Drug Metabolism and Disposition

A PREDICTIVE LIGAND-BASED BAYESIAN MODEL FOR HUMAN DRUG INDUCED LIVER INJURY

Sean Ekins, Antony J. Williams and Jinghai J. Xu

Collaborations in Chemistry, 601 Runnymede Avenue, Jenkintown, PA 19046, U.S.A. (SE)

Department of Pharmaceutical Sciences, University of Maryland, MD 21201, U.S.A. (SE)

Department of Pharmacology, University of Medicine & Dentistry of New Jersey (UMDNJ)-Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, NJ 08854. (SE)

Royal Society of Chemistry, 904 Tamaras Circle, Wake Forest, NC-27587. (AJW)

Merck & Co., Inc., 126 E. Lincoln Ave, Rahway, NJ 07065. (JJX)

Supplemental data for Bayesian model

Output from discovery studio

Leave-one-out Cross-Validation Results

This model was built using 295 samples, and validated using a leave-one-out cross-validation. Each sample was left out one at a time, and a model built using the results of the samples, and that model used to predict the left-out sample. Once all the samples had predictions, a ROC plot was generated, and the area under the curve (**XV ROC AUC**) calculated.

Best Split was calculated by picking the split that minimized the sum of the percent misclassified for category members and for category nonmembers, using the cross-validated score for each sample. Using that split, a contingency table is constructed, containing the number of true positives (**TP**), false negatives (**FN**), false positives (**FP**), and true negatives (**TN**).

Output	XV ROC AUC	Best Split	TP/FN FP/TN	# in Category
DILI new Bayesian ECFC6 EVEN more des II	0.860	0.167	115/43 17/120	158

Enrichment Results

Back to Top

This model was built using 295 samples, and validated using a leave-one-out cross-validation. Each sample was left out one at a time, and a model built using the results of the samples, and that model used to predict the left-out sample. Once all the samples had predictions, an enrichment plot was generated, and the percentage of true category members captured at a particular percentage cutoff. (For example, in a column labeled "1%" would be the percentage of true category members (e.g., actives) that were found in the top 1% of the list, when sorted by the model score.)

This table shows the output name, the percentage of samples that are in that particular category, the number of category members, and the percentage of true members found. Percentages that are less than 100% are in **bold**.

Output	Category %	1%	5%	10%	25%	50%	75%	90%	95%	99%
DILI new Bayesian ECFC6 EVEN more des II	53.559%	**1.9%**	**8.9%**	**17.7%**	**43%**	**75.9%**	**94.9%**	**99.4%**	**100%**	**100%**

Percentile Results

Back to Top

This table shows, for each model, the cutoff needed to capture a particular percentage of the good samples. For each cutoff, it shows below the estimated percentages of false positives and true negatives for the non-good samples. This table is designed to help you pick the cutoff value that best balances your desire to capture as many good samples as possible, while keeping the number of false positives at a minimum.

The rates shown in this table are estimates derived from the cross-validated data; the actual numbers you would find on your own data may vary.

Cutoff which lead to 10% or greater false positives are displayed in **bold** for ease of identification.

Model Name	99%	95%	90%	70%	50%	30%	10%	5%	1%
DILI new Bayesian ECFC6 EVEN more des II	-11.190 **62%**/38%	-7.008 **48%**/52%	-4.739 **40%**/60%	-2.230 **32%**/68%	-2.230 **18%**/82%	7.924 9%/91%	10.433 6%/94%	12.703 4%/96%	16.884 2%/98%

Category Statistics Results

Back to Top

This table shows, for each category, statistics derived from the cross-validated predictions of the model built for that category as applied to members of that category and non-members of that category. For each group, the number of members/nonmembers (N) is given; the mean prediction for each subset (Mean); and the estimate standard deviation of the predictions for each subset (StdDev).

(Categories with one or no members do not have a mean and standard deviation, as there are too few predictions upon which to base them during cross-validation. Also, occasionally categories may contain many duplicate or highly-similar compounds which predict close or identical values, causing them to have unusually low standard deviation values. These low values may be adjusted at time of use of these standard deviations for predicting, for example, percentile results.)

Output	Category N	Category Mean (±StdDev)	Noncategory N	Noncategory Mean (±StdDev)
DILI new Bayesian ECFC6 EVEN more des II	158	2.85 (±5.97)	137	-7.81 (±11.33)

Non-validated Models Results

Back to Top

Training Data Information

Back to Top

The properties used to provide the variables were: **ALogP**; **ECFC_6**; **Apol**; **logD**; **Molecular_Weight**; **Num_AromaticRings**; **Num_H_Acceptors**; **Num_H_Donors**; **Num_Rings**; **Num_RotatableBonds**; **Molecular_PolarSurfaceArea**; **Molecular_SurfaceArea**; **Wiener**; **Zagreb**

The test to identify "good" samples is:

```
property("DILI_Bins_Binary *") is defined AND property("DILI_Bins_Binary *") = 1;
```

You can extend this model by adding your own training data to it to create a new model, but because the original training data is no longer available, you will not be able to re-validate the new model. This extending is done using the *New Model from Old* component. The new training samples must already have the appropriate properties as specified above (though properties that can be calculated-on-demand will be). The "good" samples must be marked so that they will be correctly identified by the aforementioned test.

Model Construction Information

Back to Top

Model construction information:

Post-processing was performed to remove low-information bins. Low-information bins are those who have: normalized estimates in the range [-0.05, 0.05].

For each property, the following table gives the original number of bins (*Original*), the number removed due to too few samples (*TooFew*), the number removed due to a poor normalized estimate (*Noninformative*), and the final number of bins saved in the model (*Final*).

Property	Original	TooFew	Noninformative	Final
ALogP	11	0	1	10
ECFC_6	7094	0	437	6657
Apol	11	0	1	10
logD	11	0	2	9
Molecular_Weight	11	0	0	11
Num_AromaticRings	5	0	2	3
Num_H_Acceptors	8	0	3	5
Num_H_Donors	6	0	1	5

Num_Rings	5	0	3	2
Num_RotatableBonds	9	0	2	7
Molecular_PolarSurfaceArea	11	0	1	10
Molecular_SurfaceArea	11	0	2	9
Wiener	10	0	2	8
Zagreb	10	0	2	8