www.epa.gov

Development of QSAR Models for Systemic Toxicity Points of Departure with Variability in Experimental Data

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

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. Data gap filling techniques, such as quantitative structure activity relationship (QSAR) models based on chemical structure information, are commonly used to predict hazard in the absence of experimental data. However, variability in the experimental data leads to uncertainty in QSAR model predictions and impacts model quality estimates.

This study presents three sets of QSAR models developed for systemic toxicity 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 chemicals from a variety of public data sources. The first set of QSAR models were developed and evaluated to predict point estimates of POD values using structural and physicochemical descriptors. The second set of models were built to account for skewness in the training data. The third set of models were built to account for the known lab to lab variability in experimental POD values. The QSAR models were also evaluated for enrichment of most potent chemicals. These models will inform chemical screening and prioritization efforts.

DATA PREPARATION

Study Type	Species	Total number of POD values (studies)	Number of unique chemicals
Chronic (CHR)	Rat	13423	3047
	Mouse	4130	690
	Rabbit	342	240
	Rat, Mouse, Rabbit	17895	3221
Subchronic (SUB)	Rat	6696	988
	Mouse	2418	308
	Rat, Mouse	9114	1030
Reproductive (REP)	Rat	2915	425
	Mouse	244	62
	Rat, Mouse	3159	460
Developmental (DEV)	Rat	2472	416
	Rabbit	1540	273
	Rat, Rabbit	4012	511
Subacute (SAC)	Rat	1133	155
ALL (CHR, SUB, REP, DEV, SAC)	**All (Rat, Mouse, Rabbit)**	**36013**	**3762**

Table 1. Number of POD values (experiments/studies) and unique chemicals with data across different study types and species combinations with data on more than 50 chemicals.

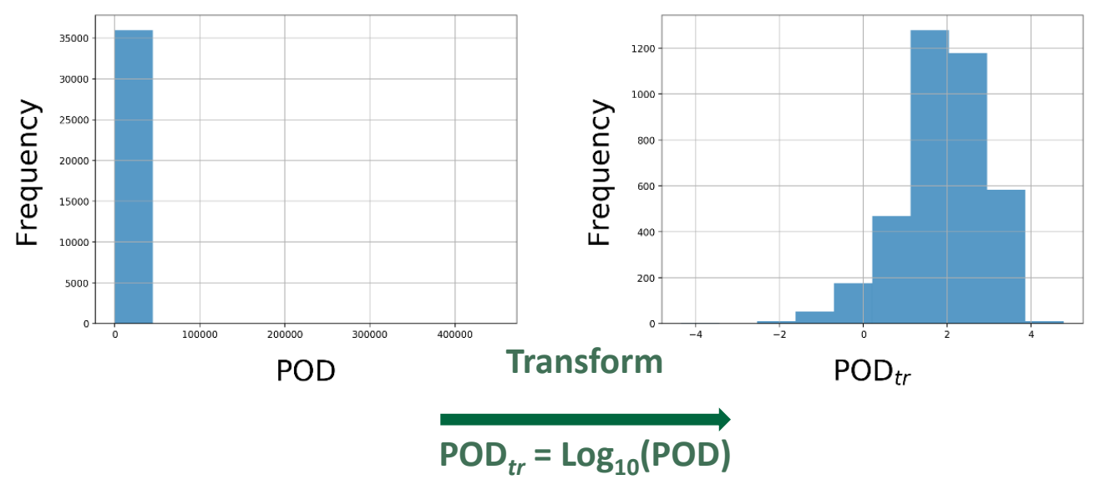


Figure 1: POD values were log-transformed before model development. (a) Histogram of untransformed POD data, (b) Histogram of transformed POD (POD$_{tr}$) data

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

Models were developed for each study type and species combination. E.g.
Model 1: study type = chronic | species = rat
Model 2: study type = chronic | species = mouse

