Ensemble QSAR Modeling to Predict Multispecies Fish Toxicity Points of Departure

Thomas Sheffield, Richard Judson
US EPA, Office of Research and Development

Thomas Sheffield | email: sheffield.thomas@epa.gov | 919-541-5781

Introduction

Goal: Prioritize chemicals for further evaluation by estimating acute and chronic points of departure in fish using QSAR (quantitative structure activity relationship) models

Approach

- Gather and clean ECOTOX[1] and ECHA[2] (European Chemicals Agency) data on points of departure in fish
- Develop two models according to similar points of departure: acute LC_{50} and acute/chronic NOEC/LOEC/LC_0/MATC (called the "NOEC" model)
- Add PaDEL[3] chemical fingerprints, OPERA[4] physical chemistry properties, NCBI taxonomy data[5], and experimental covariates as features
- Build a stacked ensemble of machine learning models for each data set to predict specified endpoints for novel chemicals as well as uncertainty estimates

Data

- ECOTOX contributed 85,634 (89%) studies after cleaning, while ECHA contributed 10,226 (11%) studies
- The final LC_{50} model contained 34,645 experiments, 2,656 chemicals, and 358 species
- The final NOEC model contained 14,484 experiments, 1,926 chemicals, and 221 species
- 33% and 35% of chemicals have only one entry in the LC_{50} and NOEC data, respectively
- 84% and 85% of chemicals have ten or fewer entries in the LC_{50} and NOEC data, respectively
- Rainbow trout, bluegills, and fathead minnows account for 43% and 50% of all studies in the LC_{50} and NOEC data, respectively
- The mean standard deviation of all chemicals' endpoints with ten or more studies was 0.53 and 0.78 $\log_{10}$(mg/L) in the LC_{50} and NOEC data, respectively

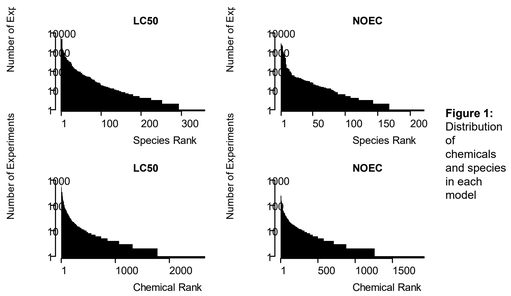


Figure 1: Distribution of chemicals and species in each model

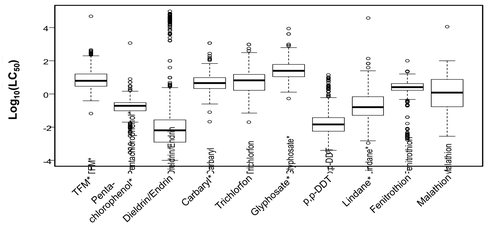


Figure 2: Distribution of endpoints for the most common chemicals in the LC_{50} data set. Asterisks indicate desalted and stereoisomer groups referred to by parent name.

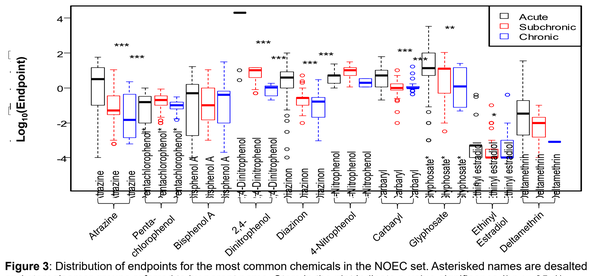


Figure 3: Distribution of endpoints for the most common chemicals in the NOEC set. Asterisked names are desalted and stereoisomer groups referred to by parent name. Stars in the plot indicate p-value significance (* p<=.05, ** p<=.01, *** p<=.001) using the two sample Wilcoxon test versus the acute distribution.

Data Preprocessing

- Study covariates, such as species, endpoint type, study type, study duration class, exposure route, and endpoint units, had to be standardized
- Rare, incongruous, or suspect experiment types were omitted
- Salts and stereoisomers were merged
- General taxonomy classifications were added as features, and only species in Actinopterygii were considered
- Features that were correlated, duplicated, near-constant, uncorrelated with the endpoint, or multicollinear were eliminated
- Studies with the same endpoint type, duration class, exposure route, study type, taxonomy groups, and chemical were merged into one experiment group; modeling was performed at the experiment group level.

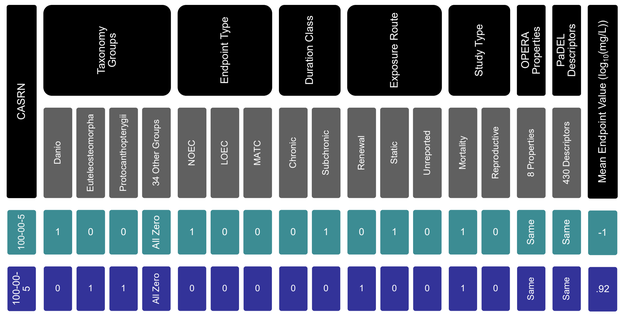


Figure 4: First two rows of the NOEC model matrix. (Note: Some covariates, such as the acute duration class and LC_0 endpoint, are encoded by leaving alternatives equal to zero.)

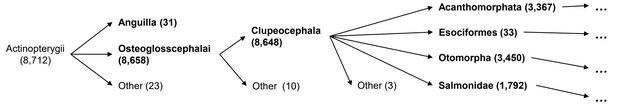
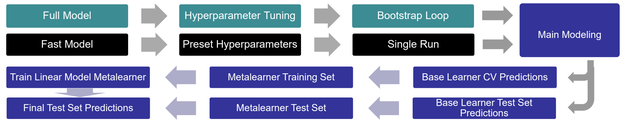


Figure 5: First levels of the LC_{50} model's taxonomy tree. Names in bold indicate categories explicitly named as model features. Numbers in parenthesis indicate how many experiment groups belong to that category.

