Evaluation of Structure Based Methods for the Prediction of LogP octanol for Agrochemicals

Pranas Japertas[a,b], Andrius Sazonovas[a,b], Eric D. Clarke[c], John S. Delaney[c]

[a] Pharma Algorithms, Inc., A.Mickeviciaus g. 29, LT-08117 Vilnius, Lithuania; [b] Chemistry Department, Vilnius University, Naugarduko g. 24, LT-03225 Vilnius, Lithuania; [c] Syngenta, Jealott's Hill International Research Centre, Bracknell, Berkshire, RG42 6EY, UK

Introduction

The organic-aqueous partition coefficient between octanol and water (*LogP*) is widely used as a measure of lipophilicity in the assessment of the movement between phases in physical or biological systems of new chemical entities. In the case of agrochemicals *LogP* serves a valuable role in the evaluation of their environmental fate. Wide establishment in the chemical industry as a useful and frequently considered property gave rise to a great number of available methods and software applications suitable for the prediction of octanol-water partition coefficient for new compounds. In this study we have compared six generally acknowledged structure based methods for the prediction of *LogP*:

- CLogP (Daylight v4.73)
- ALogP (Accelrys Diamond Descriptors v1.5)
- ACD LogP (Phys Chem Batch v6.16)
- Kowwin (EPI Suite v3.12)
- Absolv LogP (Pharma Algorithms ADME Boxes v3.5)
- AB/LogP (Pharma Algorithms ADME Boxes v3.5)

A comparison is also made with the Syngenta internal ELogP methodology which derives a consensus value from ACD LogP, CLogP and ALogP rather than using single method for the evaluation of *LogP*[1].

Finally a completely new algorithm developed by Pharma Algorithms is introduced – AB/LogP 2.0 based on 'Trainable Model' methodology (available in Pharma Algorithms ADME Boxes v4.0).

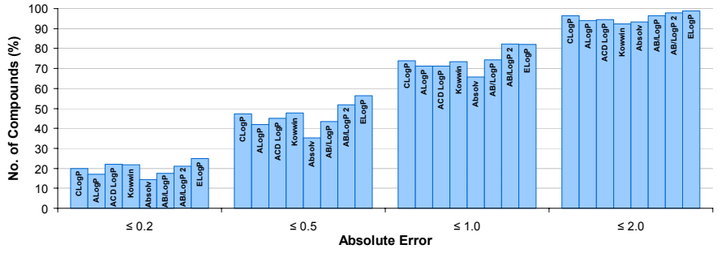
Data Set & Methodology

Selected methods have been applied to a test set of 1000 compounds randomly selected from Syngenta research projects. Nearly 4000 additional compounds with measured *LogP* values from the same source were utilized in the investigation using 'Trainable LogP' method. Predictions obtained for each method were compared to the Syngenta measured values in terms of R-squared, intercept, slope and mean absolute errors (MAE) calculated from plots normalized to give a slope of 1 and intercept of 0. In addition percentage average error values within 0.2, 0.5, 1.0 and 2.0 log units were assessed for each method.

Method Comparison

In the table the results are presented for the six listed structure based methods of *LogP* prediction along with new AB/LogP 2.0 and the Syngenta consensus ELogP methods. The bar plot displays percentage of compounds in the 1000 molecule test set having predicted absolute errors within each of the selected threshold values.

Method	R^2	Slope	Intercept	MAE
Structure Based				
CLogP	0.56	0.61	1.02	0.61
ALogP	0.51	0.57	1.25	0.66
ACD LogP	0.54	0.58	1.21	0.62
Kowwin	0.57	0.57	1.09	0.61
Absolv LogP	0.53	0.53	1.26	0.64
AB/LogP	0.57	0.62	1.02	0.62
AB/LogP 2.0	0.60	0.80	0.56	0.58
Consensus Based				
ELogP	0.67	0.73	0.75	0.52



As can be seen from the data presented above all of the selected structure based methods yield broadly comparable results on our test set. Absolv LogP gave a somewhat weaker performance in terms of the percentage of the compounds having predicted absolute errors of up to 1 log unit but the difference from other methods is really quite marginal. The Syngenta ELogP method performed noticeably better in all aspects than any of the individual methods it is based on supporting the rationale of deriving a consensus value from a number of methods rather than trusting any single one of them. However it is still restricted by some of the limitations of the structure based methods it is derived from and from the obtained results it appears unlikely that any prediction method trained on literature data sets could give rise to an $R^2 > 0.7$ and MAE < 0.5 for novel 'in house' data sets.

