Evaluation and Refinement of Quantitative Structure-Use Relationship (QSUR) Models for Chemical Function

Victoria Hull[1,2], Lindsay Eddy[1,2], Katherine Phillips[2], Kristin Isaacs[2]

[1]Oak Ridge Associated Universities, [2]U.S. EPA Office of Research and Development, Center for Computational Toxicology and Exposure

Email: hull.victoria@epa.gov, eddy.lindsay@epa.gov, issacs.kristin@epa.gov, phillips.katherine@epa.gov

HIGHLIGHTS

- Quantitative Structure Use Relationship (QSUR) models are classification models that use the structure of chemicals to predict their functional use.
- Model evaluation showed that our existing models underperformed when tested on chemicals with industrial use cases.
- New chemical use data was used to build refined QSUR models using harmonized function categories developed by the Organization for Economic Co-operation and Development (OECD).
- Descriptors for 14 product use categories (PUCs) were incorporated alongside chemical structure to allow stratification of function predictions by use case (e.g., in consumer products or in industry).
- Successful QSUR models (balanced accuracy>75%) could be built for most OECD functions.
- In these new refined models, expected use case influenced prediction of functional use.

BACKGROUND AND MOTIVATION

- The U.S. Environmental Protection Agency's (EPA's) Office of Research and Development previously developed high-throughput QSUR models that use the structure of chemicals to predict their functional use in consumer products and processes (Phillips et al., 2017).
- Evaluation of the *existing* QSURs against industry-reported consumer and industrial functional uses in EPA's Chemical Data Reporting showed that the models were successful at predicting industry-reported consumer uses. However, they underperformed on industrial chemicals that were outside the ADs of the QSUR models (Figure 1).

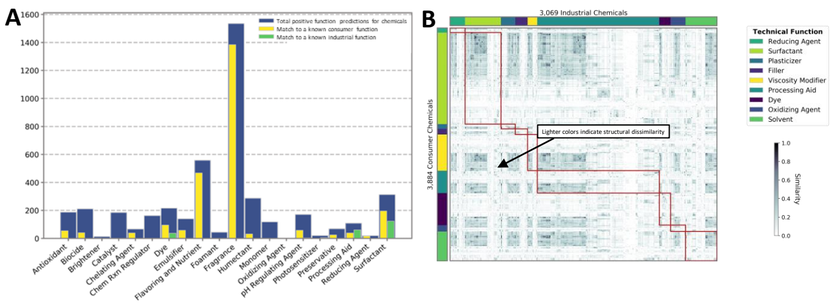

Figure 1: A) Existing QSUR models were not able to consistently identify true functional uses of CDR industrial chemicals (compared to true consumer chemicals). B) This is in part due to a lack of structural similarity between industrial and consumer chemicals with the same function.

- **Our objective was to build models that can estimate function for varying use cases (e.g., chemicals with different consumer uses or industrial uses)**

METHODS: NEW QSUR MODEL DEVELOPMENT

- Function information was curated from 146,421 documents in EPA's Chemicals and Products Database (Dionisio et al. 2018) that contained reported functional use; reported functions were harmonized to OECD categories.
- New Random Forest classification QSURs were built for 49 OECD harmonized categories having data for at least 20 chemicals; the "randomForest" and "parallel" R packages were used, and 5-fold cross validation was performed.
- Model descriptors included the structural "ToxPrint" descriptors (Yang et al. 2015), and 14 general Product Use Category (PUC) (Isaacs et al. 2020) descriptors that quantify reported occurrence in consumer product formulations and articles, occupational-related products, and other (industrial) uses. PUC descriptors were also obtained from CPDat.
- The importance of descriptors in predicting function was quantified via the mean decrease Gini score.
- Using the refined models, functional uses were predicted for a test set of chemicals having various hypothetical use cases.

Table 1: OECD functional use categories and chemical counts

Function	N	Function	N	Function	N	Function	N
Abrasive	23	Catalyst	3050	Filler	56	Pigment	331
Adhesion/cohesion promoter	35	Chelating agent	20	Flow promoter	142	Plasticizer	181
Adsorbent	295	Chemical reaction regulator	75	Fragrance	21	Plating agent	115
Anti-caking agent	31	Cleaning agent	89	Freeze-thaw additive	119	Polymerization promoter	1866
Antioxidant	38	Coalescing agent	27	Hardener	79	Solvent	148
Anti-redeposition agent	25	Corrosion inhibitor	385	Heat stabilizer	379	Surface modifier	137
Anti-scaling agent	22	Dehydrating agent (desiccant)	36	Humectant	67	Thickening agent	369
Anti-slip agent	122	Depilatory (EPA defined function)	96	Insulators	222	Tracer	561
Anti-stain agent	374	Diluent	199	Lubricating agent	28	UV stabilizer	169
Anti-static agent	74	Drier	108	Monomers	24	Viscosity modifier	35
Binder	21	Dust suppressant	363	Oxidizing agent	433	Waterproofing agent	241
Biocide	42	Dye	61	Pharmaceutical (EPA)	3264	Wetting agent	52
Brightener	592	Emulsifier	771				

RESULTS: REFINED QSUR MODEL PERFORMANCE

- 42 of 49 QSUR models had balanced accuracy > 75%.

