A Large Dataset of Acute Oral Toxicity Data Created for Testing in Silico Models

Jeremy Fitzpatrick and Grace Patlewicz

US EPA, National Center for Computational Toxicology Research Triangle Park, NC

ORCID ID:0000-0002-5401-9706

Jeremy Fitzpatrick 1 Fitzpatrick.Jeremy@epa.gov

United States Environmental Protection Agency

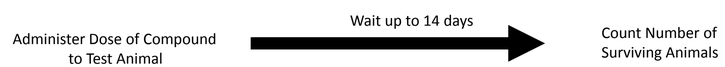
Abstract

Acute toxicity data is a common requirement for substance registration in the US. Currently only data derived from animal tests are accepted by regulatory agencies, and the standard in vivo tests use lethality as the endpoint. Non-animal alternatives such as in silico models are being developed due to animal welfare and resource considerations. We compiled a large dataset of oral rat LD50 values to assess the predictive performance of currently available in silico models. Our dataset combines LD50 values from five different sources: literature data provided by The Dow Chemical Company, REACH data from eChemportal, HSDB (Hazardous Substances Data Bank), RTECS data from Leadscope, and the training set underpinning TEST (Toxicity Estimation Software Tool). Combined these data sources yield 33848 chemical-LD50 pairs (data points), with 23475 unique data points covering 16439 compounds. These chemical-LD50 pairs have been combined with assay results from Toxcast to build a random forest model. The random forest model will help us to assess which Toxcast assays are relevant in predicting acute oral toxicity. The entire dataset of chemical-LD50 pairs was also loaded into a chemical properties database. All of the compounds were registered in DSSTox and 59.5% have publically available structures. Compounds without a structure in DSSTox are currently having their structures registered. Future work will use the structural data to evaluate the predictive performance and applicable chemical domains of three in silico models (TIMES, PROTOX, and TEST).

Background

Acute toxicity is a difficult endpoint to predict. Even animal models for acute oral toxicity can give a wide range of results due to a variety of factors. Currently we are focusing on compiling studies for rat acute oral toxicity. A typical rat study is conducted with either single or multiple doses over a 24 hour period. The dose may be administered in number of ways and using a number of substrates which can complicate the outcome when comparing across studies. The LD_{50} in acute oral toxicity studies in rats is the amount needed to kill half of the population within 14 days of being administered.

Acute Oral Toxicity Testing in Rats



Aims

Aim 1: Compile a Large Dataset of Oral Toxicity Studies From Rats
- Gather acute oral toxicity from multiple sources

Aim 2: Evaluate the Predictivity of Toxcast Assays in Predicting Acute Oral Toxicity
- Construct a random forest model using LD_{50} and Toxcast assay data to determine which clusters of Toxcast assays are the most relevant in predicting acute oral toxicity

Aim 3: Gather Structures and Evaluate Current *in silico* Models
- Collect several thousand structures of compounds with acute oral toxicity data
- Compare the predictivity for LD_{50} of currently available QSAR models for acute oral toxicity and determine their applicability domains based on structure

Overview of Current Data

Our dataset consists of data from five different sources: DOW, eChemportal, HSDB (Hazardous Substances Data Bank), Leadscope, and TEST. The DOW dataset was compiled and shared with us by researchers at DOW and contains 3790 records for rat acute oral toxicity. The eChemportal dataset was compiled from multiple searches of eChemportal for the acute oral toxicity end point in rats and contained 7284 records. The HSDB dataset held 584 records for acute oral toxicity in rats. The Leadscope and TEST datasets are the datasets used by each program to train their acute oral toxicity models and contained 3554 and 13674 records for rat acute oral toxicity respectively. Together this gives us approximately 33848 data points on acute oral toxicity with 23475 unique data points as well as information on 16439 unique chemicals.

Overview of Current Data

Unique CAS and Unique Data Points		
Database	Number of Unique CAS in Dataset	Number of Unique LD50 records for Rat AOT
DOW	8355	3790
eChemportal	3499	7284
HSDB	586	584
Leadscope	4779	3554
TEST	13548	13674

CAS was the most commonly used identifier amongst the datasets and was used for the initial comparisons. Within the individual datasets the number of unique records differed significantly, due to several factors including lacking data for a chemical in the set, having several LD_{50} values for a chemical in the set, or having a variety of species.

A comparison of chemical overlap amongst the datasets. In addition 90 compounds were found to overlap over all 5 datasets.

