Using Chemical and Biological Descriptors to Develop Predictive Models for Rat Acute Oral Toxicity

J. Fitzpatrick[1], P. Pradeep[1,2], A. Karmaus[3], and G. Patlewicz[1]

[1]US EPA, National Center for Computational Toxicology, Research Triangle Park, NC, United States. [2]Oak Ridge Institute for Science and Education, Oak Ridge, TN, United States. [3]Integrated Laboratory Systems, Inc., Research Triangle Park, NC, United States

Jeremy Fitzpatrick | Fitzpatrick.Jeremy@EPA.gov | ORCID ID:0000-0002-5401-9706

www.epa.gov

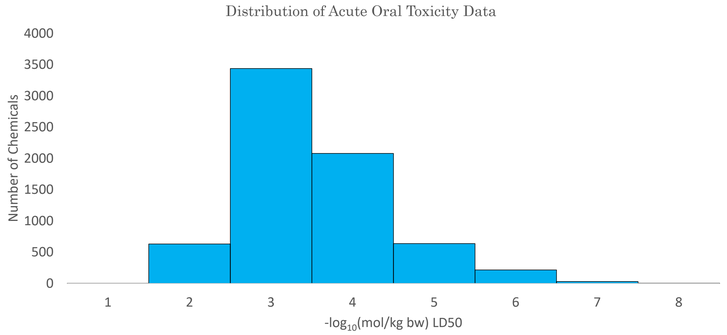
Introduction

- Assessing the acute toxic potential of a substance is necessary to determine the potential effects of accidental or deliberate short-term exposure. There are no accepted *in vitro* approaches available, and few *in silico* models, to predict acute oral toxicity.
- Until recently, a paucity of experimental *in vivo* acute toxicity data was available for model development and evaluation. Here, a large acute oral toxicity dataset totaling 11,992 unique chemicals was compiled.
- Three approaches were used to model acute oral toxicity using ToxCast™/Tox21 activities as biological descriptors as well as chemical descriptors. The first approach was a global random forest classification model, built to predict which substances would have a LD50 above or below 2000 mg/kg bw (body weight). The second was a global random forest regression model built to predict the exact LD50 for a given compound. The third was a set of 15 cluster-based local random forest models built to predict the exact LD50, the k-means algorithm was used to derive the clusters.

Acute Oral Toxicity Data

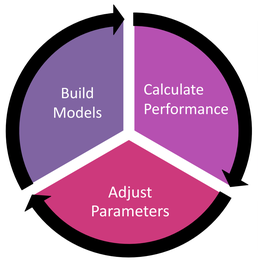
Our dataset consists of data from seven different sources: OECD eChemPortal, ECHA (European Chemicals Agency) ChemProp, NLM (National Library of Medicine) HSDB (Hazardous Substances Data Bank), Leadscope, NLM ChemIDplus via TEST (Toxicity Estimation Software Tool), EU JRC (Joint Research Centre) AcutoxBase and NICEATM PAI (Pesticide Active Ingredients database). The majority of the substances in the set (77%) have a discrete LD50 value. The remaining chemicals have outcomes from limit tests, with the most common limit test reporting a LD50 value above 5000 or 2000 mg/kg bw.

Acute Oral Toxicity Data Set		
Total Entries	Unique Substances	Substances with a Structure in DSSTox
16173	11992	9383



Creating Global Models

We constructed random forest models to predict both the which chemicals had an LD50 above or below 2000 mg/kg bw and to predict the exact LD50. We constructed and optimized our models in R using the gbm and caret packages. We used the ToxCast™ and Tox21 experimental and predicted outcomes as descriptors. A random forest is a collection of decision trees that vote for a given outcome based on a majority rule.

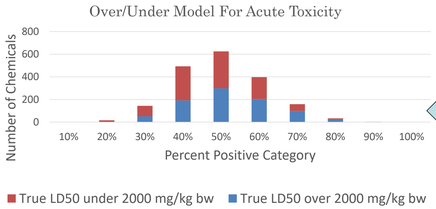


Build Models
- Each model was built using the gbm package in R
- Chemotypes, ToxCast™ and Tox21 experimental and predicted outcomes were used as descriptors
- 10-fold CV (cross-validation) was carried out using the caret package

Calculate Performance
- Accuracy was calculated and averaged over all CVs, for the classification models
- RMSE was calculated and averaged over all cross validations, for the regression models predicting the exact LD50