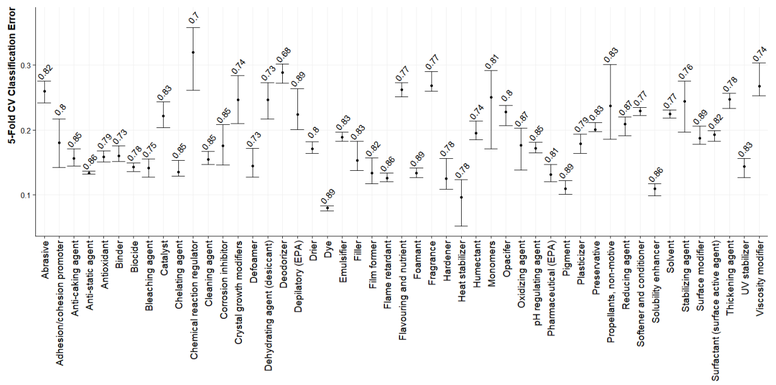

Figure 2. Average classification error for 5-fold cross-validated models. Error bars represent minimum and maximum error. Mean balanced accuracy is included above the error bars.

RESULTS: INFLUENCE OF USE CASE ON FUNCTION PREDICTION

- The mean decrease Gini score was used to determine feature importance for the 14 PUC and 729 ToxPrint descriptors. PUCs were among the top 20 important features for 32 function QSURs.

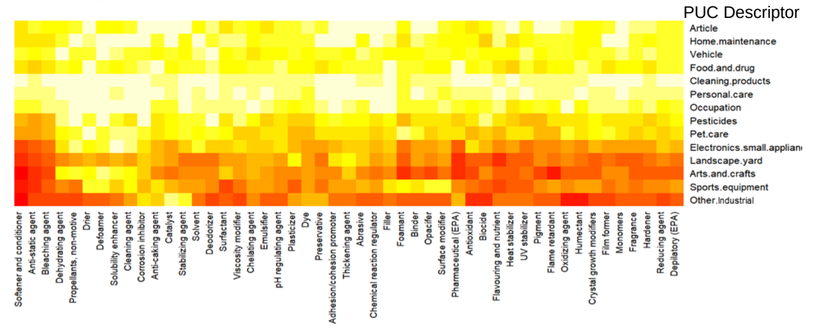

Figure 3: A heat map of feature importance for PUC descriptors. White features have a higher rank (more important) and red features have a lower rank (less important); importance of use case on function prediction for some QSURs is indicated by lighter areas.

- To demonstrate that use case is an influential component of QSUR prediction, function was predicted for 2,489 EPA CDR chemicals not in the QSUR training set assuming hypothetical use cases (e.g., use in consumer articles or industrial use).

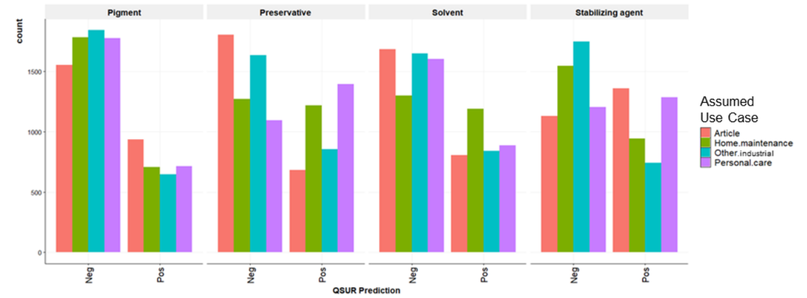

Fig 4. Positive and negative predictions in CDR for select QSUR models assuming different PUC descriptors (use cases).

CONCLUSIONS AND NEXT STEPS

- These preliminary results show that using OECD categories and PUC descriptors can result in valid QSUR models (good balanced accuracy).
- PUC descriptors rank consistently high in feature importance, indicating influence of use case on prediction of function.
- We will further evaluate the QSURs using y-randomization.
- We will assess conditional importance of predictors, allowing us to quantify the interaction of use case with structural descriptors.
- We will evaluate the final version of the refined QSURs with industrial data from regulatory partners to assess improvement in use case-specific predictions.

REFERENCES: