Predicting 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

Email: Eddy.Lindsay@epa.gov, Isaacs.Kristin@epa.gov

Introduction

- **Motivation**: 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, there are large gaps in occurrence data 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.
- **Multimedia Monitoring Database (MMDB)**:
  - EPA research database of measurements of chemical substances in dozens of environmental media
  - Includes measurements from over 20 public data sources
  - Contains over 250 million individual data records covering over 3200 unique chemicals
- **Media Models**
  - We are using chemical occurrence data from the MMDB to train predictive models.
  - For each medium, we build a random forest model which predicts chemical occurrence based on the chemical's structure.
  - We are investigating two main types of models:

Classification models

Binary prediction on whether a chemical **ever** occurs in the medium. *The classification models will be the focus of this poster.*

Regression models

Represents "severity" of occurrence. Models consider the *frequency* with which substances are detected in the MMDB.

Methods

- Random forest classification models for occurrence were built for media using data from MMDB
  - We consider a chemical to be **present** in a medium if the chemical has *ever* been detected in that medium (in the MMDB's records).
  - A chemical is considered **not present** if all its measurements are non-detects.
  - Detect measurements are disproportionally represented in the MMDB.
  - Thus, for many media, very few chemicals are "not present". (See Table 1.)

Medium	# Chemicals Present	# Chemicals Not Present
Ambient air	297	41
Aq. invertebrates	377	33
Aq. vertebrates	135	7
Birds	129	2
Blood	164	3
Breast milk	66	0
Drinking water	54	208
Dust	150	9
Fish	390	22
Food	126	0
Groundwater	677	313
Human - other	54	1
Indoor air	77	0
Landfill leachate	49	151
Livestock/meat	35	0
Other - ecological	45	0
Other - environ	4	0
Personal air	17	0
Precipitation	27	230
Raw agricultural	81	1
Sediment	626	237
Skin wipes	34	0
Sludge	84	15
Soil	68	8
Surface water	1359	346
Terr. invertebrates	46	0
Terr. vertebrates	99	15
Urine	188	1
Vegetation	39	9
Wastewater	343	487

***Table 1**: The number of chemicals present and not present in the MMDB for each medium. Media with five or fewer chemicals present are highlighted in blue.*

- To address this lack of negative data, we can build ***augmented*** models using positive unlabeled (PU) learning.
- PU learning uses **unlabeled** substances – these are substances outside of the MMDB for which we have no occurrence data.
- **Likely negatives**, selected from the unlabeled data, are used to train the final media model.
- Our unlabeled data was selected from the *TSCA Active Inventory.*

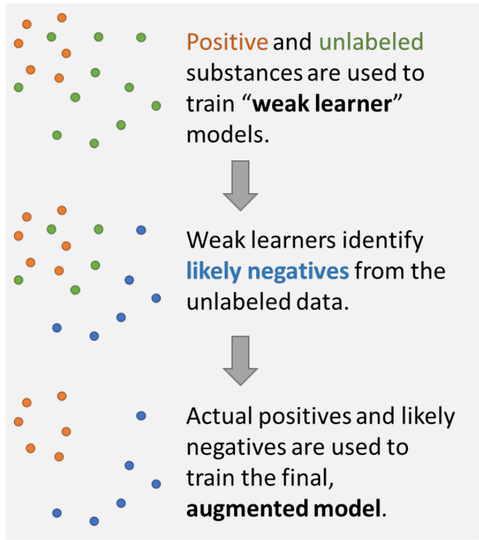

***Figure 1**: Illustration of how positive unlabeled (PU) learning is used to identify negative data for the media models.*

Results

Classification Model Performance

- To assess our models' performance, we look at their **out-of-bag (OOB) predictions**. OOB predictions represent the model's performance on chemicals outside the training set.

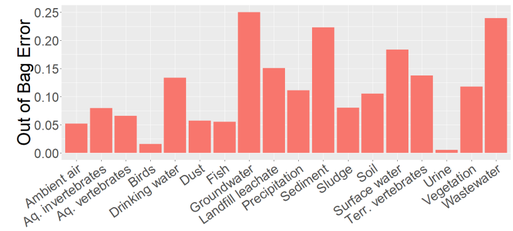

***Figure 2**: Out-of-bag error for non-augmented media models. For some media, non-augmented models could not be built due to insufficient negative data.*

- Results indicated we could build good models for media with sufficient negatives
- PU learning can be used to address remaining models

Case Study: PU learning applied to build model for blood

- We used PU learning to train an augmented blood model.

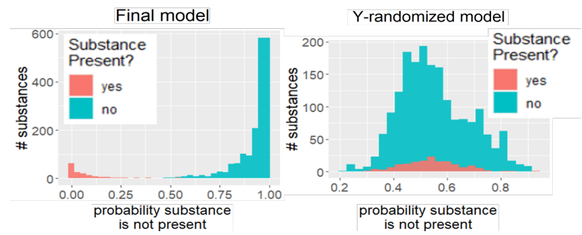

***Figure 3**: Histogram of out-of-bag predictions of the final augmented blood model and the y-randomized blood model.*

- The OOB predictions of the augmented blood model are much more accurate than those of the models trained on y-randomized (permuted) data. This indicates a generalizable model.

Next Steps

- Apply positive unlabeled learning to other media with insufficient negatives
- Test final classification models on data sets from the literature
- Incorporate model predictions into chemical decision-making workflows, e.g., prioritization of emerging chemicals of concern in drinking water and biosolids

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

AUGUST 30 - SEPTEMBER 2
ISES2021
VIRTUAL EXPERIENCE