Number of Overlapping Chemicals					
	DOW	eChemportal	HSDB	Leadscope	TEST
DOW	XXXXX	827	453	4479	3556
eChemportal	827	XXXXX	126	623	766
HSDB	453	126	XXXXX	442	552
Leadscope	4479	623	442	XXXXX	3351
TEST	3556	766	552	3351	XXXXX

Number of Overlapping Data Points					
	DOW	eChemportal	HSDB	Leadscope	TEST
DOW	XXXXX	264	22	1007	1272
eChemportal	264	XXXXX	7	115	157
HSDB	22	7	XXXXX	99	156
Leadscope	1007	115	99	XXXXX	3095
TEST	1272	157	156	3095	XXXXX

A comparison of the overlap of unique compound LD50 pairs in each dataset to find instances of the same pair being reported in multiple datasets. When comparing all 5 datasets it was found that no chemical LD_{50} pair was observed across all 5.

Available Structures

Data Source	Number with DSSTox Structure	Number No Structure in DSSTox
eChemPortal	2115	1374
HSDB	579	5
Leadscope	4769	11
TEST	9021	4527
DOW	8152	203

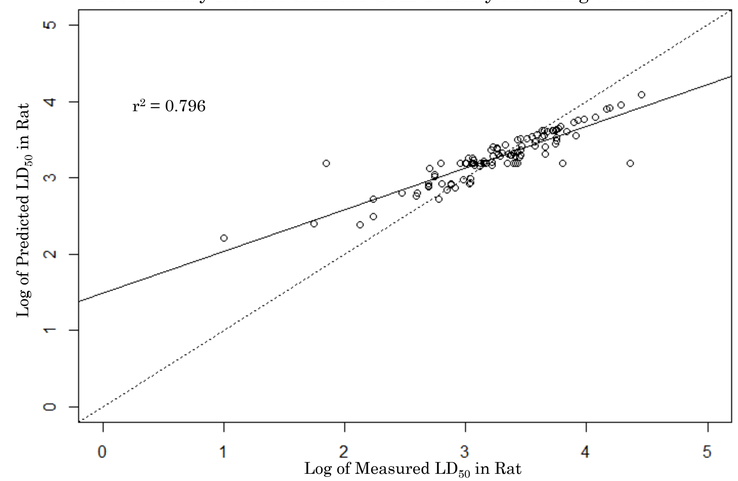
Number of compounds in each dataset that have a single structure available in DSSTox. It should be noted that compounds that are mixtures or polymers are also listed as having no structure, since a single structure is not available for them.

Random Forest Modeling

Using LD_{50} values from acute oral toxicity studies performed in rats taken from the ECHA database via eChemportal we found that 530 overlapped with chemicals that had been tested in a significant number of Toxcast assays. These were divided randomly into 5 equally sized sets, with the random seed set to one million. Four of the sets were then combined and used to train the random forest model. The fifth set was used to test the model. The model was built using the 'randomForest' package in R with 10000 trees and any missing information was filled in using the average result for the Toxcast assay.

The results of the test set in the graph below indicate that the Toxcast data are correlated with acute oral toxicity values. However it should be noted that these results are preliminary, as it will be necessary to preform repeated cross-validation to confirm the results. New chemicals will also be added to the model as we now have approximately 4000 chemicals with rat LD_{50} values and Toxcast assay data.

Preliminary Results of Acute Oral Toxicity Modeling with Toxcast



Conclusions

Aim 1: Compile a Large Dataset of Acute Oral Toxicity Studies From Rats
- 5 different datasets of acute oral toxicity data in rats have been collected and compiled from The Dow Chemical Company, REACH data from eChemportal, HSDB (Hazardous Substances Data Bank), RTECS data from Leadscope, and the training set underpinning TEST (Toxicity Estimation Software Tool)

Aim 2: Evaluate the Predictivity of Toxcast Assays in Predicting Acute Oral Toxicity
- Based on the limited modeling so far there is a correlation between Toxcast data and acute oral toxicity in our set of 530 chemicals.
- More work is needed to determine which Toxcast assays are the most informative.

Aim 3: Use These to Evaluate Current *in silico* Models
- Structures have been assigned to 9781 compounds
- Future work will use these structures to compare *in silico* acute oral toxicity models

References

A. Wallace Hayes (editor), Principles and Methods of Toxicology. Raven Press, New York, NY. Second Edition 1989

Leo Breiman, Adele Cutler, Andy Liaw, Matthew Wiener, Package 'randomForest' version 4.6-12, October 7, 2015 Availible at: https://www.stat.berkeley.edu/~breiman/RandomForests/