Experimental Variability and Uncertainty in the Context of New Approach Methodologies for Potential Use in Chemical Safety Evaluation

EPA CompTox Communities of Practice
August 27, 2020

Prachi Pradeep, PhD
ORISE Research Fellow
Center for Computational Toxicology and Exposure
U.S. Environmental Protection Agency

Disclaimer: The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA

Outline and Learning Objectives

Outline

I. Quantitative variability in repeat dose toxicity studies
II. Using experimental variability to derive uncertainty in the context of QSAR models
III. Limitations on predictivity of QSAR models owing to underlying data

Key Learning

The quality of data underpinning a predictive model is an important determinant of the robustness/quality of the model

I. Quantitative variability in repeat dose toxicity studies

Pham, L.L., et al., *Variability in in vivo studies: Defining the upper limit of performance for predictions of systemic effect levels.* Computational Toxicology, 2020.

Acceptability of NAMs for risk assessment?

- In US, Section 4(h) in the Lautenberg amendment to TSCA:
  - "…Administrator shall reduce and replace, to the extent practicable and scientifically justified…the use of vertebrate animals in the testing of chemical substances or mixtures…"
  - New approach methods (NAMs) need to provide "information of equivalent or better scientific quality and relevance…" than the traditional animal models
- "Directive to Prioritize Efforts to Reduce Animal Testing" memorandum signed by Administrator Andrew Wheeler on September 10, 2019
  - "1. Validation to ensure that NAMs are equivalent to or better than the animal tests replaced."

How do we define expectations of *in silico, in chemico,* and *in vitro* models for predicting repeat-dose toxicity?

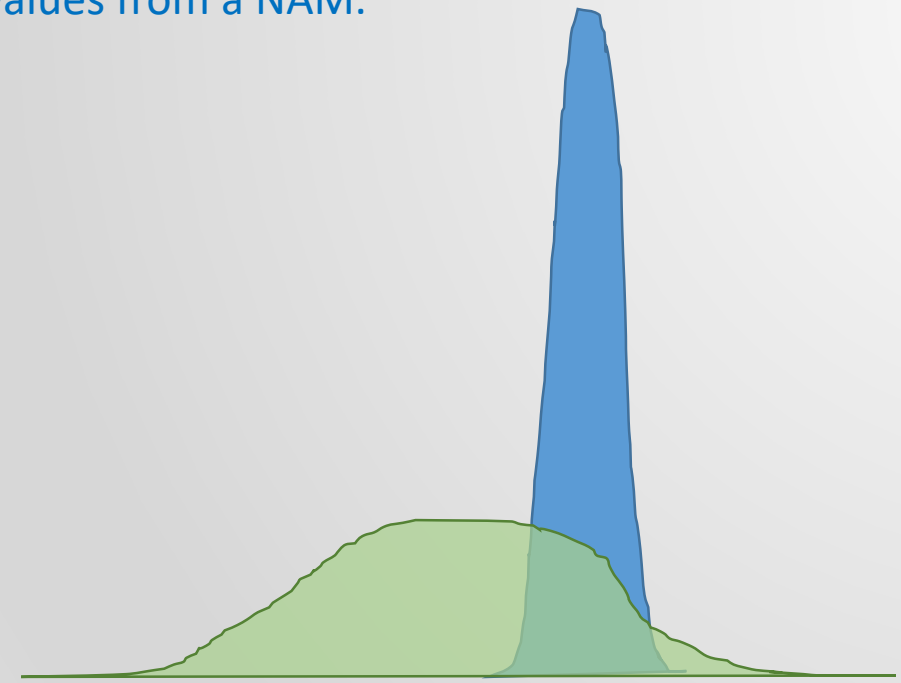
- *In silico, in chemico,* and *in vitro* models cannot predict *in vivo* systemic effect values with greater accuracy than those animal models reproduce themselves.

Slide courtesy: Dr. Katie Paul-Friedman

Variability in traditional animal toxicity tests

Quantitative: variance is a measure of how far values are spread from the average.

We need to know what the "spread" or variability of traditional effect levels (e.g., lowest effect levels, LELs, or lowest observable adverse effect levels, LOAELs) might be to know the range of acceptable or "good" values from a NAM.



Qualitative: We need to know if a specific effect is always observed or not.

		"Truth" (traditional toxicology)	
		Negative	Positive
Predicted (NAM)	Negative	True negative	False negative
	Positive	False positive	True positive

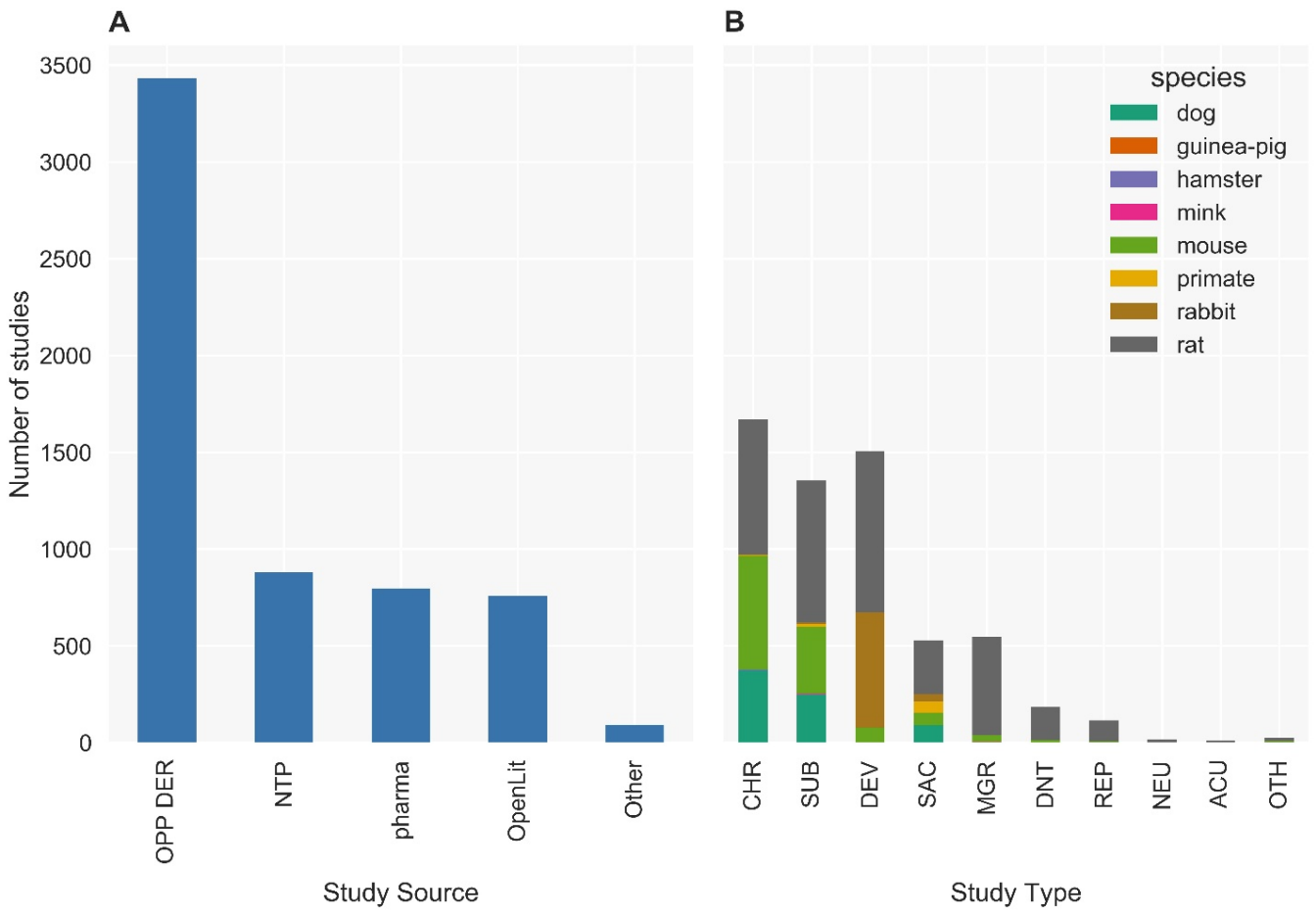
Slide courtesy: Dr. Katie Paul-Friedman

Understanding this quantitative variability

	What is the range of possible systemic effect values (mg/kg/day) in replicate studies?	What is the maximal accuracy of a model that attempts to predict a systemic effect values for an unknown chemical?
Statistical approach to the question	• Residual root mean square error (RMSE) is an estimate of variance in the same units as the systemic effect values. • The RMSE can also be used to define a minimum prediction interval, or estimate range, for a model.	• The mean square error (MSE) is used to approximate the unexplained variance (not explained by study descriptors). • This unexplained variance limits the R-squared on a new model.