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

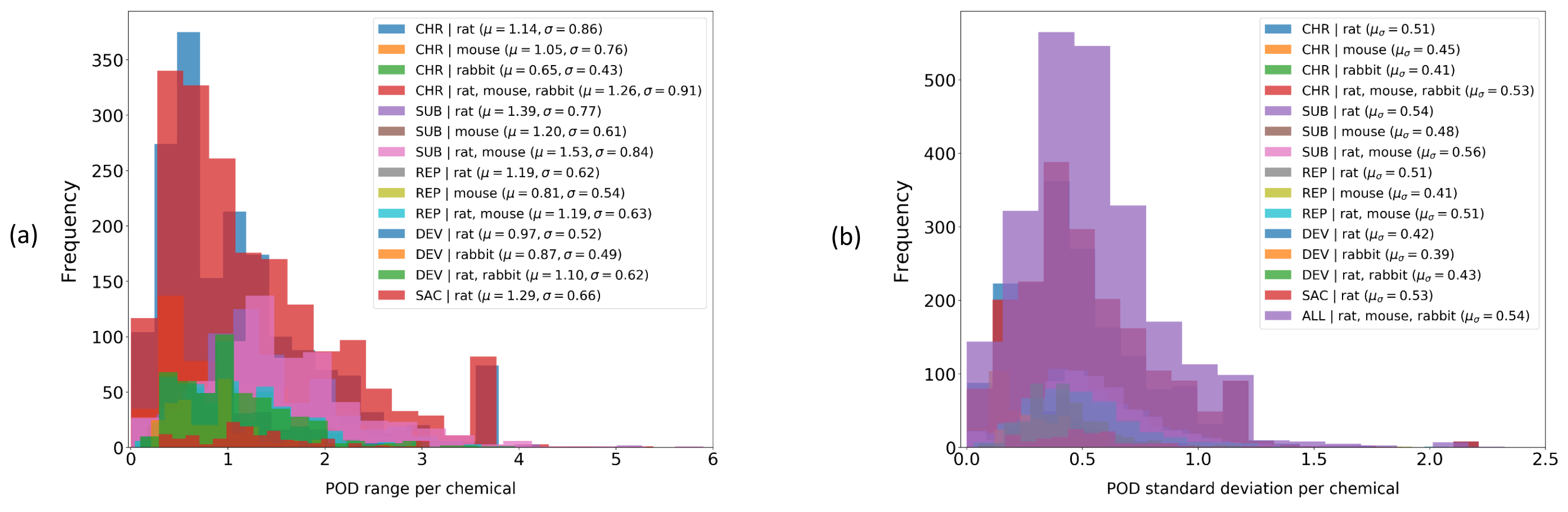


Figure 2. (a) Distribution of the range of POD values per chemical as obtained from the ToxVal database for each study type combination. The mean (μ) of the distribution is used to estimate a bound on RMSE values for the QSAR models developed using these data, (b) Distribution of the standard deviation (σ) of the POD values for each chemical per study type and species combination. The mean standard deviation (μ_σ) gives an estimate of the experimental variability in the underlying data for each combination.

TYPES OF MODELS

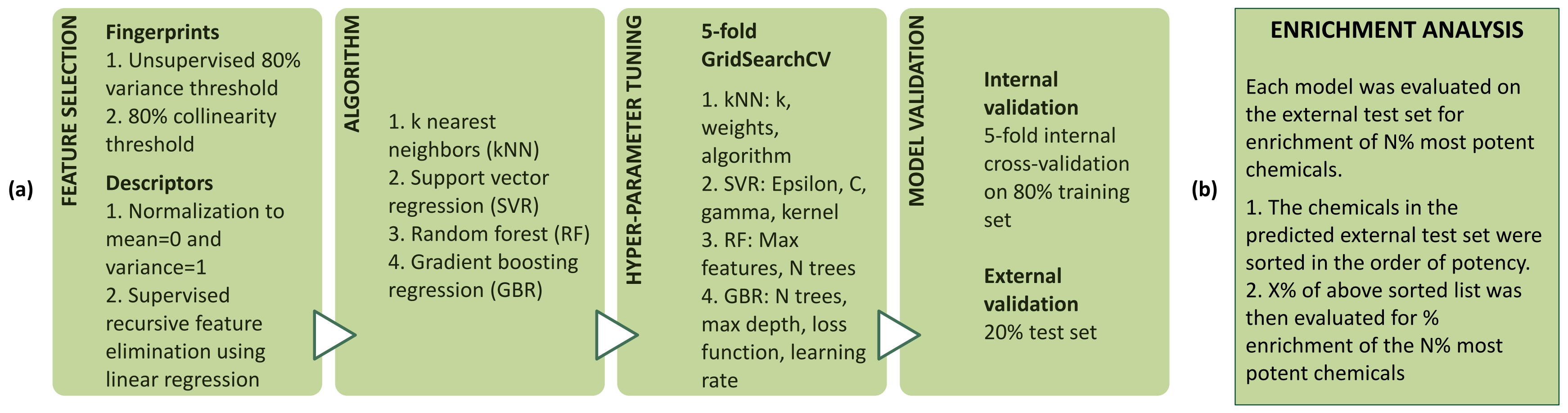
Given the variability and skewness in the training dataset, 3 types of models were developed:

1. Point-estimate Models
- A single POD value predicted for each chemical.
- Experimental POD = Median POD value from all studies.

