Development of an acute oral toxicity dataset to facilitate assessment of existing QSARs and development of new models

Jeremy Fitzpatrick[1], Agnes Karmaus[2], Prachi Pradeep[1,3], and Grace Patlewicz[1]

1. US EPA, National Center for Computational Toxicology, Research Triangle Park, NC
2. Integrated Laboratory Systems, Inc., Research Triangle Park, NC, USA*
3. Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee

ORCID ID:0000-0002-5401-9706

Jeremy Fitzpatrick 1 Fitzpatrick.Jeremy@epa.gov

United States Environmental Protection Agency

Abstract

Acute oral toxicity data are used to meet both regulatory and non-regulatory needs. Recently, there have been efforts to explore alternative approaches for predicting acute oral toxicity such as QSARs. Evaluating the performance and scope of existing models and investigating the feasibility of developing new models relies on a large set of curated acute toxicity data. We created a data set of rat oral LD50 values for 16439 substances from a variety of sources. We used a subset of this dataset to: 1) evaluate LD50 predictions of two models TIMES and TEST, and 2) investigated the feasibility of developing new models using bioactivity data from ToxCast™ and Tox21. We have processed 1787 substances through both the TIMES and TEST models, finding that 18% of the substances were within the domain of the TIMES model and 94% were within the domain of the TEST model. Our own models have been successful in using ToxCast™ and Tox21 data to predict acute oral toxicity, although testing and refinement is still on going.

Aims

- **Gather Data:** Collect the largest set of rat acute oral toxicity data that is available
- **Evaluate Models:** Evaluate currently available models for the ability to predict acute oral toxicity
- **Create Models:**
  1. Binary Classification
  2. Global Regression
  3. Local cluster-based Regression

Our Data Set

Acute Oral Toxicity Data Set		
Total # Substances	Substances with a discrete LD50 value	Substances with a defined Structure
16909	13073	11236

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 set contains a total of 42726 records. The majority of the substances in the set (77%) have a discrete LD50 value that has been measured. The remaining chemicals have outcomes from limit tests, with the most common limit test reporting a LD50 value above 5000 or 2000 mg/kg.

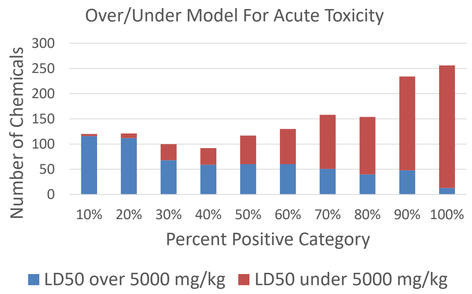
Currently Available Models

Model	Number of substances in dataset	Number of substances that could be predicted	Accuracy for substances with one Value	Accuracy for substances with multiple values	Overall Accuracy
TIMES Model	1787	315 (17.6%)	85 of 93 (91%)	206 of 222 (93%)	291 of 315 (92%)
TEST-Acute Oral Consensus Model	1787	1673 (93.6%)	433 of 490 (88%)	1092 of 1183 (92%)	1525 of 1673 (91%)

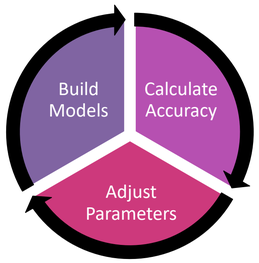
Only discrete organic chemicals were considered for the evaluation of TIMES and TEST. Other substances were found to be outside the scope of what the models were capable of predicting. The majority of substances in the dataset compiled fell outside of the applicability domain of TIMES. In contrast, TEST was able to make predictions for the majority of the dataset. To assess accuracy, we considered a prediction to be accurate if it was within one log value of the median LD50 value or if it was within the values measured in the animal data, whichever was greater. This interval for assessing the accuracy of in silico predictions will be refined further based on ongoing analysis of the variability of the animal data that has been collected.

Binary Classification Model Results

We constructed random forest models to predict each assay endpoint using the chemotypes, 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. Our random forest model could be applied before applying a continuous model to find non-toxic chemicals.



Confidence of an individual prediction can be assessed based on the prediction percentage. This is the proportion of trees which vote as active or in active. If a higher proportion of the trees vote for one outcome over another, it has a greater chance of being the correct prediction.