Slide courtesy: Dr. Katie Paul-Friedman

Using ToxRefDB to evaluate quantitative variability



Number of studies by study type and species in ToxRefDB v2.0. *The study designs include chronic (CHR), sub-chronic (SUB), developmental (DEV), subacute (SAC), multigeneration reproductive (MGR), developmental neurotoxicity (DNT), reproductive (REP), neurotoxicity (NEU), acute (ACU), and other (OTH) for numerous species, but mostly for rat, mouse, rabbit, and dog.*

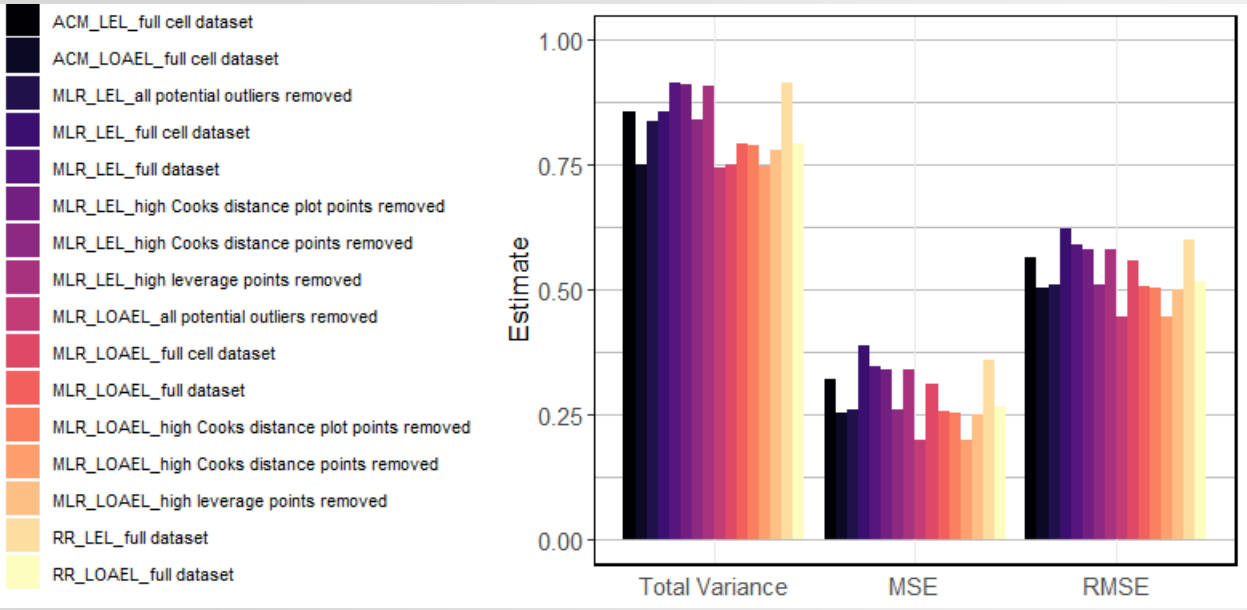
ToxRefDB v2.0 contains relevant study data to evaluate variability in traditional data for >1000 chemicals and >5000 studies.

Figure from Watford S, Pham LL, Wignall J, Shin R, Martin MT, Paul Friedman K. 2019. "ToxRefDB version 2.0: Improved utility for predictive and retrospective toxicology analyses." Reproductive Toxicology; 89: 145-158. https://doi.org/10.1016/j.reprotox.2019.07.012

