THE ROLE OF FEATURE SELECTION AND STATISTICAL WEIGHTING IN PREDICTING IN VIVO TOXICITY USING IN VITRO ASSAY AND QSAR DATA

Jessica Wignall[1], Matthew Martin[2], Joanne Trgovcich[3], Arun Varghese[3] | [1]ICF International, Fairfax, VA, [2]US EPA, National Center for Computational Toxicology, Durham, NC, [3]ICF International, Durham, NC

Abstract

Our study assesses the value of both in vitro assay and quantitative structure activity relationship (QSAR) data in predicting in vivo toxicity using numerous statistical models. The models were built on datasets of (i) 586 chemicals for which both in vitro and in vivo data are currently available in EPA's Toxcast and ToxRefDB databases, and (ii) 769 chemicals for which both QSAR data and in vivo data exist. Similar to a previous study (based on just 309 chemicals, Thomas et al. 2012), after converting the continuous values from each dataset to binary values, the majority of more than 1,000 in vivo endpoints are poorly predicted. Even for the endpoints that were well predicted (about 40 with an F1 score of >0.75), class imbalances in in vivo endpoint data or cytotoxicity across in vitro assays may be skewing results. We investigated whether use of best practices for data preprocessing and model fitting in real-world contexts would improve model predictions. This included options for dealing with missing data, including omitting observations, aggregating variables, and imputing values. We also examined the impacts of feature selection (from both a statistical and biological perspective) on performance and efficiency, and we weighted outcome data to reduce endpoint imbalances to account for potential chemical selection bias and assessed revised performance. For example, initial weighting strategies decrease the number of models with an F1 score >0.75 drastically (to 6), but these models are more able to predict nontoxic chemicals in certain contexts. The results of these analyses can be used to inform screening or other decisions, especially in the context of future data enhancements, such as more biologically relevant in vitro assays, additional in vivo endpoint data, and extension of chemical space.

Methods

COLLECT DATASETS
- Collect in vivo, in vitro, and in silico datasets from publicly available resources: ToxRefDB, ToxCast, and OpenBabel, respectively.

PROCESS DATA
- Using R Statistical Program
- **Convert all data to binary datasets**
- Impute missing data across a chemical based on activity

BUILD MODELS
- Statistical approaches include:
  - K Nearest Neighbors
  - Bernoulli NB
  - Linear SVC
  - Logistic Regression
  - Decision Tree
  - Random Forest

OTHER STATISTICAL PROCESSING
- Optimized on F1
- Feature selection
- Outcome weighting

BIOLOGICAL PROCESSING
- Combining "like" endpoints across species and study design

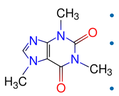
MACHINE LEARNING
- Models built using Python
- Identifies the best performing of multiple machine learning algorithms
- Builds predictive binary outcome toxicity screening models
- Measures performance based on 5-fold cross-validation

Figure 1. Illustration of Methods and Process Used to Build the Toxicity Prediction Model

1 ToxCast Data

- **Publicly available high throughput screening data from EPA on over 1800 chemicals**
- **Number of assay predictors = 711**

2 Chemical Descriptors



- **~115,000 chemicals from EPA EpiSuite™/SHEDS-HT**
- **Converted SMILES to 1024 binary fingerprint 2D descriptors calculated using Python and OpenBabel**
- **Validated by predicting chemical properties (e.g., vapor pressure)**

3 ToxRefDB Data

- **EPA's ToxRefDB includes detailed study design, dosing, and observed treatment-related effects using standardized vocabulary curated from in vivo animal toxicity studies**
- **Includes detailed study and effect information on hundreds of chemicals primarily pesticide active ingredients**
- **Across the studies, 1045 endpoints available for analysis**
- **October 2014 Release**

4 Training Set for Machine Learning

- Joined ToxCast assay data to ToxRefDB endpoint data using CAS RN

1. **Assay Only Training Set** includes chemicals with both ToxCast and ToxRefDB data (586)
2. **QSAR Only Training Set** includes chemicals with both Chemical Fingerprint and ToxRefDB data (769)
3. **Assay + QSAR Training Set** includes chemicals with ToxCast, Chemical Fingerprint, and ToxRefDB data (538)