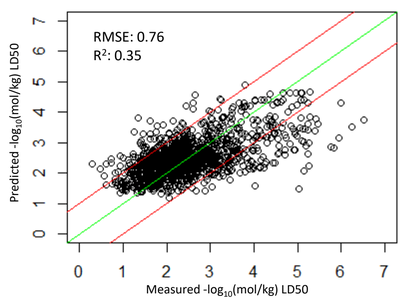


Build Model
- Each model was built using the gbm package in R
- Chemotypes, ToxCast™ and Tox21 experimental and predicted outcomes were used as descriptors
- 5-fold cross validation was carried out using the caret package

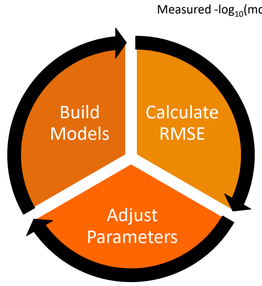
Calculate Accuracy
- Accuracy was calculated and averaged over all cross validations.

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 Regression Model Results



- Our global ridge regression model used both experimental and predicted ToxCast™ and Tox21 assay outcomes as descriptors.
- The model was built using a training set of 4164 discrete organic substances with defined LD50 values.
- The test set consisted of 1387 substances, the prediction of these substances is shown in the graph to the left.
- Ultimately 85% of the substances were found to be within one log unit of their predicted LD50 value.



Build Model
- ToxCast™ and Tox21 experimental and predicted outcomes were used as descriptors
- 5-fold cross validation was carried out using the caret package in R

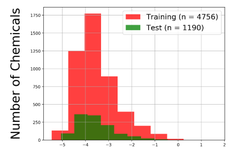
Calculate RMSE
- The RMSE (root mean square error) was calculated and averaged over all cross validations.

Adjust Parameters
- Adjust ridge coefficients, controlling the strength of the penalty terms

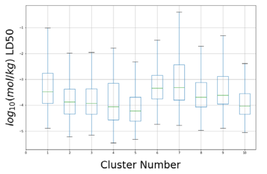
Local Cluster-based Regression Model Results

Data split:
Training = 80%, Test = 20%

K-means clustering:
k = 10 on training set



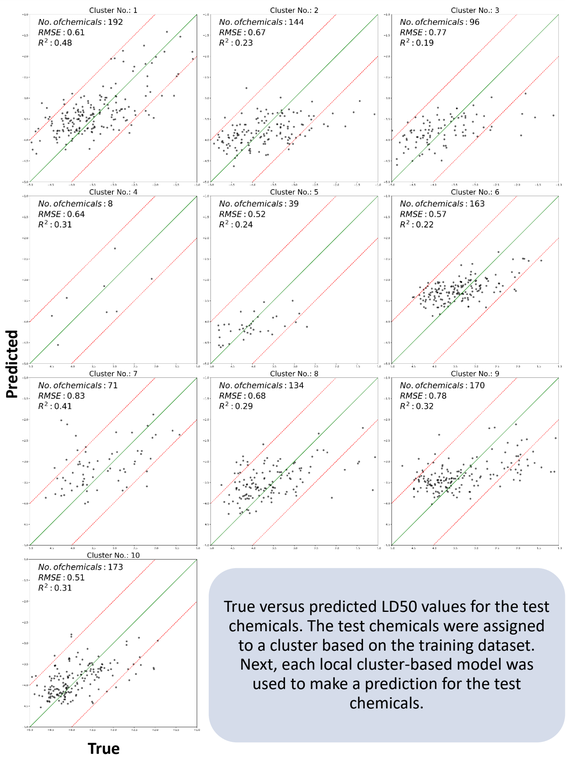
Box plot of the range of log10(mol/kg) LD50 values within each cluster.



Model Development:

Fingerprint: ToxPrints
Physchem Descriptors: 10

Random forest models were built for each cluster using the training dataset and 5-fold cross validation



True versus predicted LD50 values for the test chemicals. The test chemicals were assigned to a cluster based on the training dataset. Next, each local cluster-based model was used to make a prediction for the test chemicals.

Conclusions and Future Steps

Conclusions

- The domain of TEST is much larger than that of TIMES for acute oral toxicity predictions
- ToxCast™ and Tox21 assays contain information which are predictive of acute oral toxicity

Future Steps

- Extend assessment of currently available expert systems to the larger curated dataset
- Finalize variability assessment of the animal data
- Finalize assessment of our own models
- Compare performance of our models against available expert systems
- Compare performance of out models against the variability of the animal data

References

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

TIMES: http://oasis-lmc.org/products/software/times.aspx

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

MASS package for R: Ripley, Brian, et al. Support Functions and Datasets for Venables and Ripley's MASS. April 21, 2017. Documentation available at: https://cran.r-project.org/web/packages/MASS/MASS.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

U.S. Environmental Protection Agency
Office of Research and Development