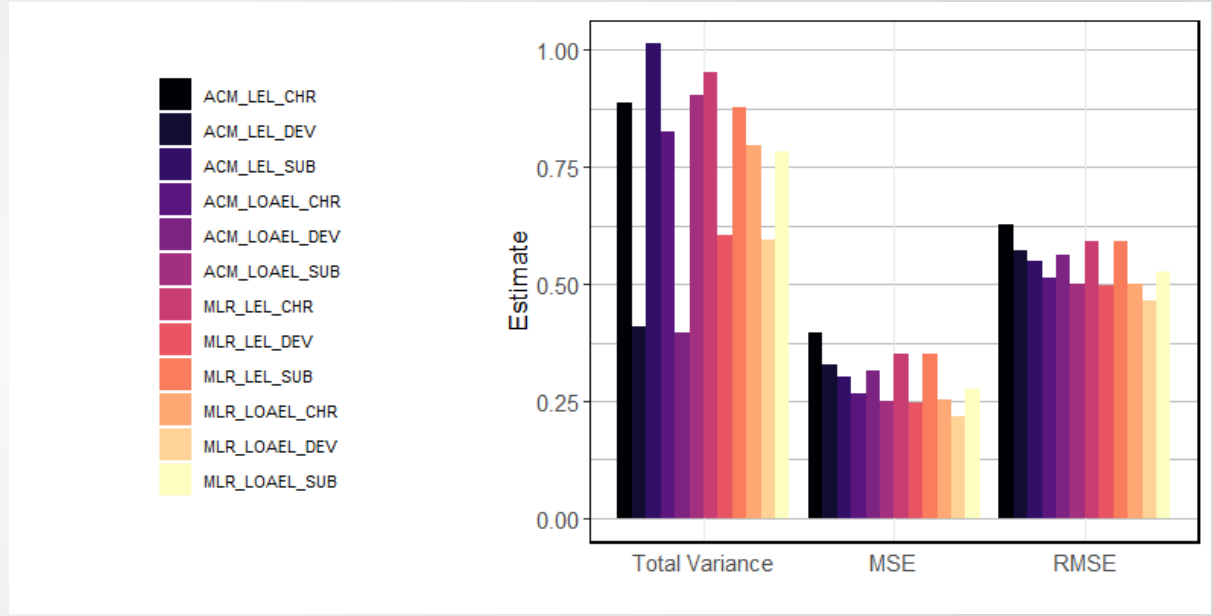
Slide courtesy: Dr. Katie Paul-Friedman

28 Models to Approximate Total Variance, Unexplained Variance, and Spread of the Residuals from the Statistical Models

Statistical models for LELs and LOAELs for the full dataset



Statistical models for LELs and LOAELs for datasets subset by study type

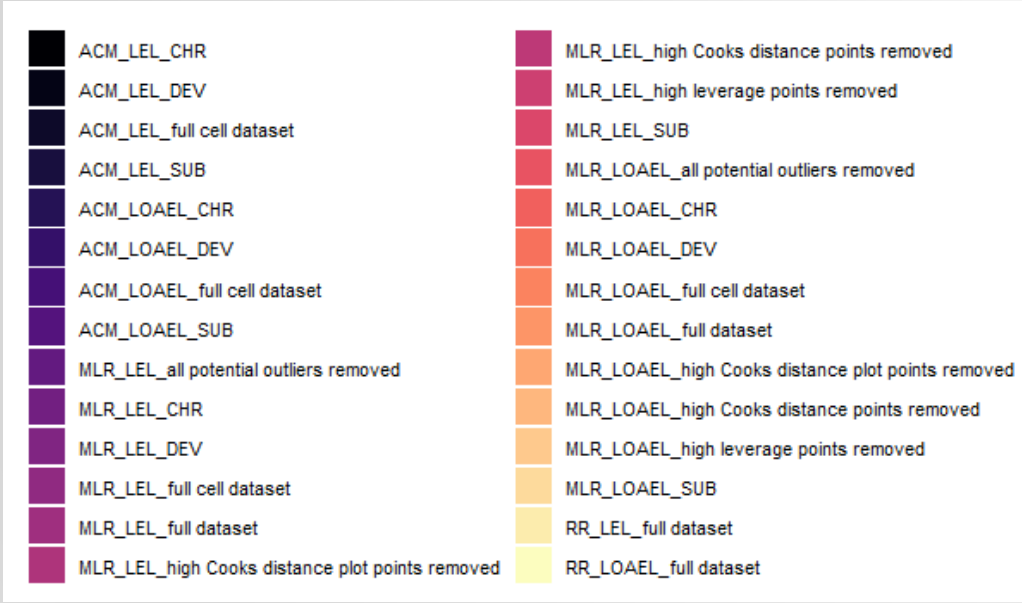


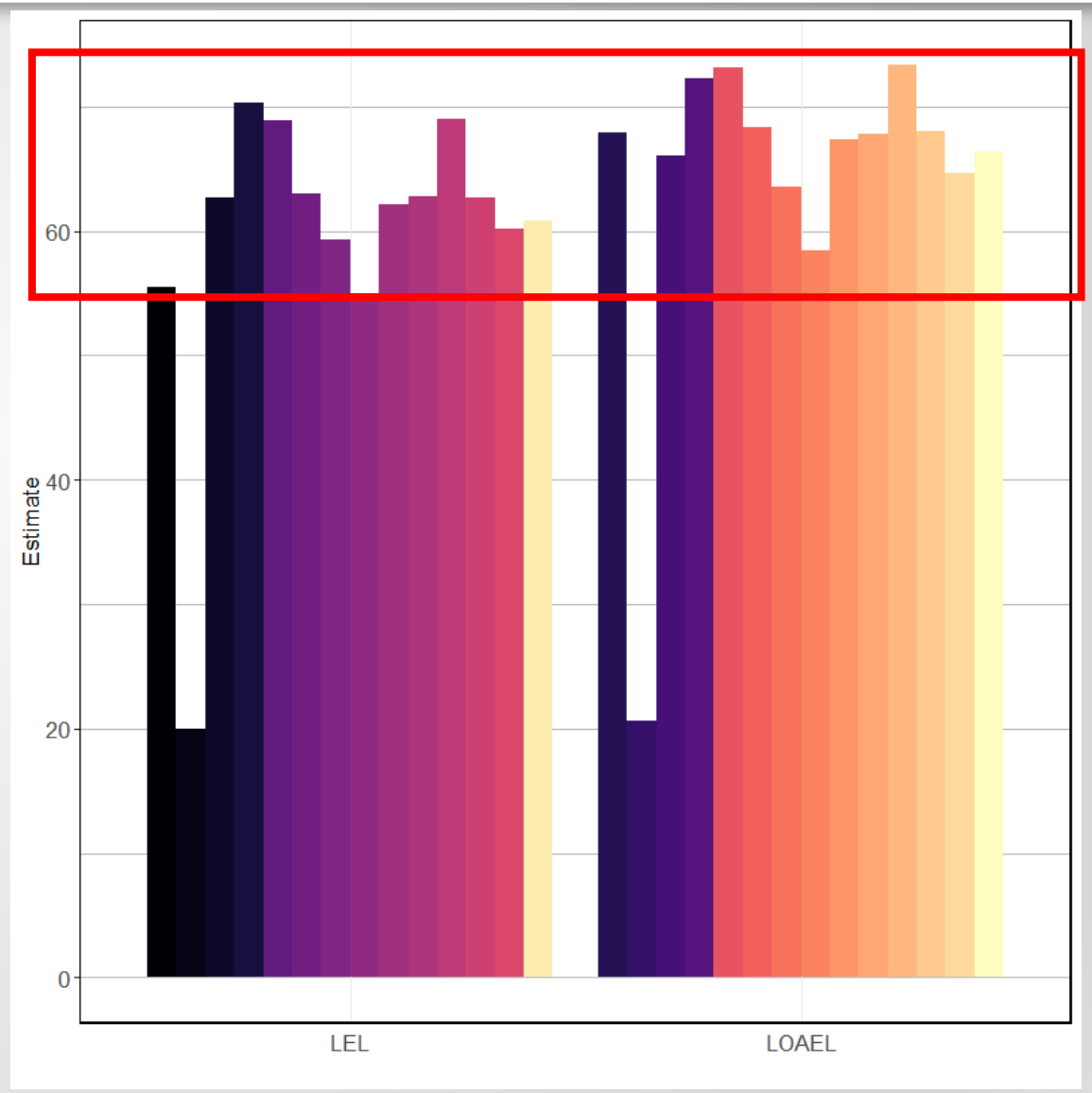
- Total variance in systemic toxicity effect values (from ToxRefDB) likely approaches 0.75-1 (units of $(\log_{10}\text{-mg/kg/day})^2$)
- MSE (unexplained variance) is 0.2 – 0.4 (units of $(\log_{10}\text{-mg/kg/day})^2$)
- RMSE is 0.45-0.60 $\log_{10}$-mg/kg/day
- RMSE is used to define a 95% minimum prediction interval (i.e., based on the standard deviation or spread of the residuals)

Slide courtesy: Dr. Katie Paul-Friedman

Percent Variance Explained

- The % explained variance (amount explained by study descriptors) likely approaches 55-73%.
- This means that the R^2 on some new, predictive model would approach 0.55 to 0.73 as an upper bound on accuracy.





Slide courtesy: Dr. Katie Paul-Friedman

Primary Conclusions and Outlook

- Variability in *in vivo* toxicity studies limits predictive accuracy of NAMs.
- Maximal R-squared for a NAM-based predictive model of systemic effect levels may be 55 to 73%; i.e., as much as 1/3 of the variance in these data may not be explainable using study descriptors.
- The estimate of variance (RMSE) in curated LELs and/or LOAELs approaches a 0.5 log10-mg/kg/day.
- Estimated minimum prediction intervals for systemic effect levels may be approximately ± 1 log10-mg/kg/day – this is without variance contributed by some NAM itself!

Slide courtesy: Dr. Katie Paul-Friedman

II. Using Experimental Variability to Derive Uncertainty in the Context of QSAR Models

Pradeep, P. et al. *Structure-based QSAR Models to Predict Repeat Dose Toxicity Points of Departure.* Submitted.

Repeat Dose Toxicity Points of Departure

Toxicity in repeat dose studies includes a range of adverse effects on one or more systems in adult animals, such as changes in body weight, or gross and/or histopathological changes in organs

Toxicity can be measured in terms of different levels of effects based on a dose-response assessment:

- The dose at which effects were first observed, lowest effect level (LEL), lowest observed effect level (LOEL) and low observed adverse effect level (LOAEL), and
- The doses at which no effects were observed, i.e. the no effect level (NEL), no observed effect level (NOEL) or no observed adverse effect level (NOAEL).

