) database. Structures were utilized provided that the following criteria were all met:

1. Measured solubility data was available for the compound in the temperature range 20–30 °C.
2. The compound was organic, i.e., contained C, N, O, S, P, and halogens only.

3. The molecule was nonreactive, i.e., acid chlorides, anhydrides and epoxides were all excluded.

4. The structure's Smiles string was processed by Daylight based make_parents utility without error. This

a. Identifies the larger fragment of disconnected structures as the parent and, in the case of salts, neutralizes the remaining ionized fragment (e.g. COO^- would be converted to COOH and NH_3^+ to NH_2),

b. Functional groups that may give rise to alternative Smiles substrings are converted into a single variant, e.g. all nitro groups are written as N(=O)(=O) (rather than $[\text{N}^+]=\text{O}[\text{O}^-]$).

Additionally, each molecule was labeled according to membership of the following classes:

Monobasic: single basic center (aliphatic primary, secondary and tertiary amines plus amidines) (11% of all molecules)

Polybasic: more than one basic center (4%)

Acidic: any number of acidic groups (CO_2H , SO_2H , and PO_2H) (14%)

Phenolic: any number of phenolic hydroxyls (8%)

Zwitterionic: any number of acidic and basic groups in the same molecule (3%)

Neutral: none of the above (60%)

The structures were so labeled to facilitate analysis of the model's performance across the range of potentially charged or neutral molecules and, if necessary, the design and inclusion of additional descriptors that would better represent the charged functionalities.

The range of the whole set, in logS terms (where S is solubility in $\mu\text{mol/L}$), was from -3.70 to $+7.60$, with a mean of 3.77 and a standard deviation of 1.83 (Figure 1).

The descriptors used are a combination of the whole molecule property, *clogP*, and 140 descriptors such as counts of atomic type fragments (e.g. aromatic halogen) and counts of functional groups (e.g. sulfonamide) and such as more complex definitions describing potential internal hydrogen bonding and fuzzy descriptors such as basic nitrogen or HBond donors/acceptors.

DEVELOPMENT OF PLS AND CUBIST MODELS

The starting set containing 3328 molecules was randomly split into the training set (2688 (81%) compounds) and the

Table 1. Comparison of the Two Models

method	training set R^2	training set AAE	test set R^2	test set AAE
Cubist	0.80	0.61	0.74	0.68
PLS	0.71	0.75	0.69	0.77

independent test set (640 (19%) compounds). An automated procedure then calculated 140 descriptors and clogP from the list of smiles, using our proprietary software based on the Daylight programming toolkit,¹² to produce a comma separated input file in a standard input format for most commercial statistical or datamining software packages.

PLS Model. Starting with a set of 140 descriptors and clogP, PLS produced a four component model (where clogP and a subset of 52 descriptors contributed to the components) with $R^2 = 0.71$ ($Q^2 = 0.7$) and an AAE of 0.745 log units for the training set. The observed R^2 value for the independent test set was 0.69 with an AAE of 0.77 log units, the R^2 value being very similar to the estimated Q^2 from the training set.

Cubist Model. Starting with the same initial descriptor sets, Cubist¹² produced a model with $R^2 = 0.8$ and an AAE of 0.61 log units for the training set, plus $R^2 = 0.74$ with an AAE 0.68 for the independent test set. Here the model consisted of a combination of four different rules of the type:

Rule 1: [883 cases, mean 2.2, range -3.7 to 6.9, est err 0.72]
if

$$\text{clogP} > 2.34$$

then

$$\log S = 6.1 - \text{coeff}_1 * \text{desc}_1 + \text{coeff}_2 * \text{desc}_2 + \dots + \text{coeff}_n * \text{desc}_n$$

