Additional Physicochemical Models for StarDrop

Introduction

This paper describes the generation and validation of Quantitative Structure Activity Relationship (QSAR) models of physicochemical properties, based on data made available by the US Environmental Protection Agency (EPA) as part of its Toxicity Evaluation Software Tool (T.E.S.T.) toolkit (1). The models were built with StarDrop's Auto-Modeller module and are available to all StarDrop users free-of-charge to download from Optibrium's on-line community at http://www.optibrium.com/community.

Overview

In the following section we will describe the data used to build the models herein, obtained from the EPA and previously used to build the models in the EPA's T.E.S.T. software (Models of the toxicity properties in T.E.S.T. have previously been developed for StarDrop and are available from http://www.optibrium.com/community/downloads/models). In the Methods section, we will provide an overview of the methods used to build models of this data in StarDrop and under Results we will summarise the validation results for the resulting models and compare these with the results for the T.E.S.T. models. Finally, we will summarise the outcome and draw some conclusions.

Data

The data used to build these models were obtained from the EPA website and are provided with version 4.0 of the EPA's T.E.S.T (1). Data sets for the following properties were used to build QSAR models in StarDrop:

- The boiling point in °C at which a chemical boils at standard atmospheric pressure (Boiling point)
- The density in g/cm^3 (Density)
- The flash point in °C, i.e. the lowest temperature at which a chemical can vaporise and form an ignitable mixture in air (Flash point)
- Thermal conductivity in mW/mK (Thermal conductivity)
- The logarithm of the viscosity, a measure of the resistance to fluid flow, in cP (Viscocity logV)
- The surface tension in dyn/cm (Surface tension)
- The negative logarithm of the water solubility in mol/L (Water solubility –logS)

Full details of how the datasets were generated can be found in the User's Guide for T.E.S.T. (version 4.0) (2). The data sets, divided into training and independent prediction sets, can be downloaded from http://www.epa.gov/nrmrl/std/cppb/qsar/DataSets.zip.

The sizes of the training and prediction data sets for each property are summarised in Table 1. The training and prediction data sets used in each case are identical to those used in T.E.S.T. 4.0 to permit direct comparison of the T.E.S.T. models with those generated in StarDrop. The data set splits were generated randomly by the T.E.S.T. project.

Table 1 Summary of data set sizes.

Property	Training set size	Test set size
Boiling point	3003	751
Density	6885	1722
Flash point	6480	1620
Thermal conductivity	275	71
Viscosity logV	346	87
Surface tension	1136	285
Water solubility -logS	4064	1015

7226 IQ Cambridge
Beach Drive, Cambridge
CB25 9TL, UK

Tel: +44 1223 815900
Fax: +44 1223 815907

Email: info@optibrium.com
Website: www.optibrium.com

Methods

StarDrop's Auto-Modeller was used to build all of the models in this study. Full details of the methods employed by the Auto-Modeller can be found in Chapter 6 of the StarDrop Reference Guide (3) and the references therein. However, these are briefly summarized here.

Descriptors

2D SMARTS based descriptors, which are counts of atom type and functionalities, and whole molecule properties such as logP, molecular weight, and polar surface area (a total of 330 descriptors) were calculated. All of the descriptors are listed in detail in Appendix 10.3 of the StarDrop Reference Guide (3).

The calculated descriptors were subjected to a descriptor pre-selection step that removed descriptors with low variance and low occurrence. Specifically, descriptors with a standard deviation less than 0.0005 and descriptors represented by less than 4% of the compounds in the training set were excluded. Also, highly correlated descriptors were excluded (when the pair-wise correlation exceeded 0.95 in the training set), such that just one of the pair remained.

Modelling Methods

The following methods may be applied by the Auto-Modeller to the training set in order to build predictive models:

- Partial Least Square (PLS) (4)
- Radial Basis Function fitting (RBF) (3)
- Radial Basis Function fitting with Genetic Algorithm descriptor selection (RBF-GA) (3)
- Gaussian Processes (GP) with the following methods for hyperparameter determination (5)
    - Fixed (GPFixed)
    - 2DSearch (GP2DSearch)
    - Forward Variable Selection (GPFVS)
    - Rescaled Forward Variable Selection (GPRFVS)
    - Optimised (GPOpt)
    - Nested Sampling (GPNest)

For the Boiling point, Surface tension and Water solubility data sets, the PLS, two RBF and GPFixed methods were applied; in the case of Density and Flash point, only the RBF and PLS methods were used. However, for the Viscosity data sets, all methods were used with the exception of GPNest. The modelling methods were restricted owing to the size of some of the data sets, which would have made the model building process intractable.

For each model the 'chemical space' which is represented by the training set is captured (also described as the 'domain of applicability'). The position of a new compound relative to the chemical space of a model is reflected in the reported confidence in the prediction for the new compound. The StarDrop models employ the Hotelling's T^2 method for representing the chemical space of the model (for more information see Section 2.6 of the StarDrop Reference Guide (3)). Furthermore, the GP algorithms also provide an estimate of the uncertainty in the prediction of each individual compound within the chemical space.

Validation

Each model was validated by application to the independent prediction set for the relevant property. This is a true measure of the predictive performance of a model, which is preferable to an internal measure of performance such as goodness of fit to the training set or cross validation.

The predictive performance of each regression model for a numerical property was assessed by calculation of the coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i^{pred} - y_i^{obs})^2}{\sum_{i=1}^{N}(y_i^{obs} - \overline{y_i^{obs}})^2},$$

and the root-mean-square error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}(y_i^{pred} - y_i^{obs})^2}$$

where y_i^{pred} and y_i^{obs} are respectively the predicted and observed values of the property for compound *i* and *N* is the number of compounds in the prediction set. For each of the numerical properties, the model with the highest R^2 and lowest RMSE was selected.

The coefficient of determination ranges from 0 to 1 and the closer it is to 1 the better the model describes the proportion of the variation in the observed property values that is explained by the fitted regression, e.g. if we have $R^2=0.85$ this means that 85% of the variation in the property is explained by the model. Note that this definition of R^2 is different from the Pearson correlation coefficient,

$$R^2_{\text{Pearson}} = \frac{\left(\Sigma_{i=1}^{N}(y_i^{pred} - \overline{y_i^{pred}})(y_i^{obs} - \overline{y_i^{obs}})\right)^2}{\Sigma_{i=1}^{N}(y_i^{pred} - \overline{y_i^{pred}})^2 \Sigma_{i=1}^{N}(y_i^{obs} - \overline{y_i^{obs}})^2},$$

which is a measure of how well the predicted versus observed values fit to a straight line, but not the ideal line. However, as the R^2_{Pearson} value was used in the validation of the models provided by the T.E.S.T. package (unfortunately also denoted R^2 in the user guide for this package), this was also calculated for each model generated using StarDrop. Similarly the mean-absolute-error (MAE),

$$MAE = \frac{1}{N}\left|y_i^{pred} - y_i^{obs}\right|,$$

the slope of the best fit line constrained to pass through the origin, *k*, and

$$\frac{R^2_{\text{Pearson}} - R^2_0}{R^2_{\text{Pearson}}},$$

where R^2_0 is the correlation coefficient of the best fit line forced to go through the origin, were calculated for comparison with the models in the T.E.S.T. package. Generally, a regression model is considered to have acceptable predictive power if $R^2_{\text{Pearson}} > 0.6$, $\frac{R^2_{\text{Pearson}} - R^2_0}{R^2_{\text{Pearson}}} < 0.1$ and $0.85 \leq k \leq 1.15$ (6).

Results

A summary of the results for the best StarDrop models on the independent predictions sets , along with a comparison with the best models in the T.E.S.T. v4.0 package are shown in Table 2. All of the StarDrop models meet the minimum standards with R^2 and R^2_{Pearson} much greater than 0.6. The performances of the best StarDrop and T.E.S.T. models for each property are very comparable.

The regression plots for the best StarDrop model for each property are shown in the Appendix to this document.

Glowing Molecule™

The StarDrop models also benefit from the Glowing Molecule visualization that highlights regions of a compound that have a significant influence on a predicted property (an example of this is shown in Figure 1). This provides a link between the predicted property and the compound structure, helping to guide the redesign of compounds with improved properties.