2. Point-estimate with Balanced Dataset Models
- Training data re-constructed to reduce skewness.
- A single POD value predicted for each chemical using the re-constructed data. The process was repeated *1000* times.
- Experimental POD = Median POD value from all studies.

3. Point-estimate with Confidence Interval Models
- A POD distribution was constructed for each chemical (μ = Median experimental POD value from all studies, σ = 0.5 log-units).
- *1000* bootstrap models were built with random sampling of POD values for each chemical from the pre-generated POD distribution.
  - Predicted POD = mean of 1000 bootstrap predictions
  - Confidence interval of POD = ±1 standard deviation of 1000 bootstrap predictions

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

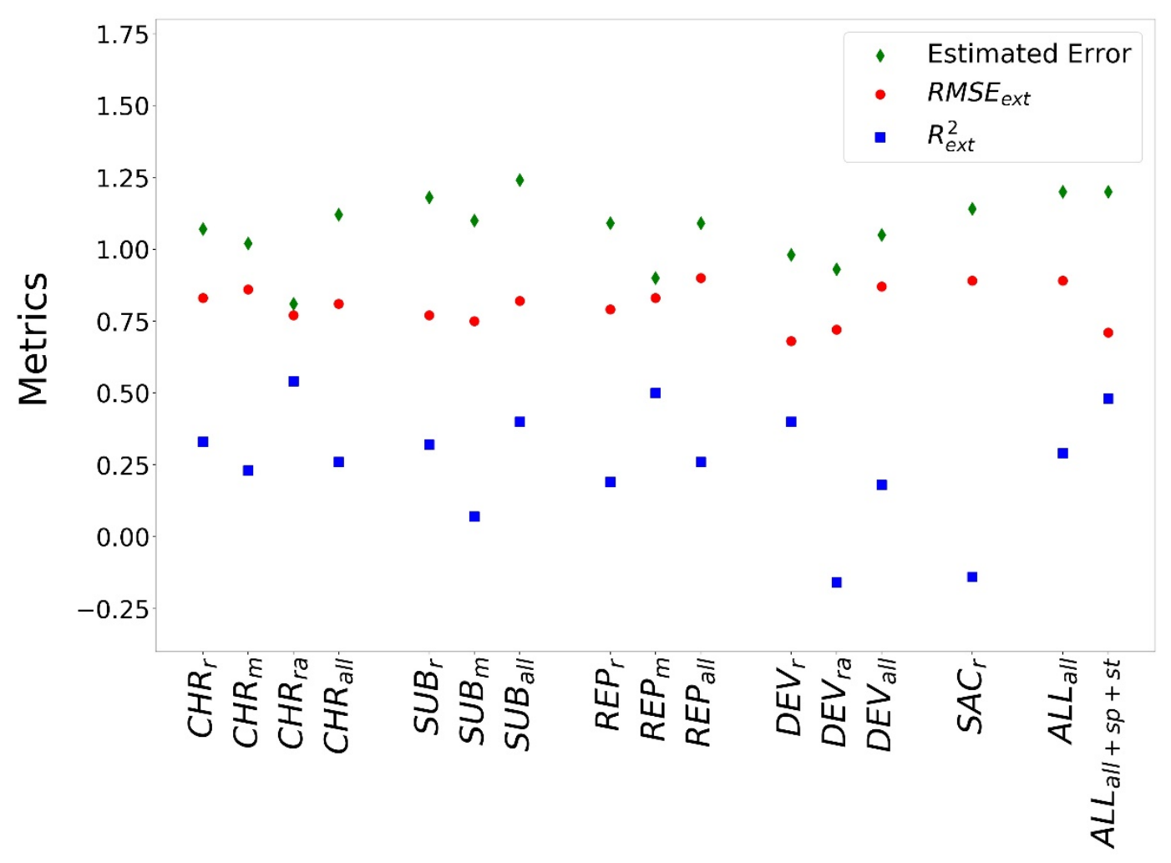
Each model was evaluated on the external test set for enrichment of N% most potent chemicals.

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 3. (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.

