Integration of Different Data-gap Filling Techniques to Facilitate Assessment of Polychlorinated Biphenyls: A Proof of Principle Case Study

Prachi Pradeep[1], Laura Macaulay[2], Geniece Lehmann[2], Richard S Judson[3], Grace Patlewicz[3]

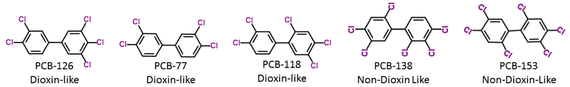
[1]ORISE, National Center for Computational Toxicology, RTP, NC
[2]National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, RTP, NC
[3]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, RTP, NC

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

www.epa.gov

BACKGROUND

Polychlorinated biphenyls (PCBs) are persistent organic pollutants associated with many adverse outcomes, including developmental neurotoxicity. A challenge in the risk assessment of PCBs includes differences in congener profiles among PCB mixtures found in the environment. These differences are important because each of the 209 PCB congeners has a unique structure and chemistry, leading to multiple modes of action (MOAs). In order to perform a human health risk assessment accounting for the toxicological activity of all PCB congeners comprised in environmental media, tools are needed to estimate the potency of each congener, including those that are relatively untested. PCB congeners can be broadly differentiated as dioxin-like (DL) or non-dioxin-like (NDL). DL PCBs activate aryl hydrocarbon receptor (AhR) signaling and elicit toxicity similar to the most toxic dioxin 2,3,7,8 Tetrachlorodibenzodioxin (TCDD). NDL PCB congeners (which make up the bulk of most PCB mixtures) have been less thoroughly tested, but there is evidence of neurotoxicity.



Example structures of DL and NDL PCB congeners.

Data gap filling techniques are commonly used to predict hazard in the absence of empirical data. The most established techniques are read-across, trend analysis and quantitative structure-activity relationships (QSARs). For PCBs, the target of data gap filling are TEFs (Toxic equivalency factors) which are applied to estimate relative potencies for mixtures of PCBs that contribute to an adverse outcome through a common biological target. For example, The TEF approach has been used for dioxin-like effects comparing individual chemical activity to that of TCDD.

MOTIVATION

Mechanistic data may be useful for estimating the relative toxicological activities of constituent congener in a given PCB mixture (commercial or environmental). However, very few PCB congeners have been evaluated *in vivo*. Thus, in practice neurotoxic equivalency (NEQ) values predicted from *in vitro* toxicity data can be used as an alternative to missing *in vivo* data for mixture assessments.

The aim of this case study was to determine whether a QSAR model can be used to predict NEQs and improve the predictive outcome for the assessment of a set of polychlorinated biphenyl (PCB) congeners and their mixtures. The dataset comprised 209 PCB congeners, out of which 87 altered in vitro Ca(2+) homeostasis from which NEQ values (range: 0-1) were derived.

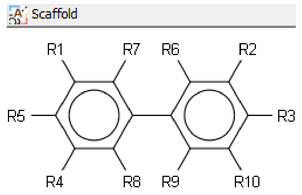
OBJECTIVE

Develop a QSAR model to predict NEQ values for the 122 untested PCB congeners to improve concordance between predictions based on *in vitro* assays and *in vivo* data

METHODS

FEATURE SET

- PCBs share a common biphenyl scaffold and differ only in the position and number of Cl substitutions.
- They can be adequately represented using a custom fingerprint where each bit encodes presence or absence of Cl substitution at each position
- Three position based fingerprints were explored



Fingerprint Number	Description	Number of bits	Bit representation details
1.	Each substitution position	11	1-10: R1 – R10 11: Total no. of Cl substitutions
2.	Ortho, meta, and para positional equivalency	4	1 (Ortho): R7, R7, R6, and R9 2 (Meta): R1, R4, R2, and R10 3 (Para): R5 and R3 4: Total no. of Cl substitutions
3.	Same ring positional equivalency	6	1 (Ortho ring 1): R7 + R8 2 (Ortho ring 2): R6 + R9 3 (Meta ring 1): R1 + R4 4 (Meta ring 2): R2 + R10 5: Total no. of Cl substitutions

MACHINE LEARNING ALGORITHM: Decision Trees

- Decision trees are a non-parametric supervised learning method used for classification and regression
- A decision model predicts the value of a target variable by learning simple decision rules inferred from the data features
- Suitable for the PCB problem since positional substitution affect biological properties
- Varied choices for tree parameters for tree growth and pruning to avoid over-fitting