Point-of-departure (POD) is the point on the dose-response curve that marks the beginning of a low-dose extrapolation

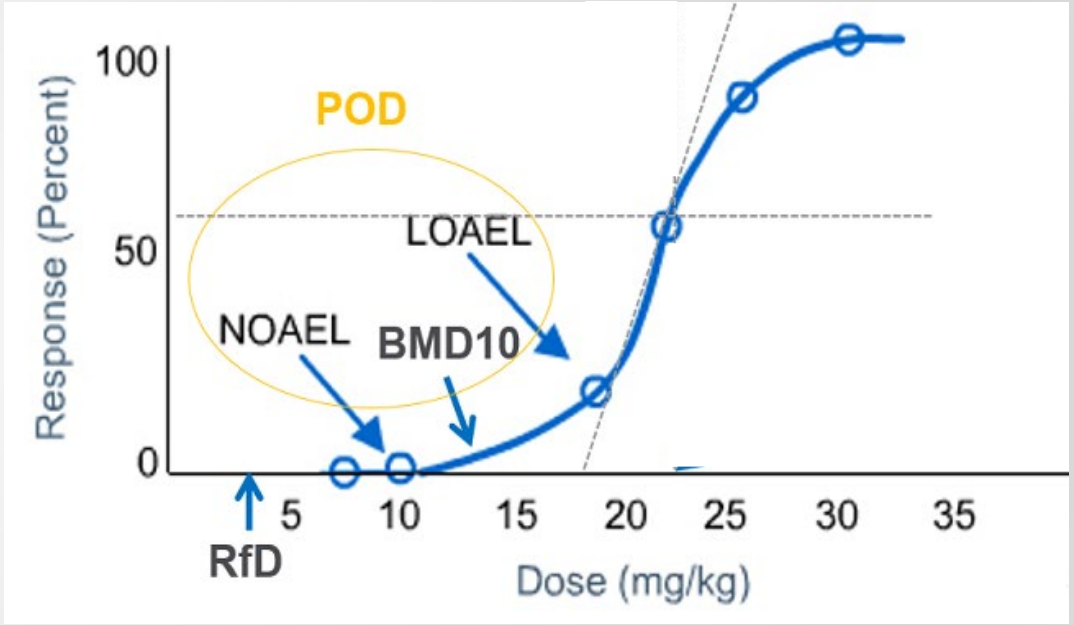


Image source: http://www.chemsafetypro.com/Topics/CRA/What_is_Point_of_Departure_(POD)_in_Toxicology_and_How_to_Use_It_to_Calculate_Reference_Dose_RfD.html

Repeat Dose Toxicity Data

US EPA's **ToxValDB,** a compilation of information on ~4000 unique chemicals from a variety of public data sources including:

- **ToxRefDB**
- IRIS
- PPRTV (ORNL)
- ECHA
- COSMOS
- CalEPA
- EPA
- ..and more.

Effect level types:

- LEL, LEC
- LOEL, LOEC
- LOAEL, LOAEC
- NEL
- NOEL, NOEC
- NOAEL, NOAEC
- BMD, BMC, BMC10
- BMDL, BMDL-01, BMDL-05, BMDL-10, BMDL-1SD, BMCL, 'BMCL-5', 'BMCL-10', 'BMCL-1SD'

Study Type	Species	Total number of effect level values	Number of unique chemicals with curated structure and descriptors
Chronic (CHR)	Rat	7172	1129
	Mouse	4029	720
	Rat, Mouse	11201	1236
Subchronic (SUB)	Rat	36017	3199
	Mouse	5030	723
	Rabbit	1516	415
	Rat, Mouse, Rabbit	42563	3306
Reproductive (REP)	Rat	5446	841
	Mouse	505	87
	Rat, Mouse	5951	889
Developmental (DEV)	Rat	6021	930
	Mouse	704	116
	Rabbit	3220	491
	Rat, Mouse, Rabbit	9945	1004
Subacute (SAC)	Rat	946	160
ALL (CHR, SUB, REP, DEV, SAC)	All (Rat, Mouse, Rabbit)	71020	3632

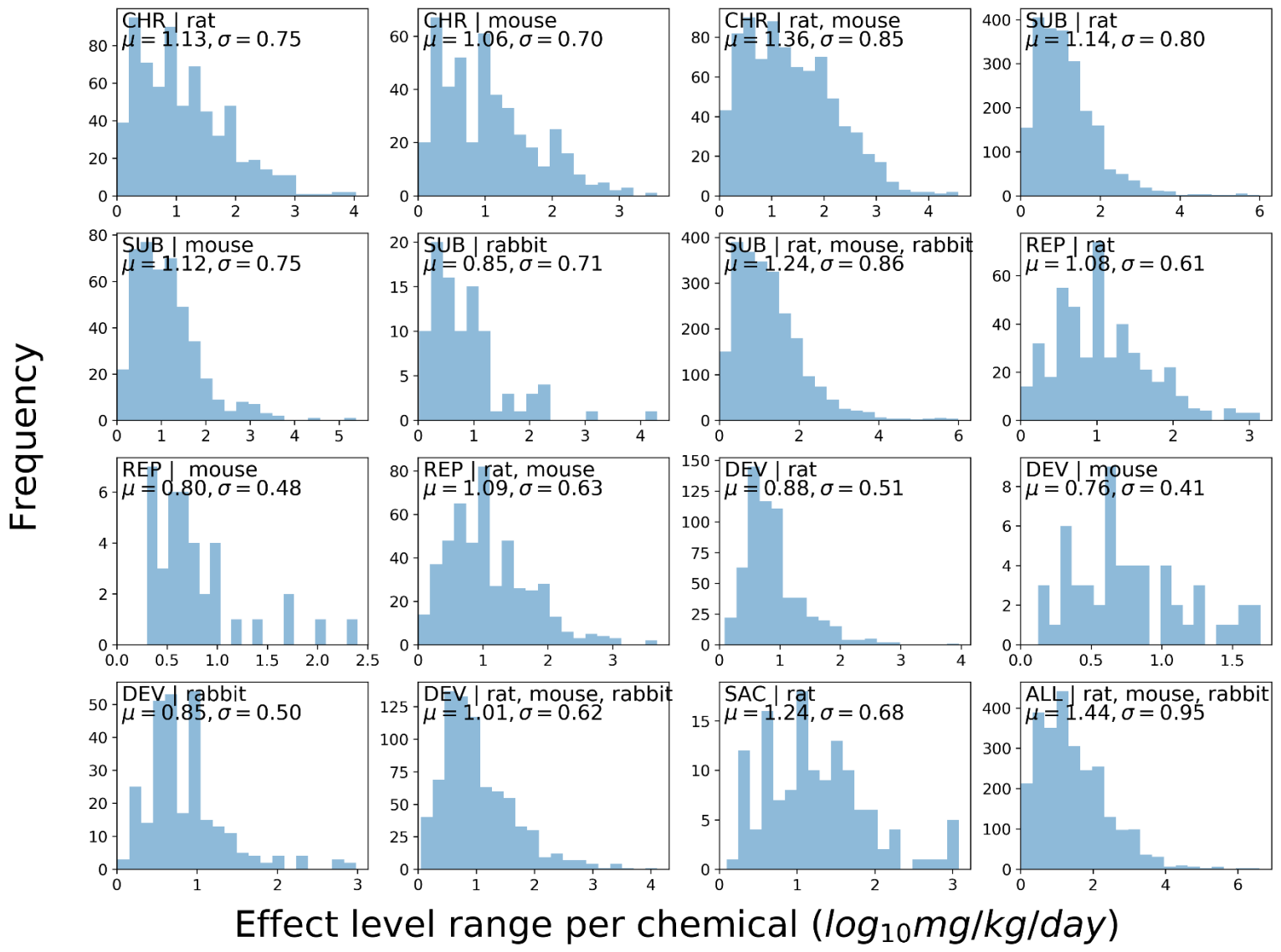
Underlying Data Variability

Experimental Variability

- Data from different labs (sources) running the "same" experiment may get different answers
- Sources of variability: biological (e.g., test species, environmental conditions) and/or technical (e.g., measurement errors, different experimental protocols)

Distribution of the range of effect level values per chemical as obtained from the ToxValDB database for each study type combination.