Trainable LogP

The last remark made in the Method Comparison discussion actually addresses one of the fundamental problems preventing the effective use of third-party predictive algorithms in the chemical industry, i.e. the literature based training set rarely covers the specific part of the chemical space occupied by the compounds that a certain company is working with or sometimes a specific experimental protocol used to measure the property of interest yields results contrasting with experimental values for the same compounds in the training set. Therefore the need has long existed for a method that would allow any company to effectively tailor a third-party predictive algorithm to its specific needs using proprietary 'in house' data.

Addressing this issue Pharma Algorithms has developed a concept of 'Trainable Models' that provides a novel solution to this problem. Each 'Trainable Model' consists of the following parts:

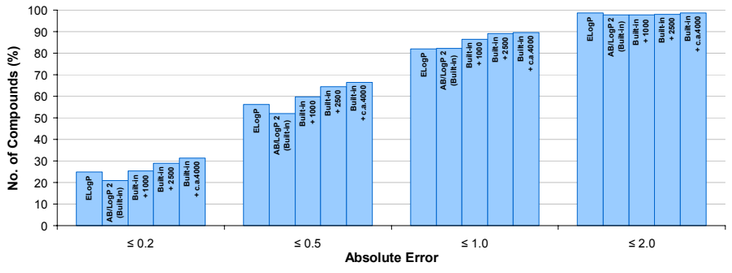
- A structure based QSAR/QSPR for the prediction of a certain property derived from a literature training set by Pharma Algorithms – the so called baseline QSAR/QSPR
- Any user defined data set with the experimental values for the property of interest – the so called Self-training Library
- A special similarity based routine that in the case of every molecule provided for prediction allows automatic selection of the most similar compounds from the Self-training Library and identification of any systematic errors produced by the baseline QSAR/QSPR for that group of compounds

The result is the final predicted value that is corrected according to the experimental results present in the user defined Self-training Library that covers the part of the chemical space not initially included in the training set or any instances of contradiction of experimental results between the literature based training set and the provided library for whichever reasons they occur.

A series of 'Trainable LogP' models using different Self-training Libraries were derived in the manner described above and applied to the test set of 1000 Syngenta compounds:

- Using Built-in Self-training Library representing the literature based training set of the baseline model
- Using Built-in Self-training Library in combination with 1000 and 2500 portions, and the complete nearly 4000 Syngenta compounds data set
- Using the different portions of the Syngenta data set as a sole source for the Self-training Libraries

Self-training Library	R^2	Slope	Intercept	MAE
Using Built-in Library				
Built-in Library (AB/LogP 2.0)	0.60	0.80	0.56	0.58
Built-in Library + 1000 Syngenta compounds	0.65	0.83	0.47	0.53
Built-in Library + 2500 Syngenta compounds	0.70	0.84	0.41	0.48
Built-in Library + c.a. 4000 Syngenta compounds	0.72	0.85	0.40	0.46
Without Built-in Library				
1000 compounds Syngenta Library	0.64	0.81	0.52	0.54
2500 compounds Syngenta Library	0.70	0.82	0.49	0.49
c.a. 4000 compounds Syngenta Library	0.72	0.83	0.46	0.48



These presented results clearly show that 'Trainable LogP' model successfully copes with the task of training itself on specific compounds from the Syngenta database as the addition of growing numbers of such compounds to the Self-training Library of the model gave a steady increase in the accuracy of the predictions and improvement of the distribution of compounds according to the absolute error values.

