Human Volume of Distribution Models

The volume of distribution (VDss) is an *in vivo* pharmacokinetic parameter representing the hypothetical volume into which the dose of drug would have to be evenly distributed to give rise to the same concentration observed in the blood plasma. This provides an indication of the distribution of the drug in the body: A low VDss indicates high water solubility or high plasma protein binding, because more of the drug remains in the plasma; a high VDss suggests significant concentration in the tissues, for example due to tissue binding or high lipid solubility.

Here we describe models of VDss that can be downloaded for use within StarDrop, built with StarDrop's Auto-Modeller and based on data published by Gombar and Hall [1].

Data

Gobar and Hall published an article describing the building and validation of models of human, clinical pharmacokinetic parameters, namely VDss and clearance. The data sets with which these models were built and validated were provided in the supplementary information to their paper [1].

The models of VDss described by Gombar and Hall were trained with a set containing 569 compounds with published clinical data. For the purposes of building and validating models within the StarDrop Auto-Modeller, this data set was divided into independent training, validation and test sets containing 399, 85 and 85 compounds respectively. The set split was performed using the Auto-Modeller's default clustering method with a cluster Tanimoto index of 0.7 (see the StarDrop Reference Guide for more details). The VDss data was transformed into log units, in common with the approach of Gombar and Hall.

The VDss models built by Gombar and Hall were tested using two external test sets: 22 compounds obtained from a paper by Berelini *et al.* [2] and 9 compounds published by Poulin and Theil [3]. These data sets were also used as external test sets in this work, to allow direct comparison with the models of Gombar and Hall.

Methods

The Auto-Modeller was applied the to the training, validation and test sets as described above. The default descriptors and parameters for descriptor selection were used and models were generated using the partial least squares (PLS), radial basis functions (RBF), random forests (RF), and four Gaussian Processes methods (GPFixed, GP2DSearch GPRFVS and GPOpt).

Details of the parameters and descriptors used are provided in the supporting information, which can be downloaded as described below.

Results

The performance of the models built with the Auto-Modeller are shown in the table below (only the best of the Gaussian Processes models is shown):

Model	Training Set					Validation Set					Test Set				
	R^2 (log units)	RMSE (log units)	Med FD	Max FD	% <3FD	R^2 (log units)	RMSE (log units)	Med FD	Max FD	% <3FD	R^2 (log units)	RMSE (log units)	Med FD	Max FD	% <3FD
PLS	0.43	0.47	1.9	149	74	0.42	0.47	2.1	29	74	0.52	0.43	1.6	37	73
RBF	N/A	N/A	N/A	N/A	N/A	0.67	0.36	1.6	9	81	0.68	0.35	1.4	12	84
RF	0.91	0.19	1.3	7	98	0.62	0.38	1.5	10	76	0.63	0.37	1.5	16	84
GPFixed	0.73	0.32	1.5	49	88	0.62	0.38	1.8	11	76	0.64	0.37	1.6	14	82

R^2 = coefficient of determination, RMSE = Root Mean Square Error, Med FD = Median Fold Difference, Max FD = maximum fold difference, %<3FD = percentage less than 3-fold different

7221 Cambridge Research Park
Beach Drive, Cambridge
CB25 9TL, UK

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

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

These models cannot be compared directly with the models generated by Gombar and Hall on the basis of these results, because only the performance of the model trained with the full data set of 569 compounds is reported in reference [1]. However, for reference, the authors report a model trained with support vector regression (SVR) had a median fold deviation of 1.62 and a maximum observed deviation of 8.86-fold on the training set. Gombar and Hall also report a multiple linear regression (MLR) model trained with 560 compounds (after removing outliers) had an R^2 of 0.78 on the training set.

To allow direct comparison of the models generated with the Auto-Modeller and previously published models, the results of applying the models to the independent test set derived from Berellini *et al.* [2] are shown in the table below:

Model	RMSE (log units)	Med FD	Max FD	% <3FD
PLS	0.36	1.7	5.7	77
RBF	0.29	1.7	4.5	91
RF	0.30	1.8	4.3	91
GP Fixed	0.35	1.6	6.8	82
Gombar and Hall SVR	0.35	1.9	4.6	86
Gombar and Hall MLR	0.63	2.1	78	59

RMSE = Root Mean Square Error, Med FD = Median Fold Difference, Max FD = maximum fold difference, %<3FD = percentage less than 3-fold different

For further comparison, the results of applying the models to the independent test sets derived from Poulin and Theil [3] are shown in the table below:

Model	RMSE (log units)	Med FD	Max FD	% <3FD
PLS	0.16	1.3	1.9	100
RBF	0.15	1.2	2.0	100
RF	0.16	1.3	1.8	100
GP Fixed	0.18	1.4	2.0	100
Gombar and Hall SVR	0.20	1.6	2.1	100
Gombar and Hall MLR	0.31	1.4	3.9	78
Poulin and Theil	0.18	1.2	2.9	100

RMSE = Root Mean Square Error, Med FD = Median Fold Difference, Max FD = maximum fold difference, %<3FD = percentage less than 3-fold different

The distribution of observed VDss values in the Poulin and Theil set is too narrow to allow a meaningful coefficient of determination (R^2) to be calculated. Therefore, the two independent test sets were combined and the resulting R^2 values are shown in the following table:

Model	R^2 (log units)
PLS	0.40
RBF	0.59
RF	0.56
GP Fixed	0.39
Gombar and Hall SVR	0.40
Gombar and Hall MLR	-0.89

R^2 = coefficient of determination

Based on the results above, the RBF model appears to have the best overall performance of the StarDrop models. However, the RF and GPFixed models also show good performance and may be worth considering. The GPFixed model offers the advantage of producing an estimate of the uncertainty in each prediction on a compound-by-compound basis, although it is notable that the performance on the Berellini and Poulin and Theil data sets is inferior to the RBF and RF models.