- The distribution of effect level values can range from 0-4 $\log_{10}$mg/kg-day for all dataset combinations
- Average range about 1 $\log_{10}$mg/kg-day

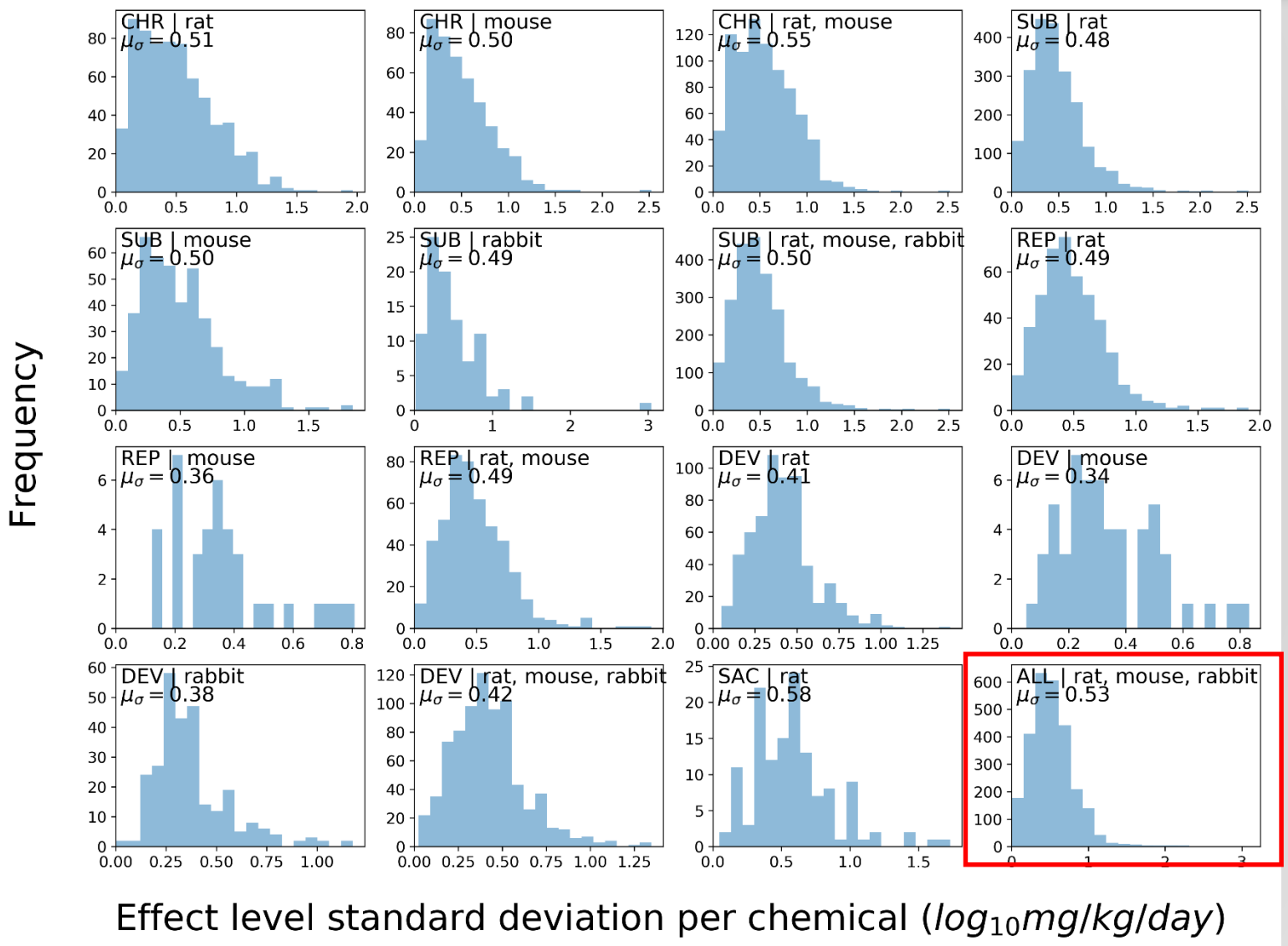


Underlying Data Variability

Distribution of the standard deviation (σ) of the effect level 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 which limits the predictive ability of any model developed on this data
- μ_σ is used as an estimate of theoretical lower bound on RMSE values for the QSAR models developed using these data.

For example, **expected RMSE** ≥ the mean standard deviation of effect level = **0.53 log_{10}-mg/kg/day (**for the highlighted **ALL combination.)**



Data Variability -> Model Uncertainty

Model Uncertainty

A model gives a result (a POD), but this is an estimate of the "true" POD. The true POD is unknown. Uncertainty in the evaluation data will lead to uncertainty in the model and our estimate of its quality. Incorporation of variability in computational model development, and subsequent quantification of data-driven uncertainty in model predictivity, are critically needed to improve the reliability and acceptance of computational models for screening level risk assessment.

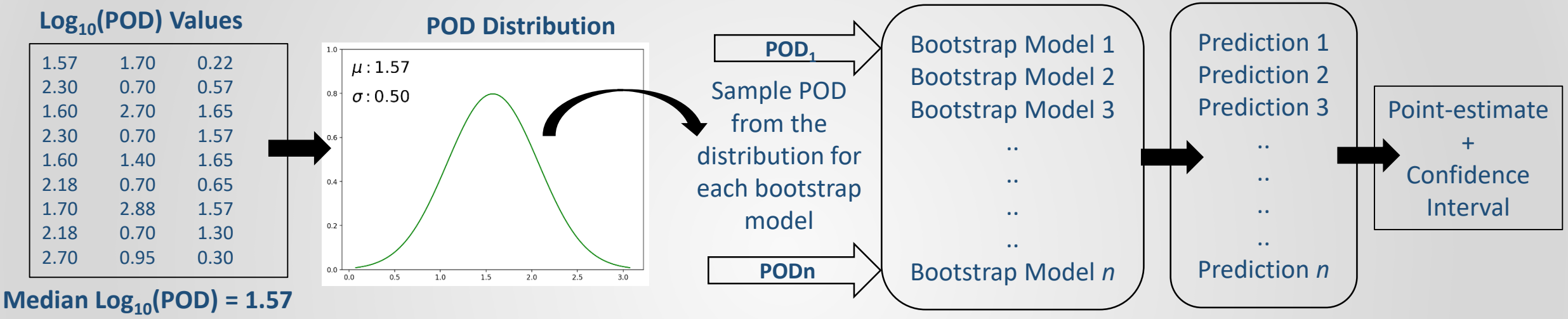
Point-estimate with confidence interval models

QSAR predictions as a confidence interval may be useful in understanding not only model performance but also in performing preliminary safety assessments, where rapid identification of a range of doses for a putative POD would enable rapid estimation of hazard to exposure ratios to identify chemicals for which additional information would be informative

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 $\log_{10}$-units)
- *100* bootstrap models were built with random sampling of POD values for each chemical from the pre-generated POD distribution.
- Predicted POD_{QSAR} = mean of 100 bootstrap predictions
- Confidence interval of POD_{QSAR} = ±2 standard deviation of 100 bootstrap predictions