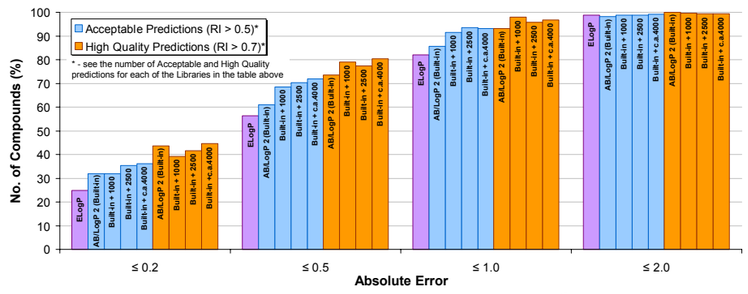
Reliability Index

In the final part of this work we address the question of the quality of the predictions. Every model, no matter what data, descriptors or modelling techniques were used building it, has a certain applicability domain, beyond which the quality of predictions becomes highly questionable. This issue is especially relevant in the application of third-party methods trained on literature data sets to the proprietary 'in house' compounds of any company and various attempts are being made to assess this aspect of the model either qualitatively or quantitatively. Among them is the implementation of Reliability Index (RI) into the 'Trainable Model' methodology by Pharma Algorithms. This index, that is provided for every prediction, can have values in the range [0 to 1] and serves as an evaluation of whether a compound the model is trying to make prediction for is in the chemical space of the model. Lower values suggest compound being further from the model space and prediction less reliable, on the other hand high RI values indicate, that one can be quite confident about the quality of the prediction. Estimation of the Reliability Index takes into account the following aspects:

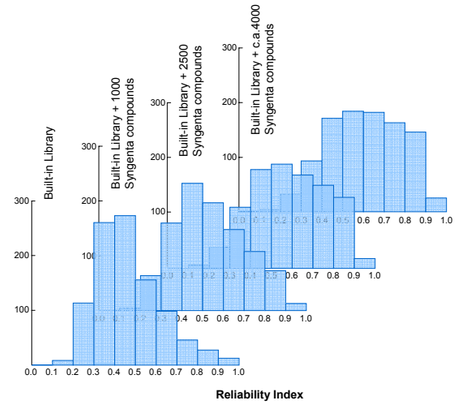
- similarity of the tested compound to the training set
- consistence of experimental values for similar compounds

The results for the four 'Trainable LogP' models with the Self-training Libraries derived from the Pharma Algorithms Built-in *LogP* library only with predictions of acceptable (RI > 0.5) and high (RI > 0.7) quality taken into account are presented and compared below in the familiar manner of statistical summary table and predicted absolute error values distribution bar chart.

Method	R2	Slope	Intercept	MAE
Consensus Based				
ELogP (1000 comp.)	0.67	0.73	0.75	0.52
Trainable LogP				
Acceptable Quality Predictions (RI > 0.5)				
Built-in Library (342 comp.)	0.65	0.80	0.51	0.53
Built-in Library + 1000 Syngenta compounds (539 comp.)	0.71	0.86	0.38	0.45
Built-in Library + 2500 Syngenta compounds (662 comp.)	0.74	0.89	0.31	0.41
Built-in Library + c.a. 4000 Syngenta compounds (700 comp.)	0.75	0.90	0.28	0.40
High Quality Predictions (RI > 0.7)				
Built-in Library (87 comp.)	0.76	0.86	0.24	0.39
Built-in Library + 1000 Syngenta compounds (194 comp.)	0.81	0.96	0.13	0.34
Built-in Library + 2500 Syngenta compounds (300 comp.)	0.79	0.95	0.13	0.35
Built-in Library + c.a. 4000 Syngenta compounds (334 comp.)	0.82	0.95	0.14	0.32



As it can be seen the Reliability index clearly correlates with the accuracy of the predictions and with the changes in the distribution according to the absolute error values. This fact justifies the use of the Reliability Index as a measure of the model applicability domain. Moreover it can be noted that the enlargement of the Self-training Library gives not only the effect of rising accuracy but most importantly the increase of the share of better quality predictions or in other words the expansion of the models applicability domain. This is clearly demonstrated by the number of acceptable and high quality predictions shown in the parentheses in the above table and the histograms of compound number distribution according to the RI value for four 'Trainable LogP' models with Self-training Libraries of increasing size presented here.



Acknowledgements

Authors wish to thank Ann Stainforth and Tom Sheldon (Industrial Placement Students, University of Bath) for assistance with data analysis.

References

1. Clarke, E. D., Delaney, J. S., *Chimia*, 2003, 57, 731-734