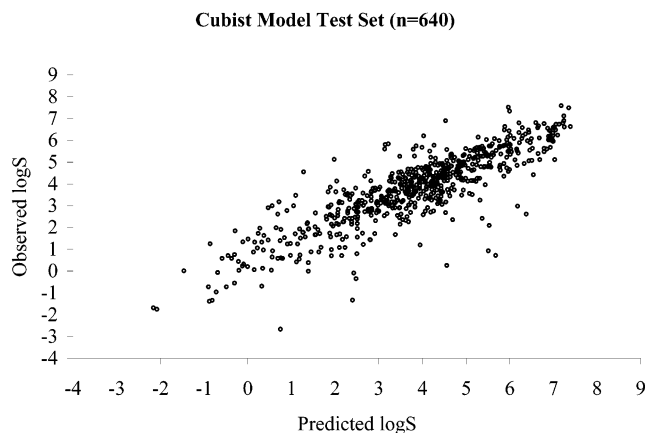
where the descriptors were clogP and subset of 51 descriptors from the initial set. The final PLS and Cubist model descriptor sets had clogP plus 28 descriptors (from the totals of 52 and 51) in common.

RESULTS AND DISCUSSION

Between the two approaches tried, Cubist outperformed PLS by 9% in the case of the training set and 5% in relation to the independent test set. In both cases, the AAE is also lower when using Cubist—by 14% for the training set and 9% for the test set (see Table 1 and Figure 2).

The performance of the models was evaluated across the predefined chemical classes, and, as can be seen from Figure 3, the Cubist model gives consistently better R^2 values and smaller AAEs for the training set than does PLS except in the case of zwitterions where R^2 is almost identical for the two methods.

While the performance of both models in respect of the independent test set as a whole is quite acceptable in terms of maintaining R^2 within 10% of that of the training set, the prediction of the solubility of acidic, polybasic, and zwitterionic molecules is less satisfactory. However, the influence of these groups was masked by a very good performance of the model in predicting solubility of monobasic molecules (Figure 4). Additionally, there is a reverse trend in both R^2

**Figure 2.** Scatter plot for the test set by Cubist model.

and AAE behavior for polybasic compounds and zwitterions in the test set.

However, since polybasic and zwitterionic molecules constitute less than 10% of the whole set, the model based on Cubist's methodology consistently outperforms that obtained from the PLS approach and thus is currently being used in house as our predictive model for aqueous solubility.

Since we do not have access to other solubility models, we have used part of a set of 21 molecules reported as an independent test set by other researchers, to compare different aqueous solubility models.^{2-4,6,7} The reason for choosing only 10 molecules from this set is that the remainders were already in our training set. Interestingly, our model gives statistical results similar to those of other known models for aqueous solubility prediction (see Table 2).

Our main objective was to build as general a model as possible, and we have not, at this stage, investigated further structural details relating to the exact nature and type of acids, polybasic molecules, and zwitterions that are present in the set. This issue will be addressed soon, and a probable outcome will be design and calculation of additional descriptors that better represent these three classes.

As mentioned earlier, the distribution of the predefined chemical classes is well preserved between the training and test sets used to build and evaluate the model (Table 3).

The most obvious explanation for the superior performance of Cubist over PLS is related to the difference in the underlying algorithms used by the two methods. Cubist tries to detect the presence of more than one cluster in the training set, identify rules describing that set, and then develop the best MLR for it, while PLS will attempt to build a single model for the whole set. It is important to emphasize here that there are facilities within the SIMCA package to perform hierarchical modeling, which would, in principle, be equivalent to the Cubist approach, i.e., identification of possible clusters in the data and construction of different PLS models for each cluster. However, so far we have not attempted this alternative approach, due mainly to the ease of using Cubist.

While we cannot disclose the full details of our model, we can report that, unsurprisingly, the calculated octanol-water partition coefficient clogP is a highly influential descriptor in the regression equation associated with each of the four rules.