Modeling



- Three base learners were used: gradient boosted trees[6], random forest[7], and support vector regression[8]
- Base learners were stacked using a linear regression metalearner
- "Full model" used 100 bootstrapped fits with noisy endpoints and tuned hyperparameters
- "Fast model" used a single fit and default hyperparameters
- 20% of data was set aside for external validation (EV); remaining training set underwent 5-fold cross-validation (CV)
- "Combined error" compares full set of endpoints to CV and EV predictions
- Standard deviation of error (σ_E) gives nearly the same values as root-mean-square-error (RMSE) and allows 95% confidence intervals to be computed

Results

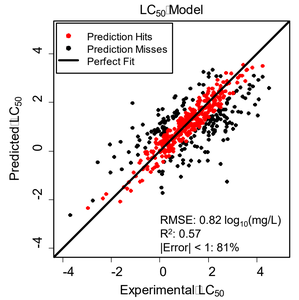


Figure 6: Scatter plot of the LC_{50} full model external validation set. "Hits" are those points for which one standard deviation in the bootstrapped predictions overlaps with the experimental value's average standard deviation based on chemicals with ten or more entries. 63% are hits.

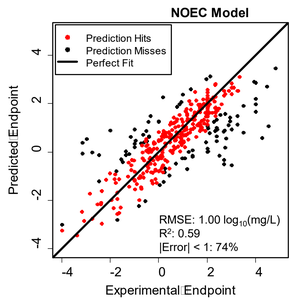


Figure 7: Scatter plot of the NOEC full model external validation set. "Hits" are defined as in Figure 6. 75% are hits, partly due to the NOEC experimental data's greater uncertainty.

Model	Method of Action	# Chemicals	R^2	RMSE	σ_E lower	σ_E	σ_E upper	|Error| < 1
LC_{50} All Studies	Narcotic	422	0.65	0.63	0.59	0.63	0.67	89.8%
	Non-narcotic	2234	0.57	0.87	0.84	0.86	0.89	79.4%
NOEC Mortality Studies	Narcotic	271	0.60	0.85	0.78	0.85	0.93	84.0%
	Non-narcotic	1655	0.58	1.01	0.97	1.00	1.04	72.3%
NOEC Growth & Reproductive Studies	Endocrine Active	110	0.16	1.03	0.89	1.03	1.23	71.6%
	Endocrine Inactive	659	0.59	1.02	0.93	1.02	1.11	74.7%

Table 1: Combined error performance of the fast model on chemicals with selected modes of action.

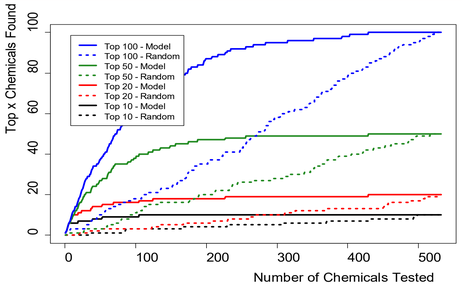


Figure 8: Prioritization performance versus random chance when searching for the 10, 20, 50, and 100 most potent compounds with the full LC_{50} model.

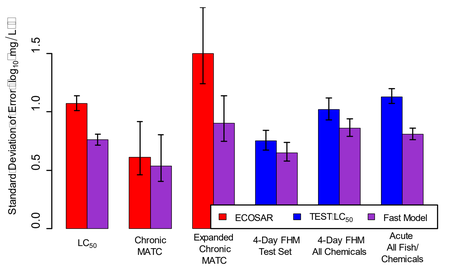


Figure 9: External validation benchmarks of the fast model against TEST[9] and ECOSAR[10] fish models using identical test sets that were not used to train either model (except in the case of the ECOSAR LC_{50}). FHM means "fathead minnows".

Conclusions

- ECOTOX experimental data tend to vary by an order of magnitude or more for a given chemical
- Experimental covariates account for some of this variation and can be added as model features
- Generalized taxonomy groups allow the model to utilize known relationships between species
- Stacked machine learning offers a small performance increase, but bootstrapping does not (not shown)
- Our model can predict acute LC_{50}'s and acute/chronic NOEC /LOEC/LC_0/MATC endpoints with RMSEs of about 0.82 and 1.00 $\log_{10}$(mg/L)
- Mean LC_{50} and NOEC are predicted within one order of magnitude for 81% and 74% of chemicals, respectively
- Predictions for narcotic chemicals show a markedly lower error than those for non-narcotics
- QSAR testing prioritization shows a sizable improvement over random guessing
- LC_{50} and chronic MATC predictions are significantly more accurate than the ECOSAR fish model
- Our model is not significantly more accurate than TEST at predicting 4-day fathead minnow endpoints, but is significantly more accurate when predicting acute LC_{50}'s for fish generally

References

1. ECOTOX: https://cfpub.epa.gov/ecotox/
2. ECHA: https://echa.europa.eu/
3. PaDEL: http://www.yapcwsoft.com/dd/padeldescriptor/
4. OPERA: Mansouri et. al. 2017, OPERA: A QSAR tool for physicochemical properties and environmental fate predictions, ACS Spring meeting
5. NCBI taxonomy: https://www.ncbi.nlm.nih.gov/taxonomy/
6. xgboost: https://cran.r-project.org/web/packages/xgboost/
7. ranger: https://cran.r-project.org/web/packages/ranger/
8. liquidSVM: https://cran.r-project.org/web/packages/liquidSVM/
9. TEST: https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test
10. ECOSAR: https://www.epa.gov/tsca-screening-tools/ecological-structure-activity-relationships-ecosar-predictive-model