RESULTS

Figure 4. A summary of the best model metrics and the estimated error for each combination of study type (CHR: chronic, SUB: subchronic, REP: reproductive, DEV: developmental, SAC: subacute, and ALL: all study types) and species (r: rat, m: mouse, ra: rabbit, sp: species, st: study type, all: all species - indicated as subscripts to the study type) on the external test set for the point-estimate models. The estimated error is derived as the square-root of the mean POD for each combination as shown in Figure 2(a). As seen, there is not much variation in the performance metrics across different model combinations and the RMSE for all the models is comparable to the estimated errors from the underlying data. The '+sp+st' in the 'ALL' combinations model indicates using species and study type as additional descriptors in the model.



RESULTS

1. Point-estimate Models

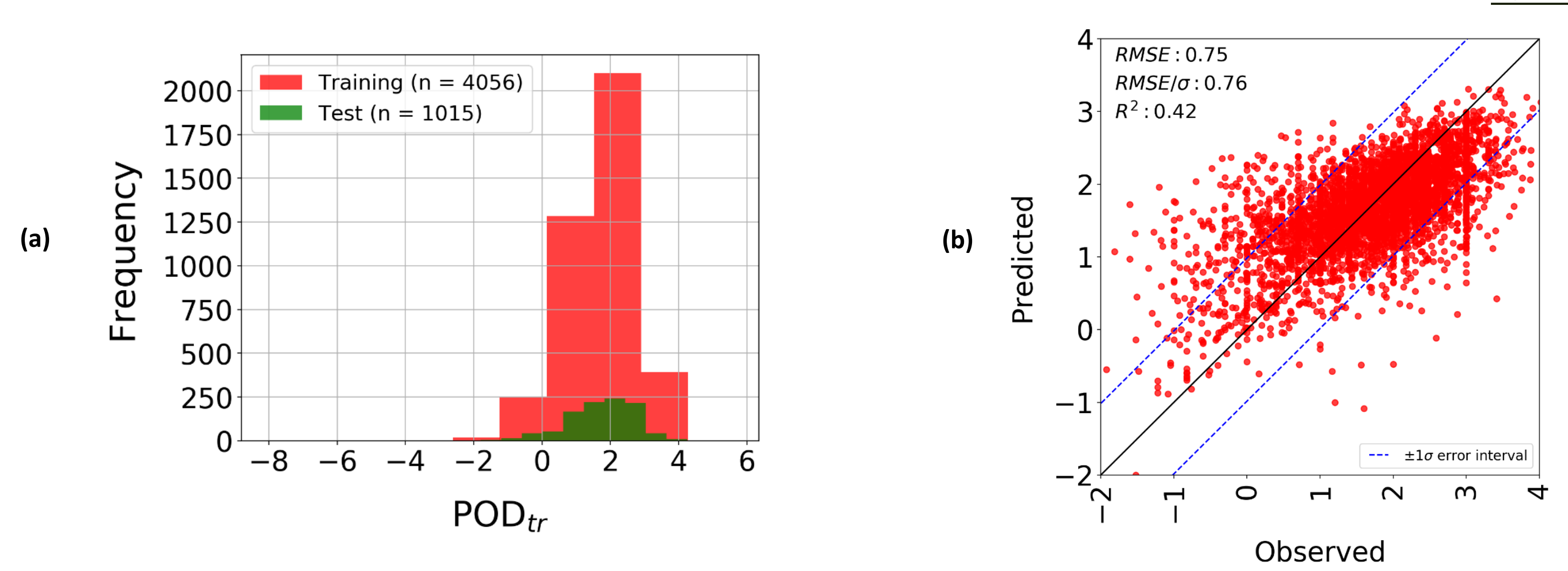


Figure 5. Data and model performance results for the best model (random forest) developed using a combination of all study types (CHR, SUB, DEV, REP and SUB) and all species (rat, mouse and rabbit) and using species and study type as additional descriptors in the model. (a) Distribution of the training and training and test dataset relative to each other. (b) Observed versus predicted POD values (transformed scale) for 5-fold internal cross-validation (red scatter plot) and external validation (green scatter plot). The legend shows the performance metrics of each set where RMSE = root mean-squared error, σ = standard deviation of the dataset, and R^2 = coefficient of determination. The blue dotted lines indicate an error interval of one standard deviation of the dataset. The points within the error interval are predicted within one standard deviation of the observed values.