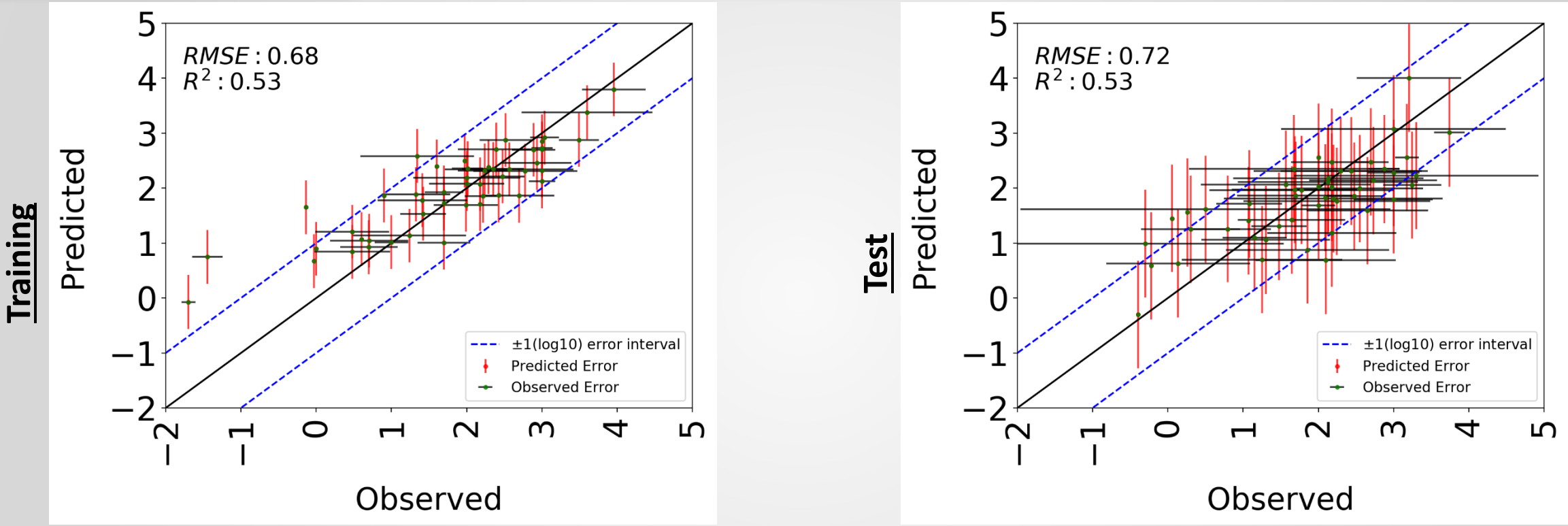
Quantifying Model Uncertainty



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

1. 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).
2. For each (of *n = 100*) bootstrapped models the POD value for Bisphenol A is randomly drawn from the pre-constructed POD distribution.
3. Each cross-validated bootstrapped model predicts a POD value resulting in *n* POD predictions. The final point-estimate POD value is the mean of *n* predictions and the confidence interval is derived as the one standard deviation of *n* predictions.

Results

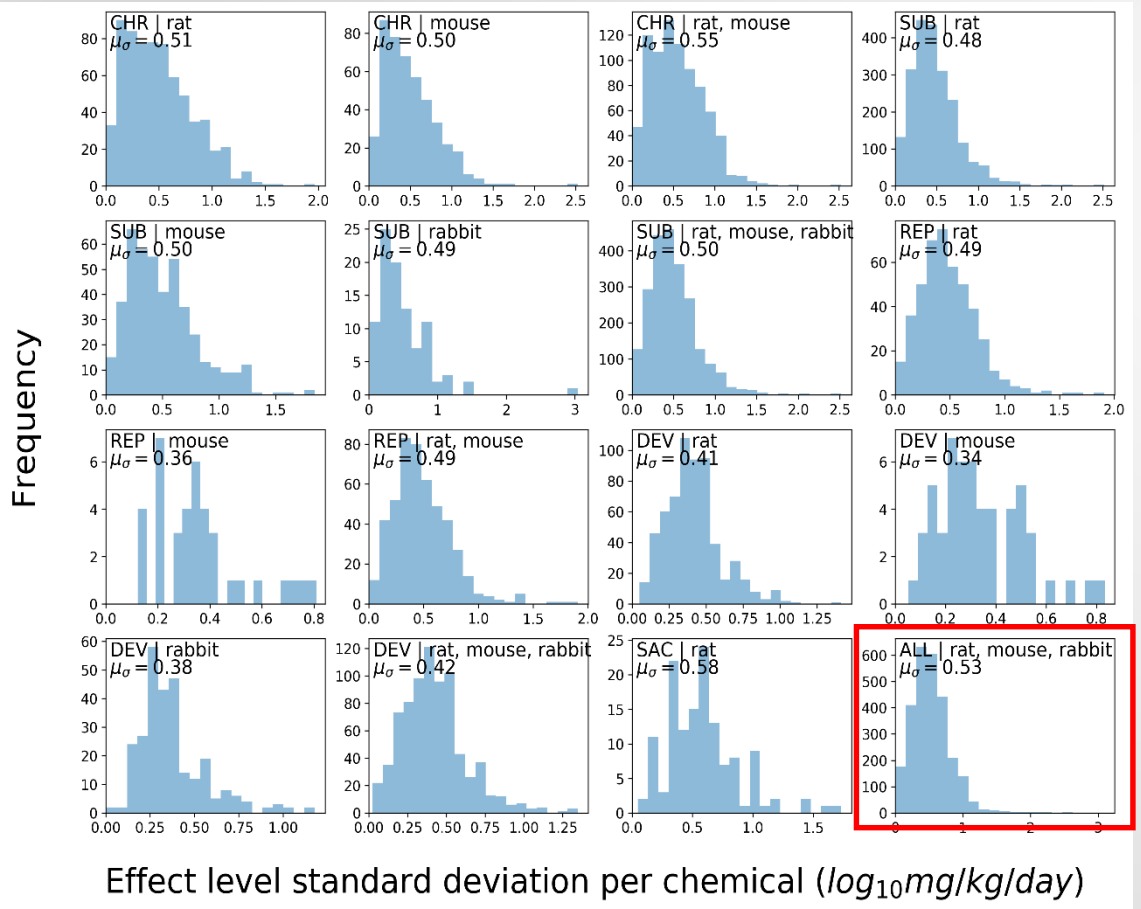


Bootstrap performance results for 50 (random) chemicals with the observed and predicted confidence intervals

- The predicted 95% confidence interval (error bar) for each chemical is calculated as two standard deviations of the predictions from the models.
- The observed 95% confidence interval (error bar) is calculated as two standard deviations of the experimental data for each chemical.

Conclusions and Outlook

Minimum expected RMSE = The mean standard deviation per chemical is **0.53 log_{10}-mg/kg/day**
Actual: RMSE = 0.72, R2 = 0.53



Pham, L.L., et al., *Variability in in vivo studies: Defining the upper limit of performance for predictions of systemic effect levels.* Computational Toxicology, 2020.

- Variability in *in vivo* toxicity studies limits predictive accuracy of NAMs.
- Total variance in systemic effect levels and the fraction explained were quantified.
- **Maximal R-squared** for a NAM-based predictive model of systemic effect levels may be **55 to 73**%; i.e., as much as 1/3 of the variance in these data may not be explainable using study descriptors.
- **The estimate of variance (RMSE)** in curated LELs and/or LOAELs approaches a **0.5 log10-mg/kg/day**.

III. Limitations on Predictivity of QSAR Models Owing to Underlying Data

Pradeep, P. et al. *Integrating data gap filling techniques: A case study predicting TEFs for neurotoxicity TEQs to facilitate the hazard assessment of polychlorinated biphenyls.* Regulatory Toxicology and Pharmacology. 2019.

