Predictive Models for Chemical Occurrence in Environmental and Biological Media

Lindsay Eddy[1,2], Katherine Phillips[1], Caroline Ring[1], Jon Sobus[1], Elin Ulrich[1], and Kristin Isaacs[1]

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

Kristin Isaacs | isaacs.kristin@epa.gov | 919-541-2785

Poster number: P647
Abstract Number: 3956

Abstract

Monitoring of chemical occurrence in various media is critical for understanding the mechanisms by which human and ecological receptors are exposed to exogenous chemicals. Since monitoring studies are expensive, data have not been exhaustively collected for the tens-of-thousands of chemicals in commerce. To fill this gap, predictive models can be used to anticipate chemical presence and inform prioritization for further study. Here we present a suite of random forest models which integrate data from dozens of public monitoring sources to predict chemical occurrence in 30 different environmental and biological media. For each medium, classifier models were built to predict the probability of any given chemical being detected in that medium. Training data for a robust classifier model must consist of examples both of chemicals that are and chemicals that are not present in the medium. However, the available training dataset disproportionately contains chemical detections; out of 30 media, 14 media had fewer than 5 true negative chemicals. To address this dearth of negative data, augmented models were built which use positive unlabeled learning to identify likely negative chemicals from an unlabeled data set (here the Toxic Substances Control Act active inventory). Likely negatives identified using the augmented models were then used to train final media models. Final 5-fold cross-validated models with a balanced accuracy of 75% could be built for 14 media. An initial validation of blood model with limited external data demonstrated an accuracy of 73%. Final versions of models for all media will be tested on identified external data sets to assess their ability to predict emerging environmental exposures. These models have the potential to inform the development of 1) workflows for environmental decision-making, and 2) methods for assessing unknown structures in non-targeted analyses of environmental and biological media.

Introduction

- Monitoring of chemical occurrence in various media is critical for understanding the mechanisms by which human and ecological receptors are exposed to exogenous chemicals. These data can inform regulatory decisions and mitigation strategies.
- Since monitoring studies are expensive, there are large gaps in occurrence data for the tens-of-thousands of chemicals in commerce. In addition, strategies are need to prioritize chemicals for measurement in different media.
- Here we present a suite of machine-learning models that integrate data from dozens of public monitoring sources to predict chemical occurrence in 27 different environmental and biological media. For each medium, classifier models were built to predict the probability of any given chemical being detected.

Approach

Our overall approach was to use available public monitoring data that has been harmonized to media and chemical identifier to train predictive machine learning random forest models for media occurrence. We employed strategies to address the lack of negative data for certain media.

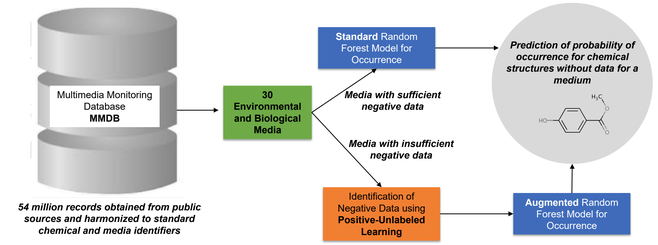


Figure 1: We developed random forest models for likely media occurrence by analyzing a database of chemicals previously measured in monitoring of various environmental and biological media. The monitoring dataset was used to train a suite of validated media-specific classification models which can be used to predict the probability of occurrence in a medium of any unknown chemical. Since negative data are lacking for many chemicals due to a reporting bias towards detections, we implemented positive unlabeled (PU) learning to identify likely negatives for some media.

Methods

- Training data for models of media occurrence were obtained from EPA's Multimedia Monitoring Database (available at https://doi.org/10.23645/epacomptox.17065024.v1). Chemicals in MMDB were harmonized to chemical identifiers and to 30 harmonized media. Data sources in MMDB include public databases developed by the U.S. EPA, the National Atmospheric Deposition Program, the state of California, the European Commission, N.C. State University, the U.S. Food and Drug Administration, the International Council for the Exploration of the Sea, the U.S. Centers for Disease Control, the U.S. Department of Agriculture, and the U.S. Geological Service.
- Classification models for occurrence in each medium were built using chemicals that were present (detected) and not present in MMDB. A chemical is considered **not present** if all its measurements were non-detects. Detect measurements are disproportionally represented in the MMDB. Thus, for many media, very few chemicals are "not present". (See Table 1.)
- To address this lack of negative data, we built **augmented** models for some media using positive unlabeled (PU) learning.[1] PU learning selects likely negative data from a library of **unlabeled** substances (those with no occurrence data).
- In the PU learning (Figure 2), 25 "weak learner" models were built using positive data and a subset of the unlabeled data and used to predict probability of occurrence (probability of being a positive) for all the other unlabeled data. The results from these models are averaged together, and chemicals that on average have the lowest probability of being a positive are identified as the **likely negatives**.
- These likely negatives were then used to train final augmented media models. In this study, unlabeled data were 23,339 commercial chemicals selected from Toxic Substance Control Act chemicals, food-related chemicals, pesticide active ingredients, pharmaceuticals, and cosmetics from lists provided on the EPA CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard/). The number of likely negatives selected was relative to the number of positives for each medium, such that the ratio of negatives to positives was 10:1.
- Standard models (for media with sufficient negatives), weak learner models, and final augmented models were built using the method of random forests.[2]