Using the VDss Models

The models can be downloaded for use within StarDrop from the following links:

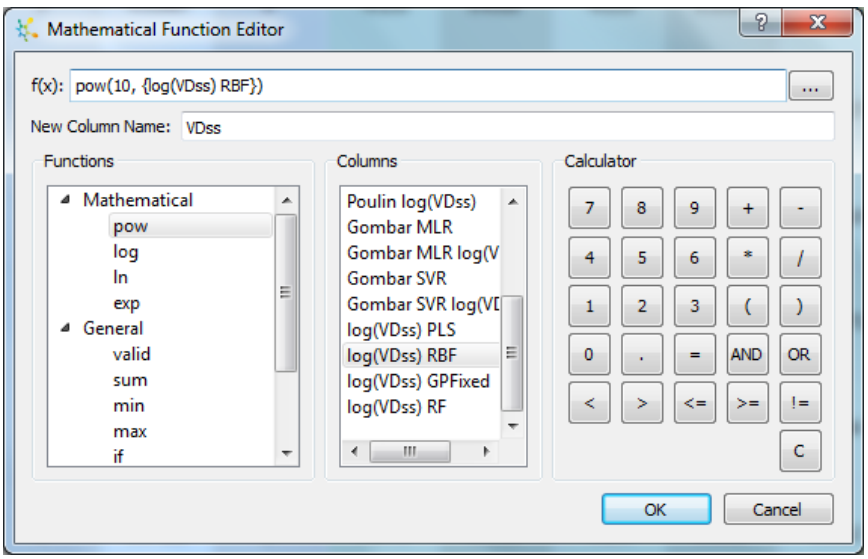
Model	Link
RBF	http://www.optibrium.com/downloads/log(VDss)_RBF.aim
RF	http://www.optibrium.com/downloads/log(VDss)_RF.aim
GPFixed	http://www.optibrium.com/downloads/log(VDss)_GPFixed.aim
PLS	http://www.optibrium.com/downloads/log(VDss)_PLS.aim

To use these models within StarDrop, download and save the model in a convenient place. Load the model into StarDrop using the  button on the **Models** tab. Alternatively, the directory in which the model file has been saved can be added to the paths from which models are automatically loaded when StarDrop starts by selecting the **File->Preference** menu option and adding the directory under **Models** in the **File Locations** tab.

The models predict the logarithm of VDss in L/kg. To convert this to a VDss in L/kg, use the mathematical function tool in StarDrop ( on the toolbar) and enter one of the following equations:

Model	Equation
RBF	pow(10, {log(VDss) RBF})
RF	pow(10, {log(VDss) RF})
GPFixed	pow(10, {log(VDss) GPFixed})
PLS	pow(10, {log(VDss) PLS})

An example is shown in the figure below:



References

1 Gombar VK, Hall SD. Quantitative Structure–Activity Relationship Models of Clinical Pharmacokinetics: Clearance and Volume of Distribution. J. Chem. Inf. Model. 2013;53(4):948–957.

2 Berellini G, Springer C, Waters NJ, Lombardo F. In silico Prediction of Volume of Distribution in Human Using Linear andNonlinear Models on a 669 Compound Data Set. J. Med. Chem. 2009;52(14):4488-4495.

3 Poulin P, Theil FP. Prediction of Pharmacokinetics Prior to In Vivo Studies. 1. Mechanism-Based Prediction of Volume of Distribution. J. Pharm. Sci. 2002;91(1):129-156.

Supporting Information

The data sets and detailed outputs from the modelling process may be downloaded in a .zip archive. The contents of this archive are as follows:

- VDss models overview.pdf: This document
- logD(VDss) summery.pdf: Summary of Auto-Modeller output for modelling of log(VDss)
- log(VDss)_RBF.aim: StarDrop RBF model of log(VDss)
- log(VDss) RBF.pdf: Detailed Auto-Modeller output for RBF model of log(VDss)
- log(VDss)_RF.aim: StarDrop RF model of log(VDss)
- log(VDss) RF.pdf: Detailed Auto-Modeller output for RF model of log(VDss)
- log(VDss)_GPFixed.aim: StarDrop GPFixed model of log(VDss)
- log(VDss) GPFixed.pdf: Detailed Auto-Modeller output for GPFixed model of log(VDss)
- log(VDss)_PLS.aim: StarDrop PLS model of log(VDss)
- log(VDss) PLS.pdf: Detailed Auto-Modeller output for PLS model of log(VDss)
- Gombar_Hall_VDss_training.csv: Training set derived from data set provided in reference [1]
- Gombar_Hall_VDss_validation.csv: Validation set derived from data set provided in reference [1]
- Gombar_Hall_VDss_test.csv: Test set derived from data set provided in reference [1]
- Berellini test set results.csv: Predictions for independent training set derived from reference [2]
- Poulin Theil test set results.csv: Predictions for independent training set derived from reference [3]