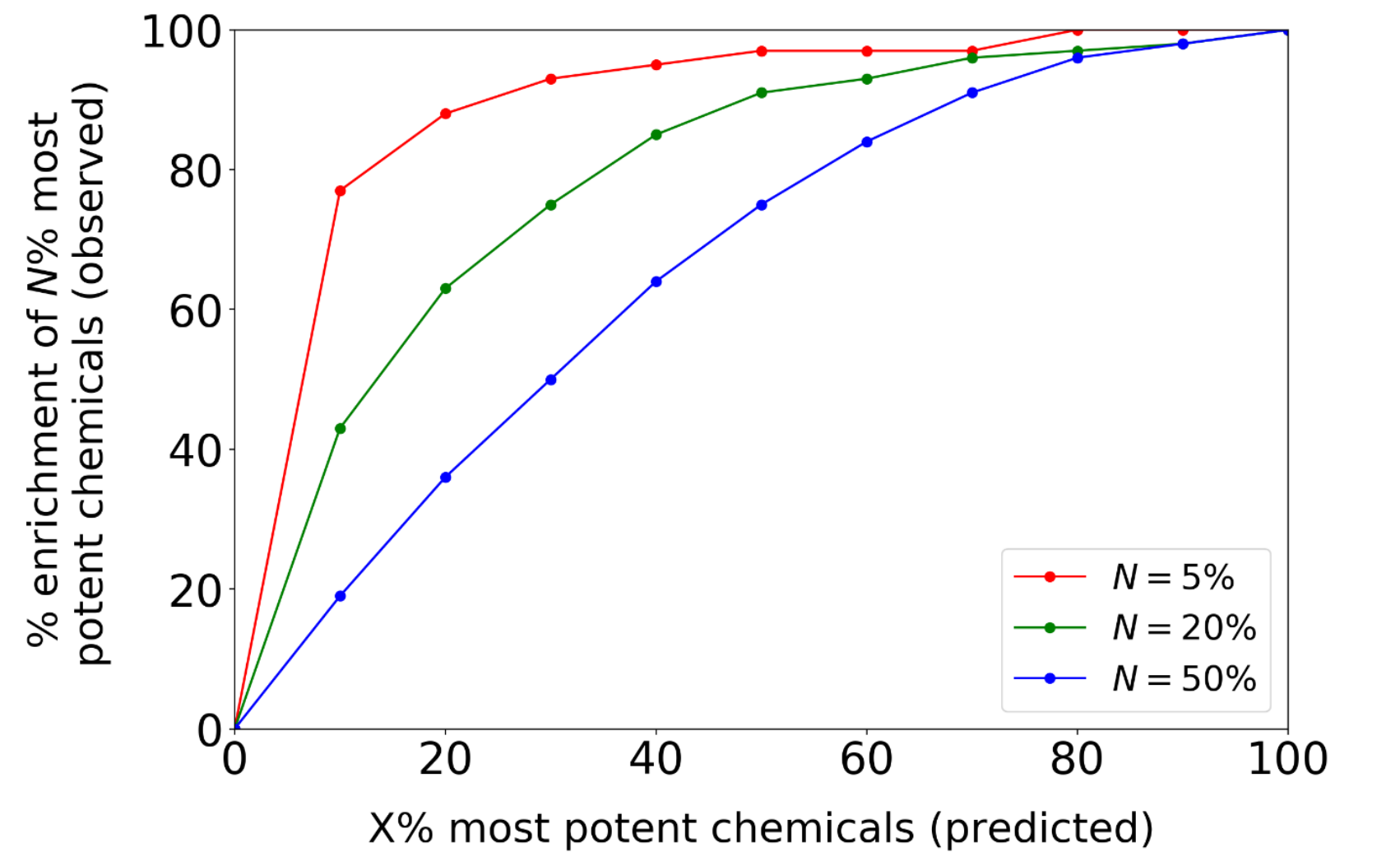
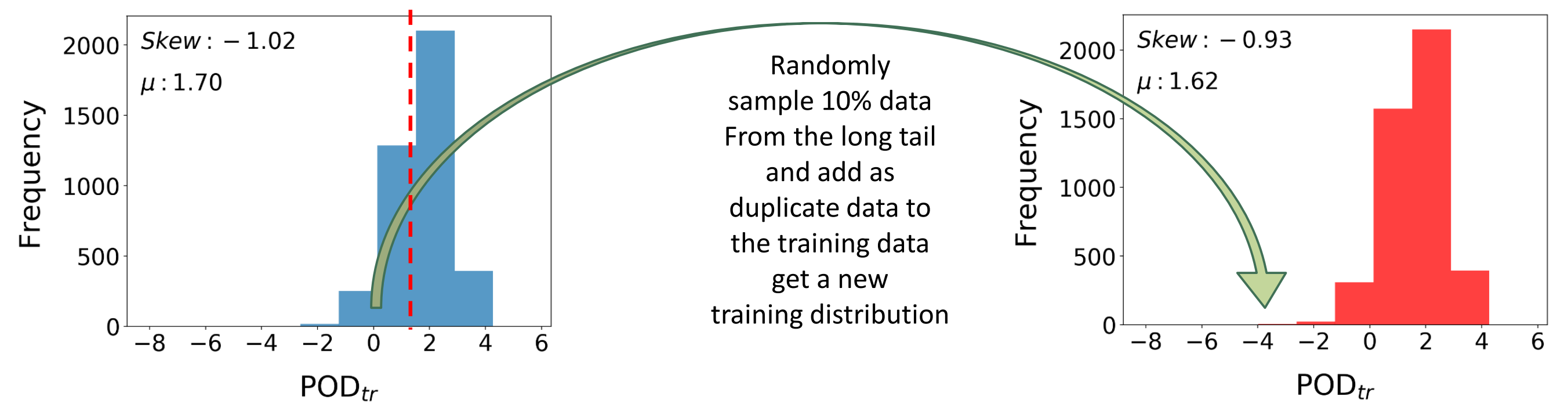