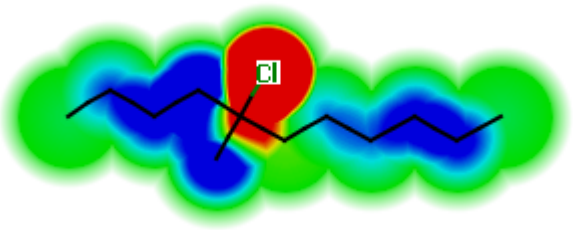

Figure 1 The Glowing Molecule highlights regions of a molecule that have a significant impact on a predicted property. In this case the chlorine is predicted to have a significant effect to increase the surface tension.

Table 2 Summary of StarDrop results on the independent prediction sets for best StarDrop models and comparison with best T.E.S.T. models.

Property	Best StarDrop Model							Best T.E.S.T. Model					
	Method	R^2	RMSE	$R^2_{Pearson}$	$\frac{R^2_{Pearson} - R^2_0}{R^2_{Pearson}}$	k	MAE	Method	RMSE	$R^2_{Pearson}$	$\frac{R^2_{Pearson} - R^2_0}{R^2_{Pearson}}$	k	MAE
Boiling point	GPFixed	0.97	14.4	0.97	0.00	1.00	8.32	Consensus	15.73	0.96	0.00	1.00	10.0
Density	RBF	0.95	0.08	0.95	0.00	0.99	0.04	Consensus	0.09	0.94	0.01	0.99	0.04
Flash point	RBF	0.90	26.8	0.90	0.01	0.97	16.1	Consensus	28.9	0.88	0.01	0.96	18.0
Thermal conductivity	GPRFVS	0.83	13.11	0.84	0.03	0.97	9.86	Consensus	13.19	0.85	0.10	0.96	8.39
Viscosity logV	GA RBF	0.94	0.13	0.95	0.00	0.92	0.09	Consensus	0.20	0.86	0.00	0.86	0.12
Surface tension	RBF	0.87	2.26	0.87	0.01	1.00	1.07	Consensus	1.82	0.92	0.02	1.00	1.23
Water solubility -logS	RBF	0.88	0.81	0.88	0.01	0.93	0.52	Consensus	0.86	0.87	0.02	0.93	0.59

Conclusions

The models described in this paper were all built using StarDrop's Auto-Modeller without manual intervention, reinforcing previous examples that illustrated the Auto-Modeller's capability to build and validate models that are comparable with those built 'manually' using other methods (7). The resulting models comfortably meet the standards for acceptable predictive power on the independent prediction sets.

Finally, from inspection of the data sets, it is apparent that the majority of the compounds are not 'drug-like'. Therefore, some of the models may be more relevant for application to potential environmental pollutants than in drug discovery; this information is captured by the domain of applicability of the models and reported in the confidence of each prediction in StarDrop.

Bibliography

1. **US Environmental Protection Agency.** Quantitative Structure Activity Relationship. *Environmental Protection Agency.* [Online] 2011. [Cited: May 20, 2011.] http://www.epa.gov/nrmrl/std/cppb/qsar/#TEST.

2. User's Guide for T.E.S.T. (version 4.0). *www.epa.org.* [Online] 2011. [Cited: May 20, 2011.] http://www.epa.gov/nrmrl/std/cppb/qsar/testuserguide.pdf.

3. **Optibrium Ltd.** *StarDrop Reference Guide Version 5.0.* Cambridge : Optibrium Ltd., 2011. Manual.

4. **Wold, S., Sjostrom, M. and Eriksson, L.** Partial Least Squares Projectoins to Latent Structures (PLS) in Chemistry. [book auth.] P von Rague Schleyer, et al. *The Encyclopedia of Computational Chemistry.* Chichester, UK : John Wiley and Sons, 1999, pp. 1-16.

5. *Gaussian processes: a method for automatic QSAR modeling of ADME properties.* **Obrezanova, O, et al.** 2007, J. Chem. Inf. Model., Vol. 47, pp. 1847-1857.

6. *Rational Selection of Training and Test sets for the Development of Validated QSAR Models.* **Golbraikh, A., et al.** 2-4, 2003, Vol. 17, pp. 241-253.

7. *Automatic QSAR modeling of ADME properties: blood-brain barrier penetration and aqueous solubility.* **Obrezanova, O, et al.** 2009, J. Comput.-Aided Mol. Des., Vol. 22, pp. 431-440.

Appendix - Regression Plots for Numerical Models

The regression plots are shown below for the independent predictions sets for each of the StarDrop models. The red line on each is the identity line, i.e. the ideal line for predicted against observed, for comparison.