

## Shen Blood-Brain Barrier Models

Blood-brain barrier (BBB) penetration is a measure of the ratio between the compound concentration in brain and blood. Good BBB penetration is required for compounds intended for targets in the central nervous system (CNS). Alternatively, for peripheral targets, poor BBB penetration reduces the risk of CNS side effects.

Here we describe classification models of BBB, built with StarDrop's Auto-Modeller based on data published by Shen *et al.* [1], that can be downloaded for use within StarDrop.

### Data

Shen *et al.* published a paper describing the generation and validation of QSAR models of human intestinal absorption (HIA) and blood-brain-barrier penetration (BBB) [1]. The data sets with which these models were built and validated were provided in the supplementary information.

Shen *et al.* used a data set of 1593 compounds classified as high or low brain penetration. This was divided into a training set of 1093 compounds (832 high and 261 low) and a test set of 500 compounds (451 high and 49 low). An additional external test set of 246 compounds (155 high and 91 low) was also identified from a paper published by Li *et al.* [2].

As discussed below, it was notable that models built with the training set selected in [1] performed poorly on the external test set from [2], suggesting that the compound structures in the external test set are not well represented in the training set. Therefore, to ensure as broad a coverage of chemical diversity as possible, the three sets were combined to create a single set containing 1838 compounds and split using StarDrop's Auto-Modeller into training, validation and test sets in the proportions 70:15:15, using a Y-based sampling. The resulting data sets are summarised in Table 1 and are included in the supporting information, as described below.

**Table 1 Overview of data set split generated using the Auto-Modeller.**

Data set	Number High	Number Low
Training	1042	246
Validation	201	74
Test	194	81

### Methods

The Auto-Modeller was applied to the original data sets provided by Shen *et al.* [1] to allow direct comparisons with the models generated in this paper. The Auto-Modeller was subsequently applied to the revised data set split, as described above.

In both cases, the default descriptors and parameters for descriptor selection were used and models were generated using the decision tree (DT) and random forest (RF) methods.

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

## Results

### Shen *et al.* data set split

Table 2 shows a comparison of the random forests (RF) model generated by the Auto-Modeller with the best model of Shen *et al.* and a model of the same data set reported in a previous work by Zhao *et al.* [3].

**Table 2** Results of RF model generated with the Auto-Modeller using data set split in [1] and compared with the models published in [1] and [3]. TP is number of true positives, TN is number of true negatives, FP is number of false positives and FN is number of false negatives.  $\kappa$  is the kappa statistic as described in Section 6.8.7 of the StarDrop Reference Guide.

Model	Training					Test					External Test				
	TP	TN	FP	FN	$\kappa$	TP	TN	FP	FN	$\kappa$	TP	TN	FP	FN	$\kappa$
Shen <i>et al.</i>	832	258	3	0	0.99	449	42	7	2	0.89	149	19	72	6	0.20
Zhao <i>et al.</i>	815	246	15	17	0.92	443	43	6	8	0.84	-	-	-	-	-
Auto-Modeller (RF)	832	261	0	0	1	449	38	11	1	0.85	145	27	64	10	0.26

### Revised Data Set Split

As noted above, the poor performance on the external test set of both the models generated by Shen *et al.* and the RF model generated by the Auto-Modeller suggests that the compound structures in the external test set are not well represented in the training set. Therefore, the Auto-Modeller was applied to the revised data set split, as described above. The best model resulting from this split was a RF model and its performance is summarised in Table 3.

**Table 3** Results of RF model generated with the Auto-Modeller the revised data set split. TP is number of true positives, TN is number of true negatives, FP is number of false positives and FN is number of false negatives.  $\kappa$  is the kappa statistic as described in Section 6.8.7 of the StarDrop Reference Guide.

Model	Training					Validation					External Test				
	TP	TN	FP	FN	$\kappa$	TP	TN	FP	FN	$\kappa$	TP	TN	FP	FN	$\kappa$
Auto-Modeller (RF)	1040	246	0	1	1.00	200	56	18	1	0.93	194	66	15	0	0.95

The results for this model are excellent, with a  $\kappa$  statistic above 0.9 on both the validation and test set.


## Using the BBB Models

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

[BBB Shen training.aim](#) : The RF model generated with the data set split in Shen *et al.* [1]

[BBB Shen full set.aim](#) : The RF model generated using the revised split of the full set published in Shen *et al.* [1]

To use these within StarDrop, download and save these files in a convenient place. Load them into StarDrop

using the  button on the **Models** tab. Alternatively, the directory in which the model files have 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.

## References

- [1] Shen *et al.* J. Chem. Inf. Model. 2010, **50**( 6) pp. 1034-1041
- [2] Li *et al.* J. Chem. Inf. Model. 2005, **45**(5), 1376-1384
- [3] Zhao *et al.* J. Chem. Inf. Model. 2007, **47**(1), pp. 170-175

## 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:

- Shen BBB models overview.pdf: This document
- BBB\_Shen\_training\_summary.pdf: Summary of Auto-Modeller output for models built with BBB training set from reference [1]
- BBB\_Shen\_training\_RF.pdf: Detailed output for random forests model built with BBB training set from [1]
- BBB\_Shen\_training.aim: StarDrop model built using random forests method using BBB training set from reference [1]
- Shen\_BBB\_training\_set.csv: Comma separated value file containing training set, as published in reference [1], including descriptors, observed and predicted values
- Shen\_BBB\_test\_set.csv: Comma separated value file containing test set, as published in reference [1], including descriptors, observed and predicted values
- Shen\_BBB\_external\_test\_set.csv: Comma separated value file containing the external test set, as published in reference [1], including descriptors, observed and predicted values
- BBB\_Shen\_full\_set\_summary.pdf: Summary of Auto-Modeller output for models built with full BBB set from reference [1], with corrected structures and split into training validation and test sets within the Auto-Modeller.
- BBB\_Shen\_full\_set\_RF.pdf: Detailed output for random forests model built with corrected and resplit data set from reference [1]
- BBB\_Shen\_full\_set.aim: StarDrop model built using random forests method with corrected and resplit data set from reference [1]
- Shen\_BBB\_full\_set\_training.csv: Comma separated value file containing training set derived from Auto-Modeller split of full data set from reference [1]
- Shen\_BBB\_full\_set\_validation.csv: Comma separated value file containing validation set derived from Auto-Modeller split of full data set from reference [1]
- Shen\_BBB\_full\_set\_test.csv: Comma separated value file containing test set derived from Auto-Modeller split of full data set from reference [1]