Figure 6. Enrichment analysis for 5%, 20% and 50% most potent chemicals (observed) in the predicted set from the random forest model developed using the dataset with all study types and species and using study type and species as additional descriptors in the model. Over 75% of the 5% most potent chemicals were enriched within 10% of predicted potent chemicals.

2. Point-estimate with Balanced Dataset Models



Randomly sample 10% data From the long tail and add as duplicate data to the training data get a new training distribution

Figure 7. Schematic outlining the process for balancing the training dataset for development of point-estimate balanced dataset models. For each (of *n*) models, the 10% data is randomly sample from the original training dataset and duplicated to obtain a new training dataset. In this example iteration, the original training dataset which is left skewed with a skewness coefficient = -1.02 is balanced to such that the balanced dataset has a reduced skewness coefficient = -0.91. The balanced dataset is then used for training the models. This process is repeated *n* times where *n* = 1000.

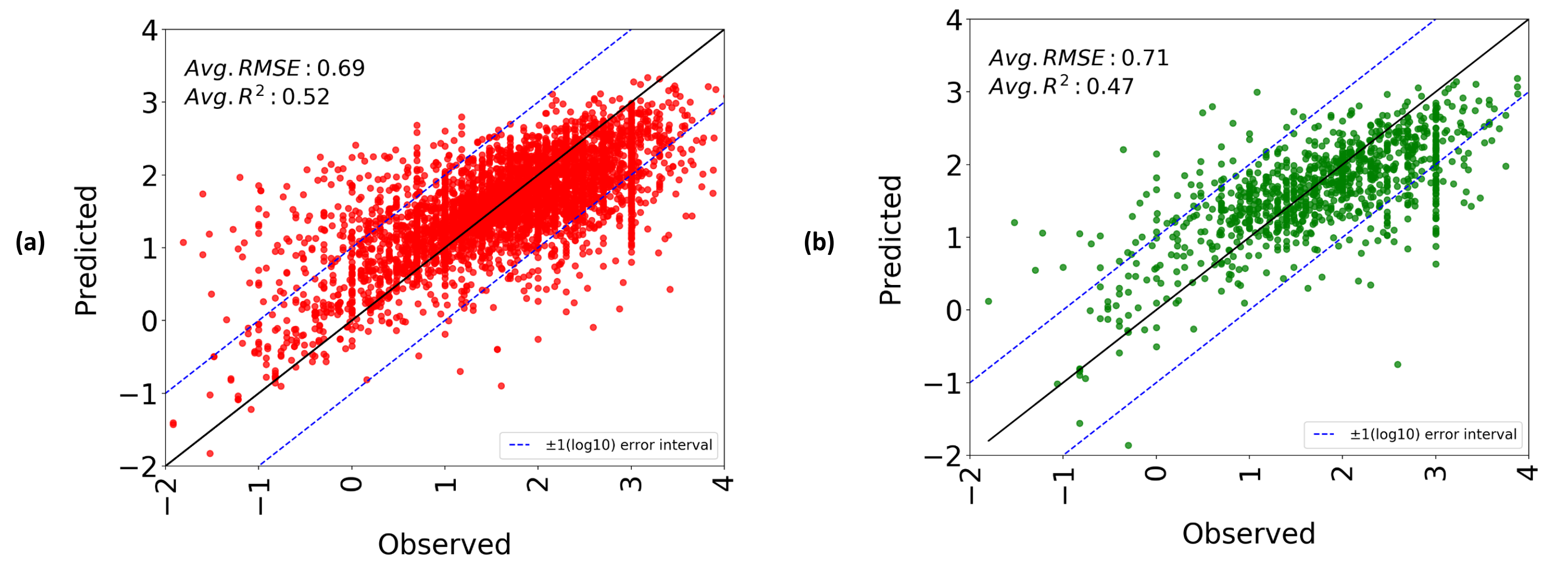
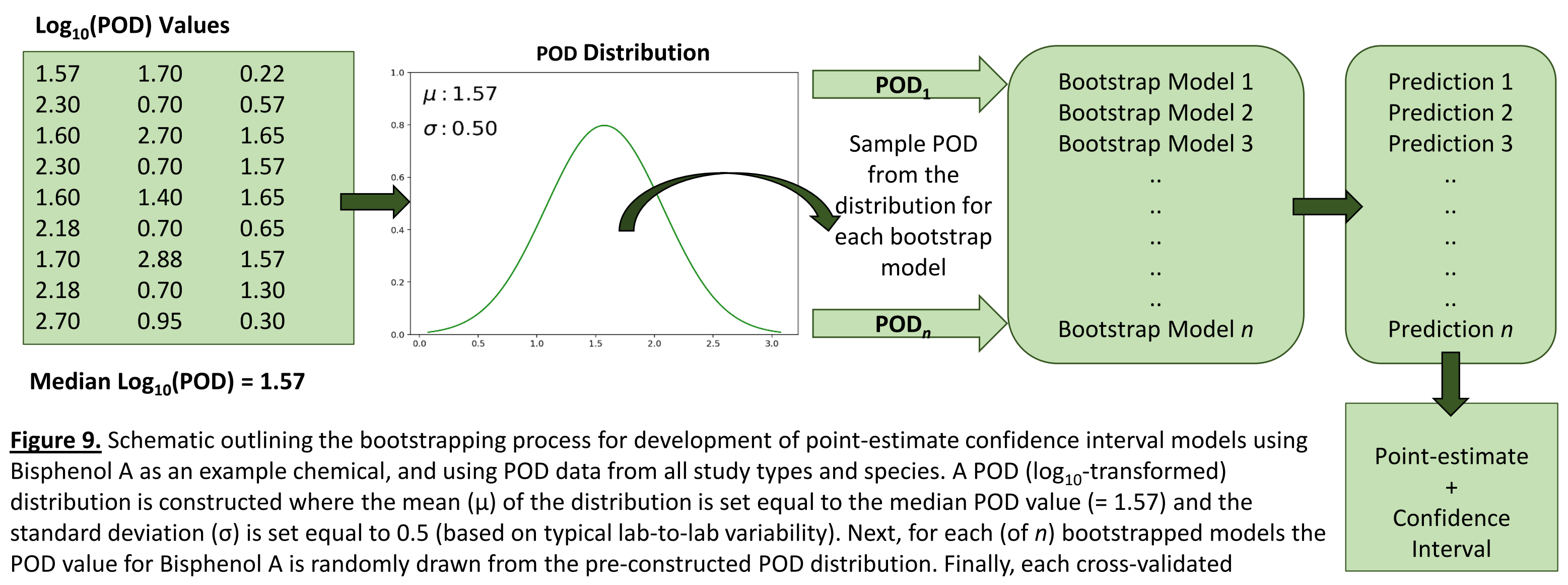


Figure 8. Example model performance results calculated using n=1000 models using a balanced training dataset for (a). training set, and (b). test set. The average RMSE and average R^2 are shown for both the training and the test set. As shown, balancing the training dataset helps improve the model metrics for the training dataset but do not significantly affect the test set. The blue dotted lines indicate an error margin of 1-log$_{10}$ unit.