Medium	*# Chemicals Present*	*# Chemicals Not Present*
Ambient air	297	41
Aquatic invertebrate	377	33
Aquatic vertebrates/mammals	135	7
Birds	129	2
Breast milk	66	0
Drinking water	54	208
Fish	390	22
Food product	126	0
Groundwater	677	313
Human (other tissues or fluids)	54	1
Human blood (whole/serum/plasma)	164	3
Indoor air	77	0
Indoor dust	150	9
Landfill leachate	49	151
Livestock/meat	35	0
Other ecological media	45	0
Raw agricultural commodity	81	1
Sediment	626	237
Skin wipes	34	0
Sludge	84	15
Soil	68	8
Surface water	1359	346
Terrestrial invertebrates/worms	46	0
Terrestrial vertebrates	99	15
Urine	188	1
Vegetation	39	9
Wastewater (influent, effluent)	343	487

Table 1: The number of chemicals present and not present in the MMDB for media. Media with five or fewer chemicals "not present" are highlighted in blue; these media required PU learning and augmented models.

- Model descriptors for chemicals included ToxPrint molecular substructures,[3] high-level use classifications from EPA's Chemicals and Products Database (CPDat)[4], and EPA's OPERA physical-chemical property predictions.[5]
- Model validation was performed using Y-randomization;[6] model error was assessed using out-of-bag error and the area under the receiver operating characteristic curve (AUROC).[7]
- The chemical domain of applicability of each model was also assessed via standard methods comparing the distance of new chemicals to the chemical space of the training set.[8]

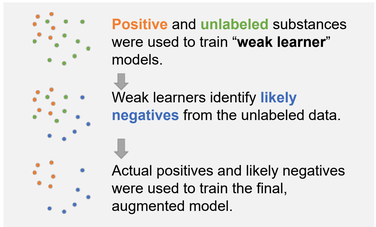
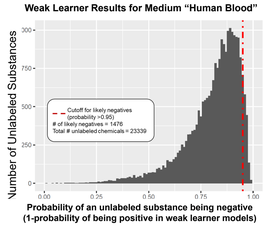


Figure 2: Illustration of how positive unlabeled (PU) learning is used to identify negative data for the media models. In this study, 25 weak learner models were built.

Results

- The PU learning successfully identified likely negatives for media without negative data (see example in Figure 3).
- Based on y-randomization validation, we could build valid prediction models for 14 media using standard models and 13 media using PU learning and augmented models (Figure 4).



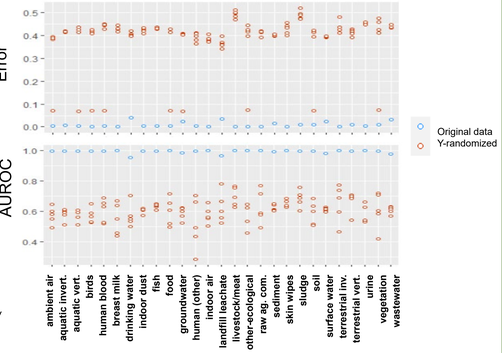


***Figure 3 (above)**: Performance of the weak learner models for identifying likely negatives from unlabeled substances for blood. The 1475 substances used to build the final augmented model for blood had a 95% probability of being negative.*

***Figure 4 (right)**: Out-of-bag error and AUROC for the final models for all media. Models built using the original data had low error and AUROC close to 1, indicating excellent ability of the models to discriminate positive and negative chemicals.*

- In an initial case study, we used the final model to predict chemicals likely to occur in drinking water by predicting occurrence for over 700,000 structures in EPA's DSSTox database.[9] Known uses for these chemicals obtained from CPDat and CompTox Dashboard lists are summarized in Table 2.

Sector of Use	Number of Chemicals
Pharmaceutical Use Only	60
Pesticide Active Use Only	6
Consumer Use Only	29
Industrial Use Only	123
Consumer and Industrial Use Only	41
Multiple Uses	31

***Table 2**: Chemicals with known uses that were predicted to occur in drinking water (probability > 0.50). The total counts of chemicals that were within domain and positively predicted are likely influenced by the chemical space of the training set for the media.*

Conclusions and Next Steps

- We could build successful standard and augmented models for predicting occurrence 27 environmental and biological media.
- We will perform external validation for models for which outside monitoring datasets can be identified. We will also incorporate available data from new non-targeted studies and investigate the feasibility of building machine-learning regression models that consider the quantitative frequency with which substances are detected in MMDB.
- Predictions for all 700,000+ structures in the EPA Computational Toxicology dashboard for all 27 media will be generated. These predictions may be incorporated into chemical decision-making workflows, e.g., prioritization of emerging chemicals of concern in drinking water and biosolids.
- These media occurrence models will also have utility in guiding structured literature searches and prioritizing chemicals for new monitoring studies.

References

1. Mordelet F., et al. *Pattern Recognition Letters 37*: 201-209, 2014.
2. Breiman, L. *Machine Learning*, 45: 5–32, 2001.
3. Yang C., et al. *J Chem Inf Model*, 55(3): 510–28, 2015
4. Dionisio K. et al. *Scientific Data* 5:180125, 2018.
5. Mansouri K., et al. *J Cheminform.* 10(1):10, 2018.
6. Rucker C., et al. *J. Chem. Inf. Model.* 47(6): 2345–2357, 2007.
7. Powers, D. *Journal of Machine Learning Technologies.* 2(1): 37–63, 2011.
8. Tropsha, A., et al. *Curr Pharm Des.* 13(34):3494-504, 2007.
9. Grulke, C., et al. *Computational Toxicology, 12, 2019.*