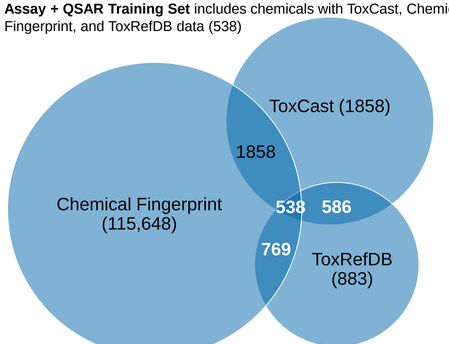


Figure 2. Source and Number of Chemicals used in the Toxicity Prediction Model

Toxicity Prediction Model Features

Data Pre-Processing

- **Imputing:** Smooths out data sets; results in larger training set than otherwise possible, but introduces assumptions to data

Statistical Measures of Fit:

- **F1 Score:** Measure of model accuracy that considers both precision and recall; optimizing on this metric minimizes False Negatives but sacrifices accuracy (ideal for screening)
- **Area under the curve of the receiver operating characteristic (ROC-AUC):** Combines specificity and sensitivity; often used as a metric to communicate predictive performance
- **Predictive improvement over random baseline (pseudo-R2):** Shows improvement in model performance over predictions based on existing ratio in data only (i.e., how much does the model improve predictions?)

Combining Endpoint Outcome Data: Assumes that an effect in one species/study design is relevant to all species/study design

Weighting

- Outcome data can be weighted to account for potential chemical selection bias or endpoint imbalances in the training datasets.
- For example, if input data is heavily weighted as toxic or non-toxic, then weighting data to counteract skewed datasets should result in more useful predictions, as our findings indicate.

Figure 3. Weighting Assay Data for Predicting Unknown Chemicals.
Green bar portion represents predicted active chemicals, yellow bar portion represents predicted inactive chemicals. Blue, orange and gray lines represent F1, ROC-AUC, and R2 data generated by the model.
In vivo endpoint models (bars) include the following:
(1) Systemic Carcinogenic: adult-OtherSystemic-InLifeObservations-BodyWeight
(2) Systemic Carcinogenic: adult-OtherSystemic-InLifeObservations
(3) Systemic Carcinogenic: adult-OtherSystemic
(4) Systemic Carcinogenic: adult
(5) Systemic Carcinogenic

Combined Endpoints

Combining endpoints results in predictive models for overall in vivo endpoints, not species or study design-specific.

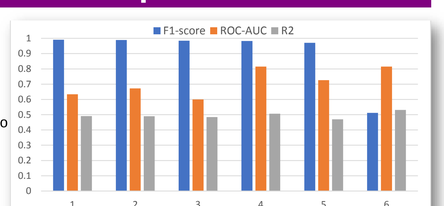


Figure 4. Combining Endpoints Results In Predictive Models For In Vivo Prediction.
Orange bars, F1 score; Gray bars, ROC-AUC; blue bars, R2. Data are shown for 6 endpoints as follows:
(1) SystemicCarcinogenic; (2) SystemicCarcinogenic_adult; (3) SystemicCarcinogenic_adult_OtherSystemic; (4) SystemicCarcinogenic_adult_OtherSystemic_InLifeObservations; (5) SystemicCarcinogenic_adult_OtherSystemic_InLifeObservations_BodyWeight; (6) Cholinesterase

Model Performance

Application of Feature Selection Increases the Number of Models that Predict in Vivo Toxicity

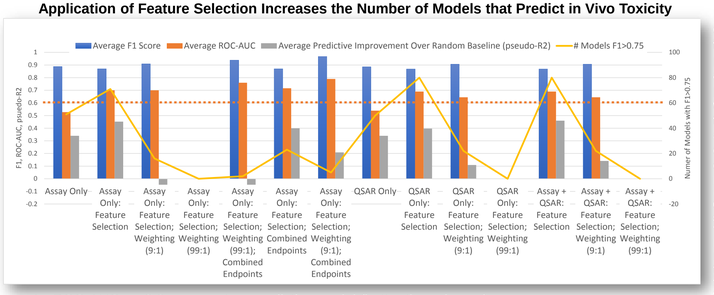


Figure 5. Comparison of toxicity prediction after data preprocessing, weighting, feature selection, and combining endpoints. The performance of models with F1 > 0.75 within each scenario is shown. The dotted orange line highlights the F1 score of 0.60 for comparison. If there is no data in the bar graph, no models passed the threshold of F1 score >0.75.