3. Point-estimate with Confidence Intervals Models


Log$_{10}$(POD) Values

1.57	1.70	0.22
2.30	0.70	0.57
1.60	2.70	1.65
2.30	0.70	1.57
1.60	1.40	1.65
2.18	0.70	0.65
1.70	2.88	1.57
2.18	0.70	1.30
2.70	0.95	0.30

Median Log$_{10}$(POD) = 1.57

Sample POD from the distribution for each bootstrap model

Bootstrap Model 1
Bootstrap Model 2
Bootstrap Model 3
..
..
..
..
..
Bootstrap Model *n*

Prediction 1
Prediction 2
Prediction 3
..
..
..
..
..
Prediction *n*

Point-estimate + Confidence Interval

Figure 9. Schematic outlining the bootstrapping process for development of point-estimate confidence interval models using Bisphenol A as an example chemical, and using POD data from all study types and species. A POD (log$_{10}$-transformed) distribution is constructed where the mean (μ) of the distribution is set equal to the median POD value (= 1.57) and the standard deviation (σ) is set equal to 0.5 (based on typical lab-to-lab variability). Next, for each (of *n*) bootstrapped models the POD value for Bisphenol A is randomly drawn from the pre-constructed POD distribution. Finally, each cross-validated bootstrapped model predicts a POD value resulting in *n* POD predictions. The final point-estimate POD value is the mean of the *n* predictions and the confidence interval derived as the one standard deviation of the *n* predictions. In this work, *n* = 1000.

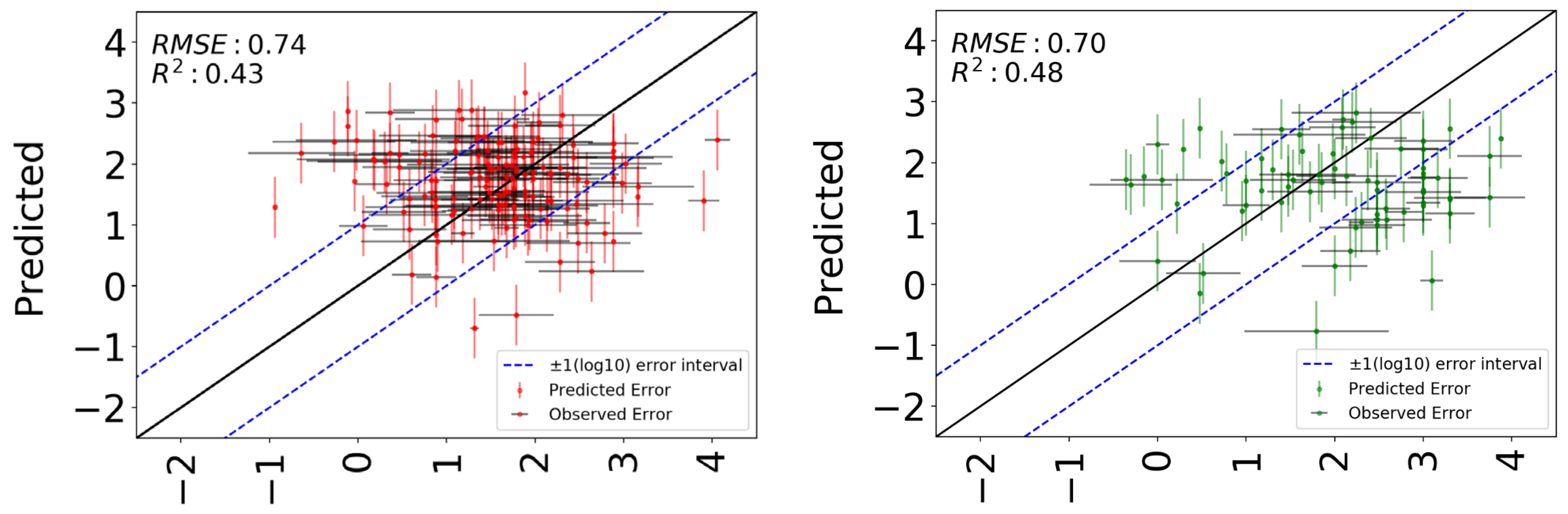


Figure 10. Model performance results calculated using n=1000 bootstrap models for (a). training set, and (b). test set. 50 chemicals from each, training and test, set were randomly selected and plotted. The predicted error bar for each chemical is calculated as the standard deviation of the predictions from the models. The observed error bar is calculated as the standard deviation of the experimental data for each chemical. The blue dotted lines indicate an error margin of 1-log$_{10}$ unit.

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 (Figures 2 and 6).
  - Enrichment analysis results demonstrate the utility of these models for chemical screening and prioritization efforts.
- Point-estimate with balanced dataset model results show improvement in 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.

Disclaimer: The views expressed are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

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