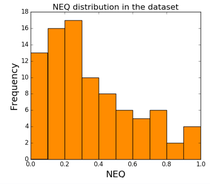
MODEL VALIDATION

External: 5-fold cross validation
Internal: Leave one out cross-validation

PERFORMANCE METRIC

Root mean squared error (RMSE)

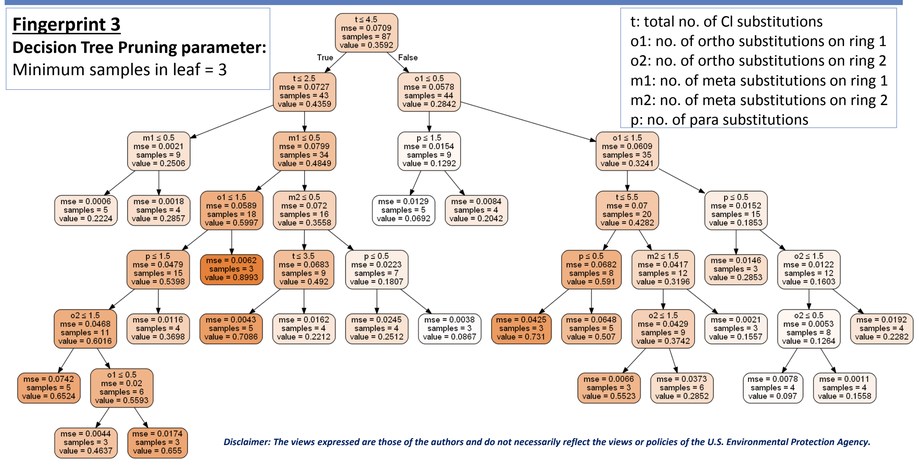
DISTRIBUTION OF NEQ VALUES IN THE TRAINING DATASET



DECISION TREE

Fingerprint 3
Decision Tree Pruning parameter:
Minimum samples in leaf = 3

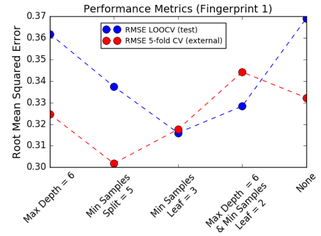
t: total no. of Cl substitutions
o1: no. of ortho substitutions on ring 1
o2: no. of ortho substitutions on ring 2
m1: no. of meta substitutions on ring 1
m2: no. of meta substitutions on ring 2
p: no. of para substitutions



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.

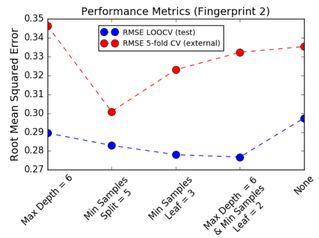
RESULTS

Performance metrics of the decision tree model using all three fingerprints and varying parameters in the decision tree model



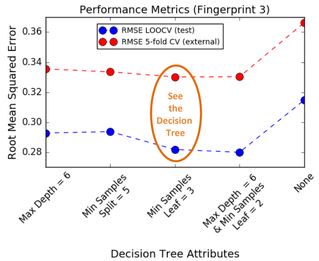
Fingerprint 1 Pruning parameter: Minimum samples in leaf = 3	
RMSE Training (LOOCV)	0.32
RMSE Validation (5-fold CV)	0.32

Considering all positions in the fingerprint to over-fit the model



Fingerprint 2 Pruning parameter: Minimum samples split = 5	
RMSE Training (LOOCV)	0.28
RMSE Validation (5-fold CV)	0.30

Considering all ortho, meta and para positions equivalent seem to overestimate equivalency between substitution positions



Fingerprint 3 Pruning parameter: Minimum samples in leaf = 3	
RMSE Training (LOOCV)	0.28
RMSE Validation (5-fold CV)	0.33

Considering ortho and meta positions equivalent on each ring and both para positions equivalent seem to be the best choice of fingerprint

CONCLUSIONS

This case study presents a simple QSAR model to predict molecular toxicity employing scaffold-based features using PCBs and NEQ values as an example dataset and endpoint, respectively

Feature set

- Using all positions in the fingerprint results in an over-fitted tree
- Using ortho, meta, para substitutions appears to overestimate equivalence between substitution positions
- Considering that some positions may not be equivalent based on other substitutions, same ring equivalence seems to be the best choice of fingerprint for the PCB problem

QSAR Model

- Decision tree model is simple to understand, transparent, and results in mechanistically interpretable predictions