Structure-based QSAR Models to Predict Systemic Toxicity Points of Departure

Prachi Pradeep[1,2] and Richard Judson[2]

[1]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee
[2]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina

Prachi Pradeep | pradeep.prachi@epa.gov | ORCID iD: 0000-0002-9219-4249 | Phone: 919-541-5150

www.epa.gov

INTRODUCTION

Human health risk assessment associated with environmental chemical exposure is limited by the tens of thousands of chemicals with little or no experimental in vivo toxicity data. A complete battery of regulatory tests for risk assessment associated with a single chemical relies on multiple animal testing's and can costs millions of dollars. A more effective and alternative way to evaluating a large number of environmental chemicals is to prioritize them for testing based on alternative methods for predicting experimental data.

Data gap filling techniques, such as quantitative structure activity relationship (QSAR) models based on chemical structure information, are commonly used to predict risk in the absence of experimental data. This study presents a set of QSAR models developed using chemical structural and physicochemical properties for in vivo points of departure (POD, the point on the dose-response that marks the beginning of a low-dose extrapolation). The in vivo data is taken from the EPA's ToxValDB, a compilation of information on ~3000 unique chemicals from a variety of public data sources. The QSAR models presented here provide estimates of POD and are evaluated for enrichment of most potent chemicals. These models will be used to inform chemical screening and prioritization efforts.

DATA PREPARATION

ToxValDB

Study type| Species (Number of chemicals)
- [chronic (CHR), subchronic (SUB) | rat (1873), mouse (719)]
- [reproductive (REP)| rat (633), mouse (107), rabbit (219)]
- [developmental (DEV)| rat (668), mouse (90), rabbit (437)]
- [subacute (SAC) | rat (689)]

toxval_units = mg/kg-day

toxval_type != [RfD, RfC]

use_me > 1

Figure 1: Schematic of data selection from the ToxValDB for developing predictive models for POD. Models were developed for each risk assessment class and species combination. E.g.
Model 1: study type = chronic, sub-chronic | species = rat
Model 2: study type = chronic, sub-chronic | species = mouse
Model 3: study type = chronic, sub-chronic | species = rat and mouse
(use_me category is an expert assigned measure of data quality.)

For demonstration of research conducted for this poster, data and results for only chronic, sub-chronic | rat and mouse models (models 2 and 3) are presented.

MOLECULAR FEATURES

- PubChem fingerprints (881 bits)
- Chemistry development kit (CDK) descriptors (18)
- PaDEL descriptors (1875)

Models were developed using combinations of PubChem, CDK and/or PaDEL descriptors

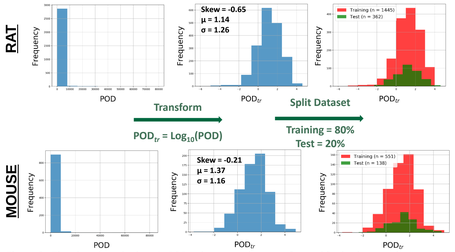

Figure 2: The chronic, sub-chronic POD values were log-transformed for both rat and mouse datasets. (a) Histogram of untransformed POD data, (b) Histogram of transformed POD (POD_{tr}) data, and (c) Histogram of training and test data relative to each other.

CHALLENGES

1. Experimental Variability
- Data from different labs (sources) running the "same" experiment may get different answers
- Sources of variability: Species, strain, dose range, dose spacing, length of study etc.

2. Model Uncertainty
- A model gives a result (a POD), but this is an estimate of the "true" POD. The true POD is mostly unknown.
- Uncertainty in the evaluation data will lead to uncertainty in the model and our estimate of its quality

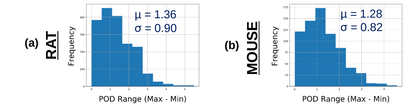

Figure 3: Distribution of the range of POD values for the CHR-SUB (a) rat and (b) mouse datasets. Variability in experimental data leads to uncertainty in the model predictions. Roughly, the root mean squared error (RMSE) in the models can be estimated to be around 1.17 (=V1.36) for rat and 1.13 (=V1.28) for mouse models by just looking at the distribution of POD values for both datasets.

TYPES OF MODELS

Given the variability and skewness in the training dataset, 3 types of models were developed:

1. Point-estimate models
- Regression models were built to predict a POD value for each chemical.
- Experimental POD = Minimum POD value from all studies (same species, study type)

2. Point-estimate, balanced-dataset models
- The training data was re-constructed to reduce skewness by adding 10% duplicate data from the long tail.
- Regression models were built to predict a POD value for each chemical using the re-constructed data. The process was repeated *n* times.
- Experimental POD = Minimum POD value from all studies (same species, study type)

3. Point-estimate with confidence interval models
- A POD distribution was constructed for each chemical using mean = experimental POD value (= Median POD value from all studies) and standard deviation = 0.5 log-units, based on the typical lab to lab variability.
- *n* bootstrap models were built with random sampling of POD values for each chemical from the pre-generated POD distribution.

METHODS

(a)

FEATURE SELECTION
Fingerprints
1. Unsupervised 80% variance threshold
2. 80% collinearity threshold

Descriptors
1. Normalization to mean=0 and variance=1
2. Supervised recursive feature elimination using linear regression

ALGORITHM
1. k nearest neighbors (kNN)
2. Support vector regression (SVR)
3. Random forest (RF)
4. Gradient boosting regression (GBR)

HYPER-PARAMETER TUNING
5-fold GridSearchCV
1. kNN: k, weights, algorithm
2. SVR: Epsilon, C, gamma, kernel
3. RF: Max features, N trees
4. GBR: N trees, max depth, loss function, learning rate

MODEL VALIDATION
Internal validation
5-fold internal cross-validation on 80% training set

External validation
20% test set

(b)

ENRICHMENT ANALYSIS

Each model was evaluated on the external test set for enrichment of N% most potent chemicals.

