National Institute of Environmental Health Sciences
Division of Translational Toxicology

Characterizing the Impacts of Assay Design on Cytotoxic Concentration Range

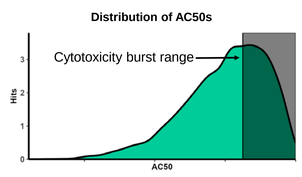
K To[1], AL Karmaus[1], V Hull[1], D Allen[1], N Kleinstreuer[2]

[1]Inotiv, RTP, NC; [2]NIH/NIEHS/DTT/NICEATM, RTP, NC

Introduction

- The Tox21 and ToxCast high-throughput screening (HTS) programs have provided an abundance of in vitro assay data.
- Chemicals evaluated in these HTS assays are evaluated for occurrences of "cytotoxicity burst." This is a change in activity at a concentration threshold where cytotoxicity and generalized cell stress are likely to impact measured assay activity and confound hazard estimates.
- Evaluating cytotoxicity for burst activity provides additional context when interpreting bioactivity data by highlighting assay results of questionable validity.
- Bioactivity data in HTS assays is often expressed as AC50: the concentration of test chemical that induces 50% of the maximum assay activity. The range of AC50s observed for a given chemical could be impacted by assay design or physicochemical characteristics of the tested chemical.
- The parameters that are associated with heightened cytotoxicity could be used to improve predictions of cytotoxicity burst activity.

Hypothetical Assay Activity Showing Cytotoxicity Burst Range



Objectives

- Use machine learning to identify variables, including assay design parameters and physicochemical characteristics, that could inform on predicting cytotoxicity which would ultimately provide additional context for evaluating data validity when analyzing HTS data.
- As a proof of concept, a limited data set comprising data rich chemicals and assays were used to fit an ensemble tree model for predicting positive or negative hit calls.
- By using chemicals that have a low frequency of missing data, the parameters that maximize accuracy will provide context for future model development.

Data Selection and Processing

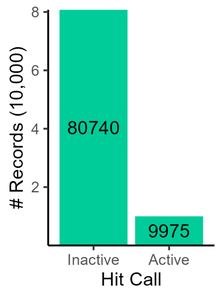
- Cytotoxicity assay data were obtained from the U.S. EPA's *invitrodb.v3.4* using the tcpl R package.
- The selected cytotoxicity assays were reviewed by subject matter experts to include only the data where the AC50s were derived from endpoints analyzed in the expected direction.
- Quantitative structure-activity relationship predictions for 13 chemical properties generated by OPERA (https://ntp.niehs.nih.gov/go/opera) were obtained from the Integrated Chemical Environment (ICE, https://ice.ntp.niehs.nih.gov/).
  - Missing data for pH-dependent lipid-aqueous partition coefficients and acid dissociation constants were imputed with "20."
- Nine assay design parameters were obtained from tcpl.
  - Missing data for *key assay reagent type* and *cell growth mode* were imputed with "N/A" and treated as unique categories for the given variable.
- Data were removed if any of the following conditions were met:
  - Flag indicating data were of questionable quality.
  - For a negative hit call, the highest tested concentration was less than 1.9 log(µM).
  - Fewer than 500 chemicals were tested using an assay (all data for that assay removed).
  - Chemicals were tested in fewer than 90% of remaining assays (all data for that chemical removed).
  - Chemicals had no available property data (experimental or predicted by OPERA, see below).
- **After filtering, there were 90,715 remaining assay results for 1,078 chemicals, spanning 89 unique assay endpoints.**

Variables Selected for Modeling

Assay Design Parameter	OPERA Chemical Property Prediction

• Cell growth mode
• Content readout type
• Detection technology type
• Key assay reagent type
• Maximum tested concentration
• Minimum tested concentration
• Organism
• Timepoint (hour)
• Tissue
	• Boiling point
	• Henry's Law constant
	• Melting point
	• Molecular weight
	• Negative log of acid dissociation constant:
	  • Number of ionizations (pKa ionizations)
	  • Strongest acidic pKa (pKa acidic)
	  • Strongest basic pKa (pKa basic)
	• K(OA): octanol-air partition coefficient
	• Octanol-water distribution coefficient (LogD) at pH 5.5
	• Octanol-water distribution coefficient (LogD) at pH 7.4
	• LogP: octanol-water partition coefficient
	• Vapor pressure
	• Water solubility

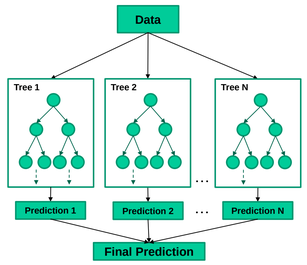
Summary of Hit Calls After Filtering



Methods

- The 90,715 cytotoxicity assay results were split into training and test sets using an 80/20 split.
- An extreme gradient-boosted tree model for predicting assay hit calls was tuned using the training data.
- The final model was selected by evaluating the test set accuracy.
- Results from the trained model were used to identify parameters that could distinguish between positive or negative hit calls.
- Splitting decisions for the variables with the highest importance were visualized to evaluate the association of split locations with positive and negative hit calls.

Ensemble Tree Model



General structure of a model with N trees. Circles represent splits for a single variable.

Model Results

- The final model predicted assay hit calls for both the training and test datasets with accuracy of 95%, sensitivity of 74%, and specificity of 98%.

