

Modeling Rat Liver Toxicity Signature Using Machine Learning Techniques

Robert Fraczekiewicz*, Jinhua Zhang, Marvin Waldman, Walter S. Woltosz
 Simulations Plus, Inc., 42505 10th Street West, Lancaster, CA 93534, USA

OBJECTIVES

Out of the 76 EPA ToxCast(TM) *in vivo* endpoints we selected four chronic rat liver toxicities: Hypertrophy, Necrosis, Proliferative Lesions, and Tumors. We assembled an input file containing after filtering 241 structures in SMILES format. The data in this file are rat liver toxicities converted into a binary format (with any measured effect defined as "Toxic", and no effect labeled "Nontoxic"), and results from ~500 *in vitro* assays in ToxCast also represented as binary descriptors (quantitative representation of the above did not lead to better models). The objective of this study was to build predictive classification models of rat liver toxicities in two combinations of *in vivo* endpoints: Rat_Hypertrophy by itself (62 Toxic / 145 Nontoxic) and Rat_Any – a union of all four (96 Toxic / 145 Nontoxic). Please note the data sets are highly unbalanced.

METHODOLOGY

ADMET Predictor(TM) 3.1.0 (Simulations Plus, Inc.) was used to calculate 325 molecular descriptors and ~80 predicted properties - these were appended to the input file. Next, the ADMET Modeler(TM) module within ADMET Predictor was employed to train thousands of binary classification models based on Artificial Neural Network Ensembles (ANNE) and Support Vector Machine Ensembles (SVME). Each model was validated with an external test set. An early stopping technique was used to prevent overtraining. SVME did not lead to better results, hence only ANNE results are reported. Standard algorithms have failed to produce models of sufficient performance and generalizability. Moreover, standard descriptor selection techniques were overwhelmed with ~900 descriptors to choose from and only 241 examples. Therefore, we have embarked on creating a new generation of algorithms capable of tackling this problem. The following statistics were used to assess models:

$$\text{Concordance} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Balanced Accuracy} = \frac{1}{2}(\text{Sens} + \text{Spec})$$

$$\mu = \frac{1}{2} \left(\frac{TP}{TP + FP} + \frac{TP}{TP + FN} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN} \right) - 1$$

where:

TP, FP = true, false positives

TN, FN = true, false negatives

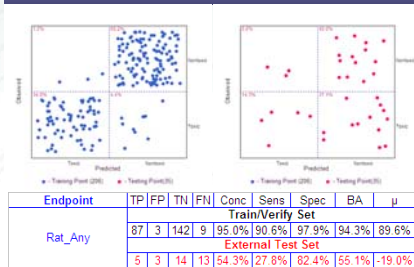
The new statistic, μ , has been developed by Marvin Waldman with measuring the overall model performance on unbalanced data in mind. It ranges from -1 (worst) through 0 (random) to +1 (best scenario).

CONCLUSIONS

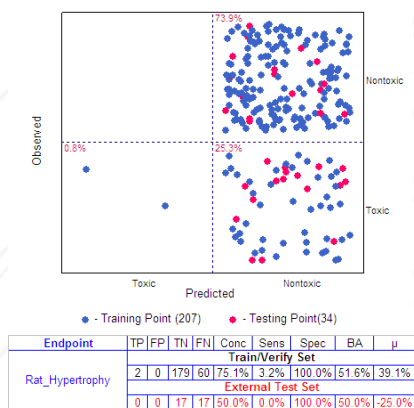
ToxCast Phase I is an extremely challenging data set for modeling. Standard classification algorithms get overwhelmed with an excessively large number of descriptors, too small number of independent examples, and highly unbalanced data. Nevertheless, some of the standard models may look good without **external** validation. We have been able to develop a series of new modeling methodologies, including the new μ statistic with high diagnostic power. This enabled building successful classification models predicting rat liver toxicity endpoints with overall accuracy better than 74% (molecular descriptors only) and better than 80% (mixed molecular and *in vitro* descriptors).