1. The chemicals in the predicted external test set were sorted in the order of potency.
2. X% of above sorted list was then evaluated for % enrichment of the N% most potent chemicals

Figure 4: (a). Workflow for development of QSAR models, and (b). Algorithm for model enrichment analysis. All the models were developed and evaluated for enrichment for each combination of study type and species. A summary of comparative predictive ability of the best model for each of the combinations is presented in the results.

RESULTS

1. Point-estimate models

Species [Molecular Features (number)]	Random Forest Hyper-parameters	5-fold Internal Cross-validation			External Validation		
		RMSE	RMSE/σ	R^2	RMSE	RMSE/σ	R^2
Rat [PubChem (43) + CDK (5) + PaDEL (5)] Coverage = 1807	Max features = auto Number of trees = 1250	1.02	0.80	0.35	0.98	0.78	0.39
Mouse [PubChem (49) + CDK (5)] Coverage = 719	Max features = auto Number of trees = 750	0.99	0.85	0.27	1.02	0.86	0.26

Table 1: Performance metrics for the best (random forest) point-estimate models for 5-fold internal cross validation and external validation for CHR-SUB rat and mouse datasets.

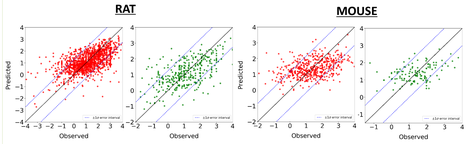

Figure 5: Plots of observed versus predicted POD values (transformed scale) for the best rat and mouse model (random forest model) for 5-fold internal cross-validation (red scatter plots) and external validation (green scatter plots).

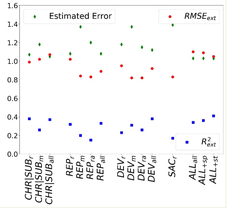

Figure 5: A summary of the best model metrics and the estimated error (Figure 3) for each combination of study type and species on the external test set. As seen, there is not much variation in the metrics across different model combinations. (r: rat, m: mouse, ra: rabbit, sp: species, st: study type)

RESULTS

1. Point-estimate models: Enrichment Analysis

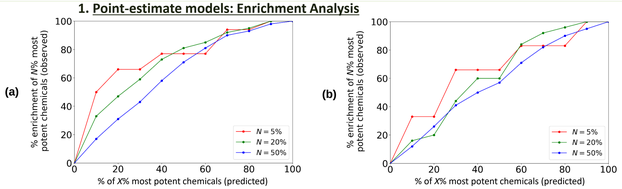

Figure 6: Enrichment analysis for (a) rat, and (b) mouse data. 80% of the 20% most potent chemicals are enriched within 40% of predicted data for the rat model and 60% of predicted data for the mouse model. These results demonstrate the utility of these models for chemical prioritization and testing.

2. Point-estimate with balanced dataset models

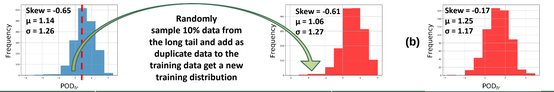

Figure 7: Sample histogram of re-constructed data for (a) rat and (b) mouse. Total number of samples in the re-constructed training dataset for rat data is N = 1589 and mouse data is N = 606.

2. Point-estimate, balanced-dataset models	Random Forest Hyper-parameters	5-fold Internal Cross-validation			External Validation		
		RMSE	RMSE/σ	R^2	RMSE	RMSE/σ	R^2
Rat	Max features = auto/sqrt Number of trees = 1000/1250/1500	0.95	0.75	0.45	1.00	0.80	0.36
Mouse	Max features = auto/sqrt Number of trees = 1000/1500	0.90	0.79	0.39	1.03	0.87	0.24

Table 2: Performance metrics for the point-estimate balanced-dataset models for 5-fold internal cross-validation and external validation.

3. Point-estimate with confidence interval models

Illustrative example: Bisphenol A (Mouse, POD_{tr})

0.699
0.952
-0.215
2.176
2.875
2.176
2.875

Median POD: 2.176

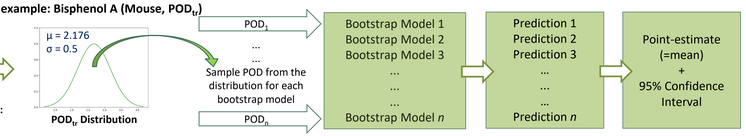

Figure 9: Plots of observed versus predicted POD values for the best rat and mouse model (random forest model) for 5-fold internal cross-validation (red scatter plots) and external validation (green scatter plots) with 95% confidence intervals (green error bars) for n = 100 bootstrap models.

CONCLUSIONS

- Point-estimate model results demonstrate that independent study type and species combinations did not result in significantly better models than combining the data for all the classes and species together.
  - The RMSE for the all the models are within the variance in the underlying POD data (Figure 3 and 5).
  - Enrichment analysis results demonstrate the utility of these models for chemical screening and prioritization efforts.
- Point-estimate with balanced dataset model results improved the training set results but did not show improved results on the external test sets.
- Point-estimate with confidence interval models presented a technique to estimate uncertainty associated with model predictions. The results demonstrate the impact of variability in training data (experimental POD) on uncertainty associated with model results.

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