Polychlorinated Biphenyls and Neurotoxic equivalency

Polychlorinated biphenyls (PCBs) are persistent organic pollutants associated with many adverse outcomes, including developmental neurotoxicity

- PCBs share a common scaffold and differ only in the position or degree of chlorination leading to 209 unique congeners
- Human exposure to PCBs happens primarily through inhalation and dietary sources of environmental PCB mixtures
- Risk assessment of PCB mixtures is challenging because very few PCBs (congeners) have been evaluated *in vivo*

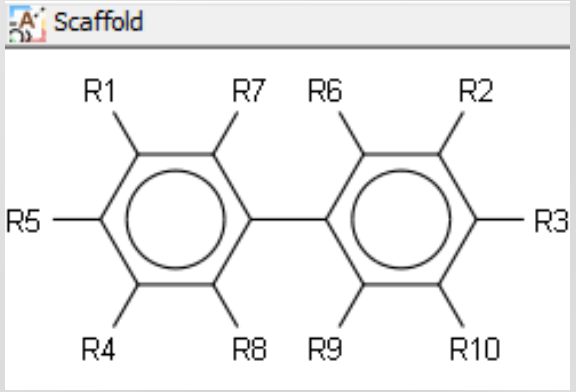
Toxic equivalency factor (TEF) is a way to express the toxicity of dioxins, furans and PCBs in terms of the most toxic form of dioxin, 2,3,7,8-TCDD. TEF approach can be used to predict **Neurotoxic equivalency (NEQ)** values from *in vitro* neurotoxicity data that can be used as an alternative to missing *in vivo* data for mixture assessments.

Experimental data

- NEQ value from each tested congener from 8 *in vitro* data sources for a total of 87 congeners (Simon et al., 2007)

Objective

- Develop a QSAR model to predict NEQ values for the 122 untested PCB congeners

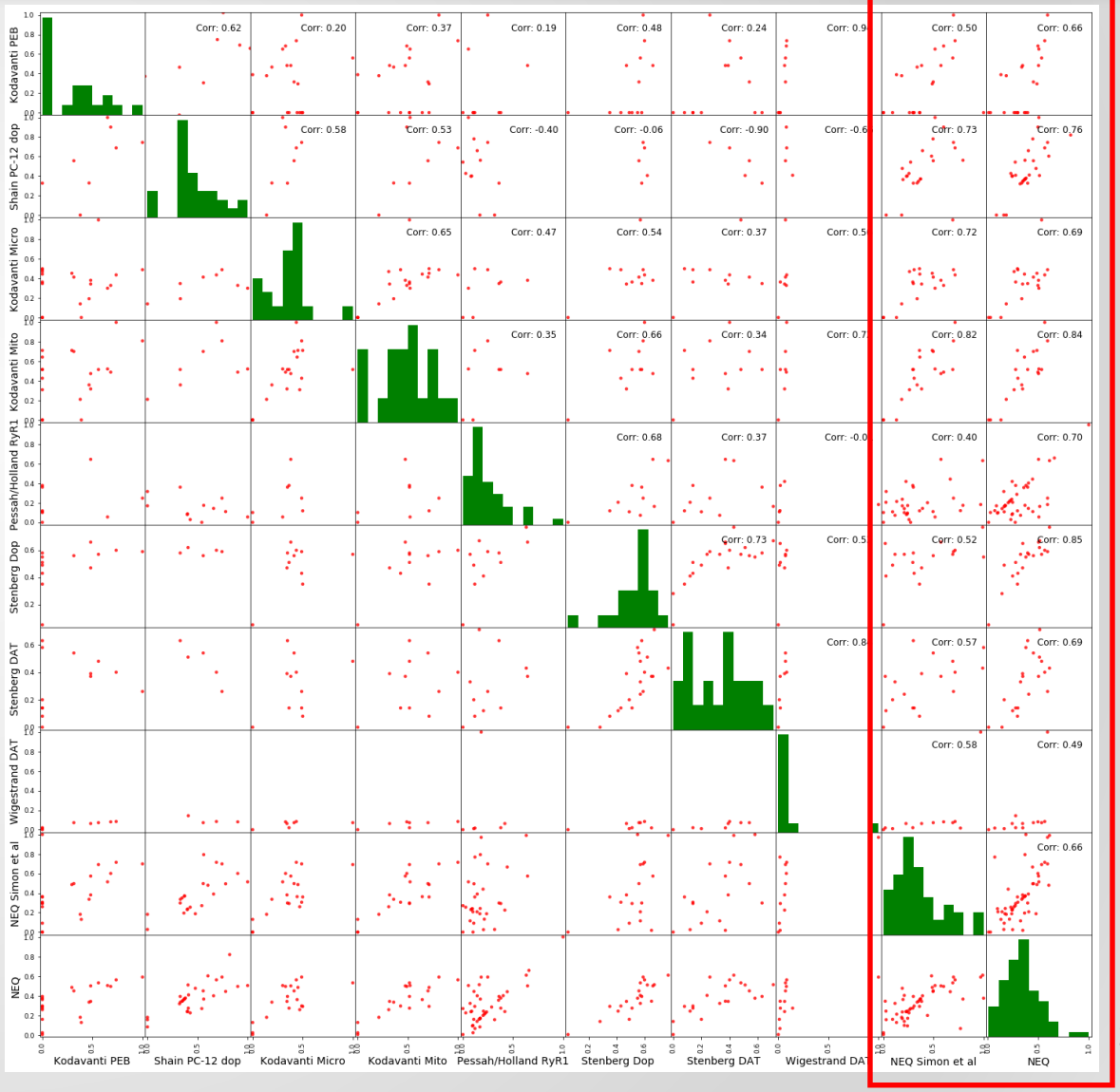


Neurotoxicity Data

Scatter plot matrix demonstrating the correlation between

- Experimental values from 8 experimental sources,
- NEQ values derived by Simon et al. (Simon et al., 2007), and
- NEQ values derived in our work

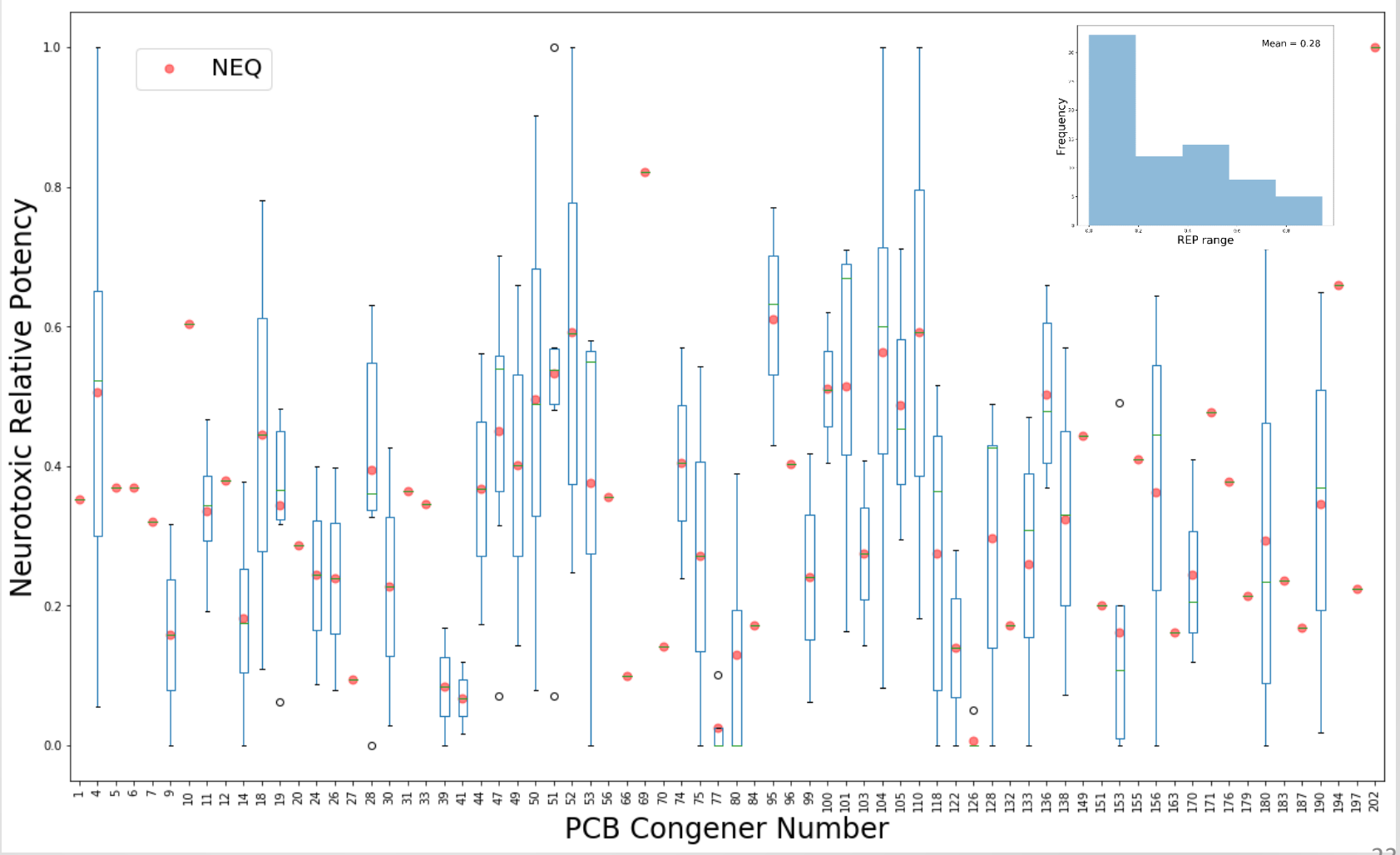
The Pearson's correlation coefficient for each pair of values indicates how poorly correlated different experiments in their measurement of neurotoxic effects for the PCBs