Confusion Matrix for Training Set Predictions

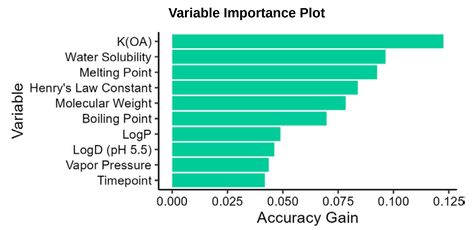
Predicted \ Actual	Negative	Positive
Negative	63506	2115
Positive	1076	5875

Confusion Matrix for Test Set Predictions

Predicted \ Actual	Negative	Positive
Negative	15879	523
Positive	279	1462

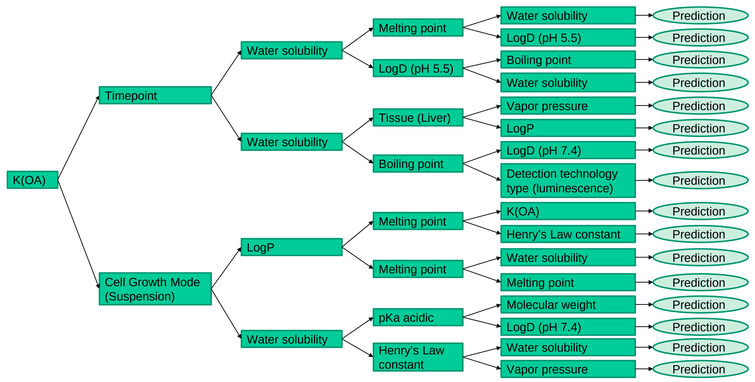
- Variable importance was measured by the gain in prediction accuracy achieved by introducing a split on that variable in a decision tree.
- Among the ten variables with the highest gain, the first nine were OPERA-predicted variables and the tenth was the assay timepoint.

Accuracy Gain for 10 Variables With Highest Importance



- The final model comprised 1000 individual decision tree models, each using different combinations of parameters and splitting decision rules.
- A summary of the split decisions shows that most splits were performed with chemical property parameters.
- Although the assay design parameters have lower importance than the chemical property parameters, 998 individual decision trees used at least one assay design parameter to split the data.

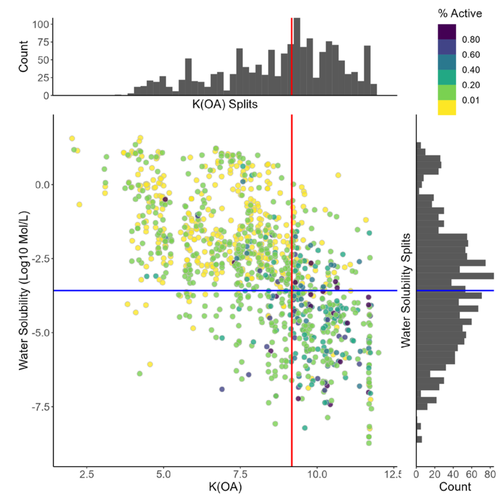
Summary Ensemble Tree Showing Most Frequent Node Splitting Decisions Across All Model Trees



Visualizing Splitting Decisions

- The scatterplot shows the proportion of positive hit calls per chemical, with each point plotted by the given chemical's K(OA) and water solubility.
- The horizontal axis is annotated by the distribution of splitting decisions made with K(OA) across all trees in the model. The median split is indicated by a red vertical line.
- The vertical axis is annotated by the distribution of splitting decisions made with water solubility across all trees in the model. The median split is indicated by a blue horizontal line.
- Chemicals with lower K(OA) and water solubility have higher proportion of negative hit calls.
- The ranges of split locations for K(OA) and water solubility show how decision tree splits maximize prediction accuracy, but additional parameters can provide more information for characterizing cytotoxicity.

% Positive Hit Calls per Chemical



Conclusion

- We trained an extreme gradient-boosted tree model for logistic regression using 13 OPERA-predicted chemical properties and nine assay design parameters.
- When using data with low missingness, the extreme gradient boosted tree predicts chemical-assay cytotoxicity hit calls with high accuracy for both the training and test data.
- Evaluation of variable importance and representative tree diagrams show that most decision tree splits are performed on chemical property variables and that the inclusion of assay design parameters can help to further discretize chemical-assay pair predictions.
- Visualization of the data shows how the splitting decisions attempt to maximize accuracy.

Future Directions

- Based on the accuracy of this proof-of-concept model, an extreme gradient-boosted tree model will be trained with less stringent filtering criteria on the cytotoxicity assays using the full HTS database.
- In order to characterize cytotoxic ranges, future model development will include consideration of tested concentrations by using individual chemical-assay data points and/or points of departure.
- Additional assay design parameters will be identified for inclusion in the model to broaden coverage for assays without paired cytotoxicity data and to assure coverage in cases where assays have a high frequency of missing metadata.

References

Chen T, Guestrin C. 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Judson R et al. 2016. Editor's highlight: Analysis of the effects of cell stress and cytotoxicity on in vitro assay activity across a diverse chemical and assay space. Toxicol Sci 152(2):323–339.

Mansouri K et al. 2018. OPERA models for predicting physicochemical properties and environmental fate endpoints. J Cheminform 10, 10. https://doi.org/10.1186/s13321-018-0263-1.

Acknowledgments

We thank Catherine Sprankle, Inotiv, for editorial input. Technical support was provided by Inotiv under NIEHS contract HHSN273201500010C.

The views expressed above do not necessarily represent the official positions of any federal agency.