Each of the four rules produced by Cubist applied to at least 630 training set molecules, and the corresponding MLR

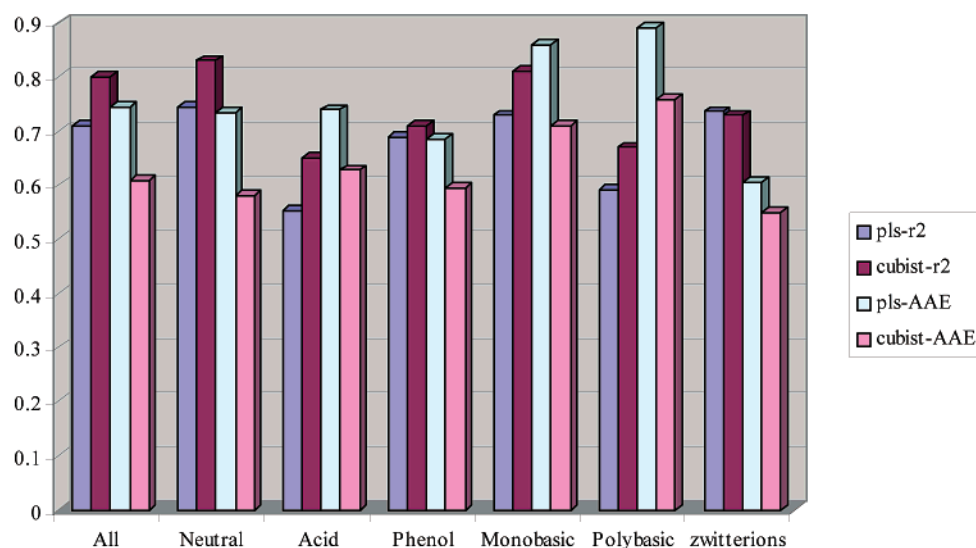


Figure 3. Histogram comparing the two models for the training set.

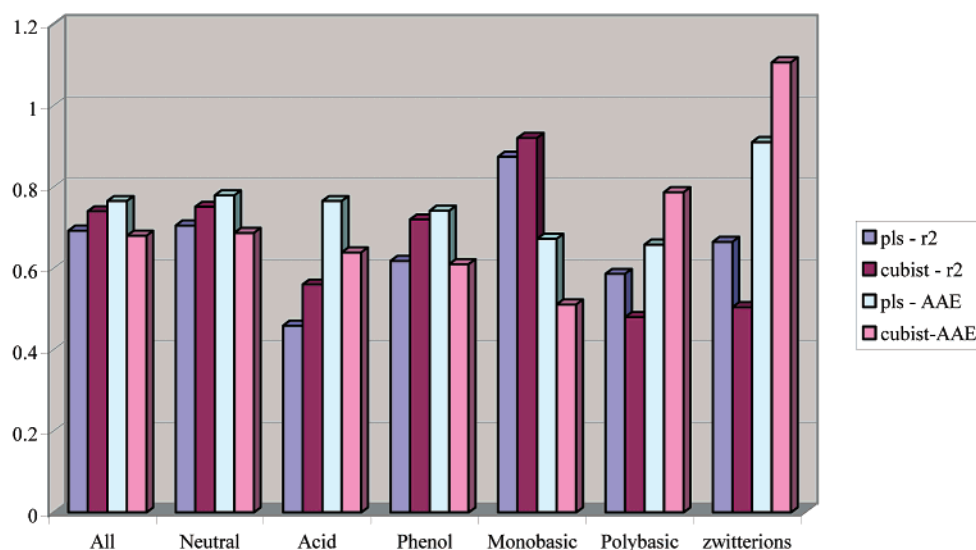


Figure 4. Histogram comparing the two models on the test set.

Table 2. Observed and Predicted Aqueous Solubility (log S in mol/L) for the Test Set of 11 Compounds

name	logS obs ^c	ClogS ^d	Jain ^e	Huuskonen ^f	Huuskonen ^g	Liu ^h	Klopman ⁱ	Kuhne ^j
2,2',4,5,5'-PCB	-7.89	-6.93	-6.87	-7.21	-7.40	-7.55	-7.90	-7.47
4,4'-DDT	-8.08	-6.08	-7.25	-7.67	-7.82	-7.93	-8.00	-7.75
antipyrine ^a	0.39	-2.04	-0.59	-1.29	-1.20	-1.41		-1.90
chlordane	-6.86	-5.40	-5.50	-7.29	-8.35	-7.32	-7.55	-6.51
chorpyriphos	-5.49	-4.68	-4.50	-5.61	-5.46	-4.50	-5.77	-3.75
diazinon	-3.64	-3.50	-3.75	-4.01	-4.10	-3.56	-5.29	-4.98
lindane	-4.64	-4.19	-4.10	-4.71	-5.34	-4.91	-4.88	-5.08
parathion	-4.66	-3.67	-3.33	-4.13	-3.98	-3.64	-3.94	-4.59
phenolphthalein	-2.90	-4.15	-4.52	-3.99	-4.05	-4.16	-4.48	-4.61
prostaglandin E2 ^b	-2.47	-2.63	-2.74	-3.29	-4.35	-3.80	-4.21	
<i>R</i> ²		0.89	0.89	0.96	0.89	0.89	0.84	0.82
SE		0.94	0.93	0.55	0.93	0.94	0.89	1.19