RESULTS

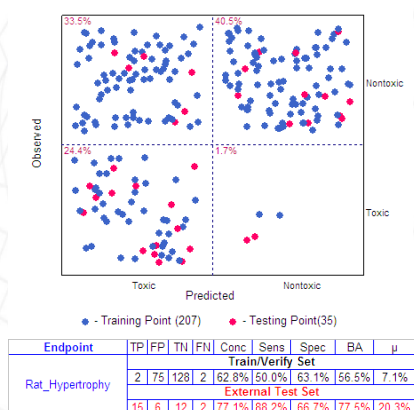
PART 1: WHAT DOESN'T WORK!



This deliberately overtrained example shows why the presence of an external test set is **so important**. Although the training set statistics are good, the test results are poor, indicating that the model is not predictive. In conclusion, any reports where models were not validated with an **external** test set should not be trusted.

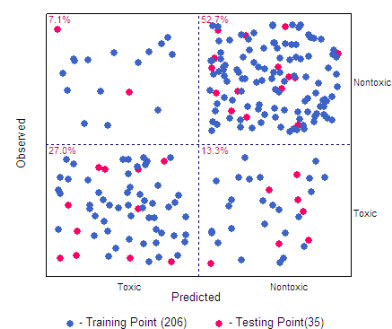


Standard, unweighted classification methodology effectively assigns almost all examples to the Nontoxic class resulting in deceptively good predictivity and specificity, but low sensitivity and low μ .



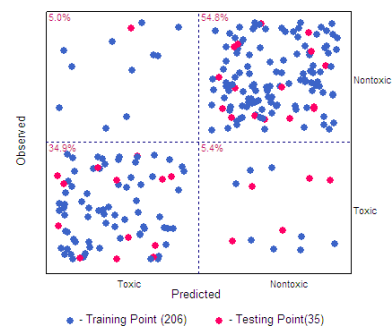
A load-balanced ANNE classifier with strong weighting scheme is able to discriminate Toxic compounds, but at the cost of creating a high number of false positives.

PART 2: WHAT WORKS



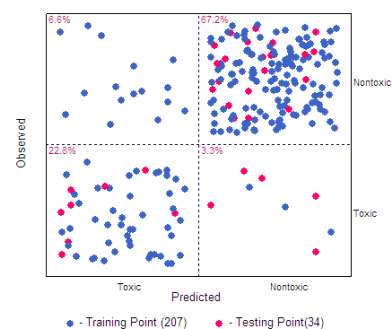
Endpoint	TP	FP	TN	FN	Conc	Sens	Spec	BA	μ	
Rat_Any	Train/Verify Set									
	54	15	112	25	80.6%	58.4%	88.2%	78.3%	58.3%	
Rat_Any	External Test Set									
	11	2	15	7	74.3%	81.1%	88.2%	74.7%	8.8%	

The load-balanced ANNE classifier with a weak weighting scheme and single threshold optimized against μ gives the best results. This model of Rat_Any uses 32 networks with 40 molecular descriptors and 1 hidden neuron.



Endpoint	TP	FP	TN	FN	Conc	Sens	Spec	BA	μ	
Rat_Any	Train/Verify Set									
	71	11	116	8	90.8%	89.9%	91.3%	90.6%	80.7%	
Rat_Any	External Test Set									
	13	1	16	5	82.9%	72.2%	94.1%	83.2%	21.3%	

Same classifier and endpoint as above, but including *in vitro* descriptors selected by the cascading input gradient method. This model uses 32 networks with 23 molecular and 57 *in vitro* descriptors, and 2 hidden neurons.



Endpoint	TP	FP	TN	FN	Conc	Sens	Spec	BA	μ	
Rat_Hypertrophy	Train/Verify Set									
	47	16	141	3	90.8%	94.0%	89.8%	91.9%	78.2%	
Rat_Hypertrophy	External Test Set									
	8	0	21	5	85.3%	81.5%	100.0%	80.8%	21.2%	

This model of Rat_Hypertrophy uses 32 networks with 30 molecular and 70 *in vitro* descriptors, and 2 hidden neurons.