- An ROC-AUC > ~0.6 represents a predictive model (better than chance).
  - 9/14 scenarios have an average ROC-AUC > 0.6
  - 3/14 have an average ROC-AUC > 0.7.
- The number of models with an F1>0.75 and an ROC-AUC>0.6 maxes out after feature selection and before weighting, and do not have a relationship with the pseudo-R2.

Feature Selection

- We combined endpoint data from ToxRefDB across species and study design for a given endpoint and analyzed feature selection to identify the most "predictive" features of the ToxCast assay set. We analyzed Selected Features (ToxCast assays) across all models. Those assays found to be most predictive across all models are shown in Table 1. These assays are specific inasmuch as they fail to be identified as predictive if in vitro data are randomized in this model.
- This Table includes 8 of the 10 ToxCast Assays that were the most predictive for carcinogenicity studies, and 9 of 10 assays most frequently used in liver-related models.
- These assays in Table 1 were performed in liver cells, or relate to function of the liver in detoxification of xenobiotic compounds, suggesting that in vitro assays targeting hepatic endpoints are predictive of in vivo activity.

Table 1. Top 10 Selective Features (ToxCast Assays) Across All Models

ToxCast assay	ToxCast biological process target (assay type)	ToxCast intended target sub family	ToxCast intended target gene	Gene Info
NVS_NR_hPXR	receptor binding (cell free nuclear run on assay)	non-steroidal	Human NR1I2	Transcriptional regulator of the cytochrome P450 gene CYP3A4
APR_MitoMass_72h_up	cell cycle (mitochondrial morphology in HepG2 cells)	organelle conformation	NA	NA
APR_NuclearSize_24h_up	cell cycle (nuclear morphology in HepG2 cells)	organelle conformation	NA	NA
NVS_ADME_rCYP3A1	regulation of catalytic activity (cell free enzyme activity assay)	xenobiotic metabolism	Rat Cyp3a23/3a1	Encodes a member of the cytochrome P450 superfamily of enzymes. Steroid-inducible member of p450 subfamily 3A
NVS_MP_rPBR	receptor binding (cell free ligand binding)	cholesterol transporter	Rat Tspo	Benzodiazapene receptor that may be involved in the neoadal response to hypoxia.
NVS_ADME_hCYP1A2	regulation of catalytic activity (cell free enzyme activity assay)	xenobiotic metabolism	Human CYP1A2	Encodes a member of the cytochrome P450 superfamily of enzymes.
NVS_ADME_rCYP2A1	regulation of catalytic activity (cell free enzyme activity assay)	xenobiotic metabolism	Rat Cyp2a1	Encodes a member of the cytochrome P450 superfamily of enzymes with testosterone 7 alpha-hydroxylase activity.
NVS_ADME_rCYP3A2	regulation of catalytic activity (cell free enzyme activity assay)	xenobiotic metabolism	Rat Cyp3a2	Encodes a member of the cytochrome P450 superfamily of enzymes; catalyzes the conversion of testosterone to 6-beta-hydroxytestosterone.
NVS_MP_hPBR	receptor binding (cell free ligand binding)	cholesterol transporter	Human TSPO	Present mainly in the mitochondrial compartment of peripheral tissues.
APR_MitoMass_24h_up	cell cycle (mitochondrial morphology in HepG2 cells)	organelle conformation	NA	NA

Conclusions and Future Research

- We provide a method for building successful models of binary predictions of in vivo activity for screening or prioritization applications, including data preprocessing, use of machine learning, statistical modelling, and biological analysis.
- Additionally, we determined that several factors can improve utility of the models, including weighting of input data according to activity or inactivity of chemicals, combining in vivo endpoints across species, and feature selection.
- Weighting decisions should be made based on expected ratios of active to inactive chemicals in the prediction dataset of interest, though this information will not always be available.
- Feature selection analysis may point to biologically-relevant rationales for inclusion of in vitro assays in these models.
- These findings suggest that predictive performance can be improved through statistical, toxicological, and biological advances. Future work to improve model performance can include (1) improving input data to include biologically-relevant in vitro assay data (e.g., metabolically-competent assays) (2) developing methods to generate more homogenous in vivo data sets (e.g., collecting additional in vivo dose-response data can supplement data gaps). The use of multinomial representations of in vitro predictors to capture threshold effects in assay response.

For more information: Jessica Wignall | jessica.wignall@icfi.com
Arun Varghese | arun.varghese@icfi.com