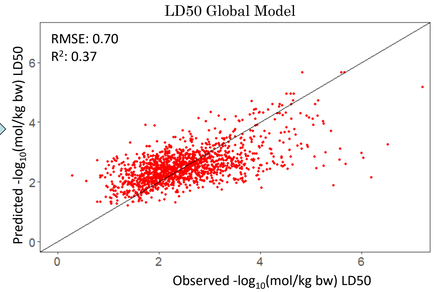
Adjust Parameters
- Results were used to adjust parameters for best fit
- Adjusted parameters were: number of trees, interaction depth, minimum number of samples to split, impact of each tree

Global Modeling Results



- Our model for predicting compounds over and under a LD50 of 2000 mg/kg bw had an accuracy of 57%, a balanced accuracy of 56%, a sensitivity of 57%, and a specificity of 56%.
- Although not shown the model performs significantly better when using the Toxcast™ and Tox21 data if a chemical groups were based on molar LD50s instead mg/kg bw, this is likely because the ToxCast™ and Tox21 data are in molar units.

- Our global model for predicting the exact LD50 performed significantly better than the Y-randomized model which had an R^2 of 0.00.
- The prediction of the LD50 was heavily influenced by the distribution of the LD50 data, see the data distribution graph to the left.



Local Cluster-based Regression Model Results

METHODS

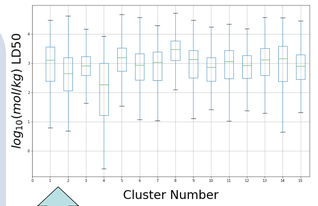
Statistical validation:
5-fold CV on training set, (80%)
External test set validation (20%)

Clustering method:
K-means clustering
Number of clusters, k = 15 (on training set)

Modeling Method:
1. Local (cluster-based) random forest models were developed for each cluster developed (using the training data) using the fingerprints and CDK descriptors as the feature set.
2. The test chemicals were assigned to a pre-determined cluster. Next, each local cluster-based model was used to make a prediction for the test set chemicals

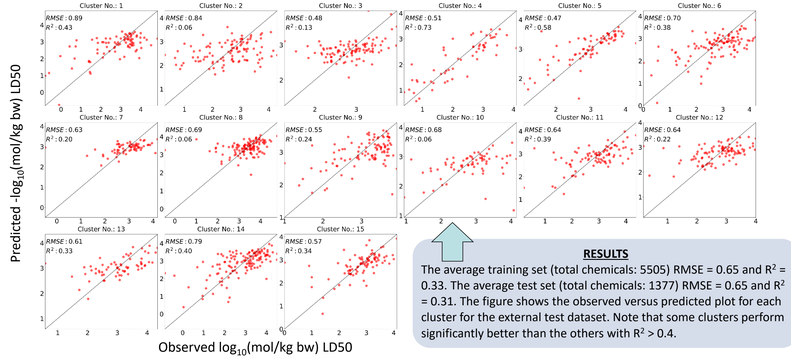
DATA

Chemical Descriptors:
ToxPrint (729) and PubChem (881) Fingerprints (Total bits: 51)
CDK Physchem Descriptors (5)



RESULTS

Boxplots showing the distribution of $\log_{10}$(mol/kg bw) LD50 values in each of the 15 clusters developed using the training dataset.



RESULTS

The average training set (total chemicals: 5505) RMSE = 0.65 and R^2 = 0.33. The average test set (total chemicals: 1377) RMSE = 0.65 and R^2 = 0.31. The figure shows the observed versus predicted plot for each cluster for the external test dataset. Note that some clusters perform significantly better than the others with $R^2 > 0.4$.

Conclusion and Future Steps

- The ToxCast™ and Tox21 assays contain information which are predictive of acute oral toxicity
- The cluster based models performed much better than the global models
- Future work will determine which ToxCast™ and Tox21 assays are the most informative.

References

TEST: https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

ToxCast™ and Tox21 latest data releases available from: https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data

E1071 package for R: Meyer, David, et al. Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. February 2, 2017. Documentation available at: https://cran.r-project.org/web/packages/e1071/e1071.pdf

gbm package for R: Ridgeway, Greg. Generalized Boosted Regression Models. March 21, 2017. Documentation available at: https://cran.r-project.org/web/packages/gbm/gbm.pdf

Caret package for R: Kuhn, Max, et al. Classification and Regression Training. April 18, 2017. Documentation available at: https://cran.r-project.org/web/packages/caret/caret.pdf

*U.S. Federal funds from NIEHS/NIH/HHS contract HHSN273201500010C supported this study