^a Outliers in Klopman's model, see ref 5. ^b Predicted values not given in ref 7. ^c See ref 3. ^d This work. ^e See ref 2. ^f Neural network, see ref 3. ^g Multilinear regression, see ref 3. ^h See ref 4. ⁱ See ref 5. ^j See ref 7.

equations were generated using a maximum of 36 descriptors. It is very important to monitor the ratio of cases (molecules used) per descriptor used when building MLR equations, and in our model, that ratio was well above the recommended 5:1 level for each rule, with the minimum ratio for any rule being at 19:1.

CONCLUSION

A model based on 2688 organic compounds, calculated octanol–water partition coefficients, and 51 fragment-based descriptors, deploying a methodology that combines a rule-based decision tree with MLR, has been described. The model relies on molecular structural features alone, and all

Table 3. Distribution of the Chemical Classes within the Training and Test Set

chemical class	training set (%)	test set (%)
neutral	60	63
monobasic	11	8
acids	14	12
polybasic	4	4
phenols	8	8
zwitterions	3	4

required descriptors are easily calculated in house by using automated Unix based scripts available via an intranet interface. The model is applicable to a variety of potentially charged functional groups and deals especially well with those molecules containing a single basic nitrogen center. Finally, this model gives statistical results similar to those of other known models for aqueous solubility prediction.

ACKNOWLEDGMENT

Our thanks go to Dr. Mike Tarbit for supporting this work, to Hazel Davies for assistance in gaining access to the Syracuse aqueous solubility database, and Simon Lister for proofreading the manuscript.

REFERENCES AND NOTES

- (1) Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. A Fuzzy ARTMAP Base on Quantitative Structure–Property Relationships

- (QSPRs) for predicting Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1177–1207.
- (2) Jain, N.; Yalkowsky, S. H. J. Estimation of the Aqueous Solubility I: Application to Organic Nonelectrolytes. *Pharm. Sci.* **2001**, *90*, 234–252.
 - (3) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773–777.
 - (4) Liu, R.; So, S.-S. Development of Quantitative-Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633–1639.
 - (5) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474–482.
 - (6) Klopman, G.; Zhu, H. J. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439–445.
 - (7) Kuhne, R.; Ebert, R.-U.; Kleint, F.; Schmidt, G.; Schuurmann, G. Group Contribution Methods to Estimate Water Solubility of Organic Chemicals. *Chemosphere* **1996**, *33*, 2129–2144.
 - (8) clogP is our proprietary calculator for octanol–water partition coefficient.
 - (9) Geladi, P.; Wold, Herman, The father of PLS. *Chemometrics Intelligent Laboratory Systems* **1992**, *15*, 1, R7–8.
 - (10) Simca-P 8.0, UMETRICS, www.umatrics.com.
 - (11) Quinlan, J. R. *Learning with Continuous Classes*; In Proc. AI '92, Adams, Sterling, Eds.; 1992; pp 343–348.
 - (12) For all details on the Daylight software and smiles see www.daylight.com.
 - (13) Cubist, RULEQUEST RESEARCH, www.rulequest.com.
 - (14) Wold, S.; Sjostrom, M.; Eriksson, L. Partial Least Squares Projections to Latent Structures (PLS) in Chemistry. *The Encyclopaedia of Computational Chemistry*; J. Wiley and Sons: 1999.
- CI020279Y