Data Variability Analysis

Box plot distributions of the NEQ values derived using experimental data for each PCB congener

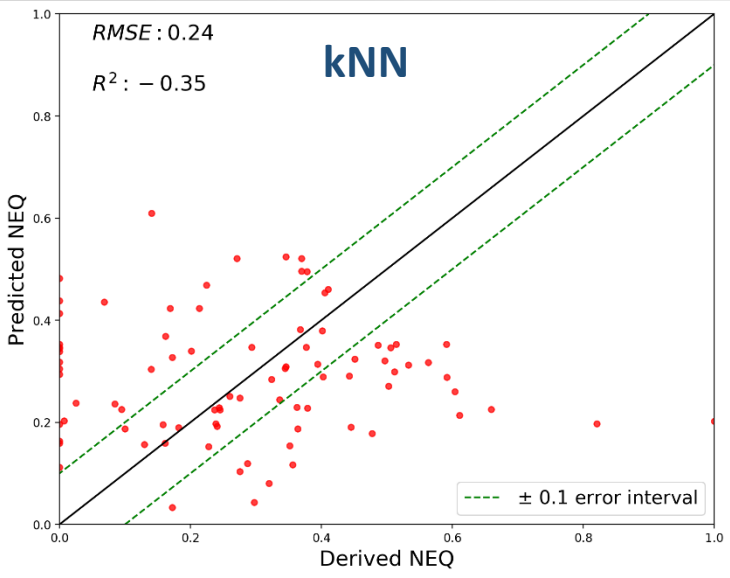
- The experimental values are represented by the boxplot
- The red dots superimposed on each box plot corresponds to the derived NEQ value for each PCB congener
- There is high variance in the NEQ values for congeners which have data from more than one experimental source
- The figure in the inset shows the distribution of the range of NEQ values (mean range = 0.28)

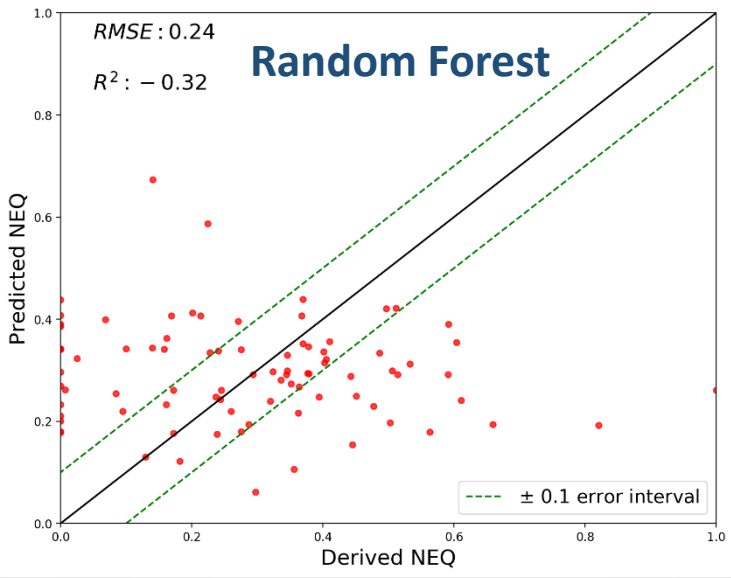


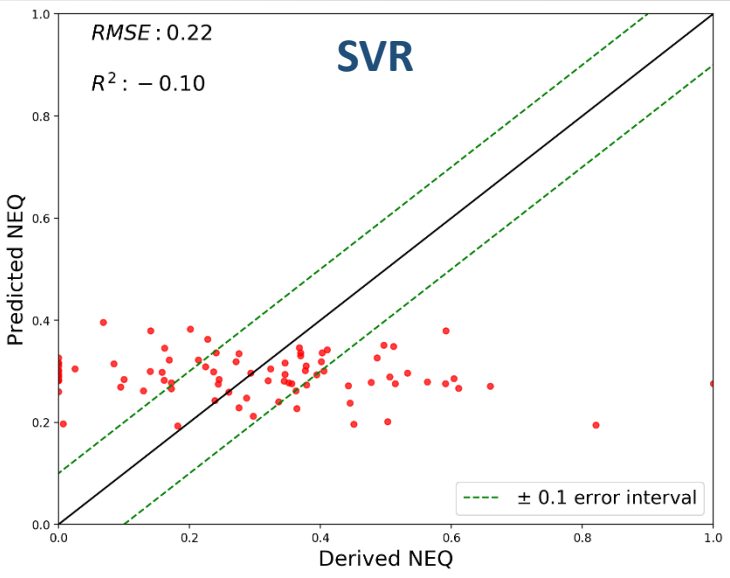
QSAR Model Results

QSAR models were developed using 3 different machine learning algorithms - *k* nearest neighbors (kNN), Random Forests and Support vector regression (SVR)

Even though the RMSE values are within the variability of NEQ values in the underlying dataset, the negative values of R^2 indicate that the QSAR models show no predictivity







Comparison of the derived neurotoxicity equivalent factors (NEQs) with the predicted values using 5-fold cross validation. The black solid line denotes perfect correspondence between derived and predicted values. The green dotted lines denote the ±0.1 error interval. The legend on top left in each graph shows the RMSE and R^2 values from each method.

Conclusions and Outlook

The failure to build robust and reliable PCB QSAR models for NEF prediction can be attributed to two major factors:

1. The quality of data and derived NEQ values

- The experimental data from each of the sources do not have high concordance with each other
- The data used to derive the NEQ values are taken from 8 datasets obtained from several experimental sources. There are multiple mechanisms through which PCBs can exert their neurotoxic effect, and each individual experiment used here measured PCB neurotoxic potential via a different mechanism

2. Quantity of available experimental data

- There are a limited number of PCBs that were tested in each of the experimental assays. The number of data points limits the ability of a machine learning algorithm to learn the structure-activity relationships well.

The derived NEQs and the QSAR predicted NEQs to fill data gaps for PCB risk assessment should be used keeping these facts in mind

Thanks for Listening!

Acknowledgements
Ly Ly Pham
Richard Judson
Katie Paul-Friedman
Grace Patlewicz
Laura Carlson
Geniece M Lehmann

Office of Research and Development
Center for Computational Toxicology & Exposure (CCTE)
Bioinformatic and Computational Toxicology Division (BCTD)
Computational Toxicology and Bioinformatics Branch (CTBB)