AI/ML for PFAS Contamination Source Attribution

Debbie Fagan, Jessie Yaros, Danish Hussain, Eva Brayfindley, Charlotte Roiger, and Tim J Johnson, PNNL

Introduction

Historic use of per- and polyfluoroalkyl substances (PFAS) and subsequent concerns about toxicity and carcinogenicity have led to efforts to limit exposure[1]. Among the estimated 12,000 PFAS compounds that may be in use today, less than 100 can be identified analytically[2]. Artificial Intelligence/Machine Learning (AI/ML) models capable of detecting both known and emerging contaminants using limited historical data are of keen interest. In particular, the ability to discriminate among PFAS sources, such as Aqueous Film Forming Foam (AFFF) and other commercial formulations (CF), is of interest to facilitate responsible environmental management. This study uses AI/ML methods on mass spectral data to achieve source attribution.

Data

Data for this study comes from the NIST PFAS database that uses the Database Infrastructure for Mass Spectrometry (DIMSpec) toolkit and contains LC-MS/MS spectra for 104 PFAS samples resulting in a total of 7,194 high-resolution MS/MS spectra (Figure 1). All MS/MS data are transformed into fixed-length numeric encodings by using intensity binning.

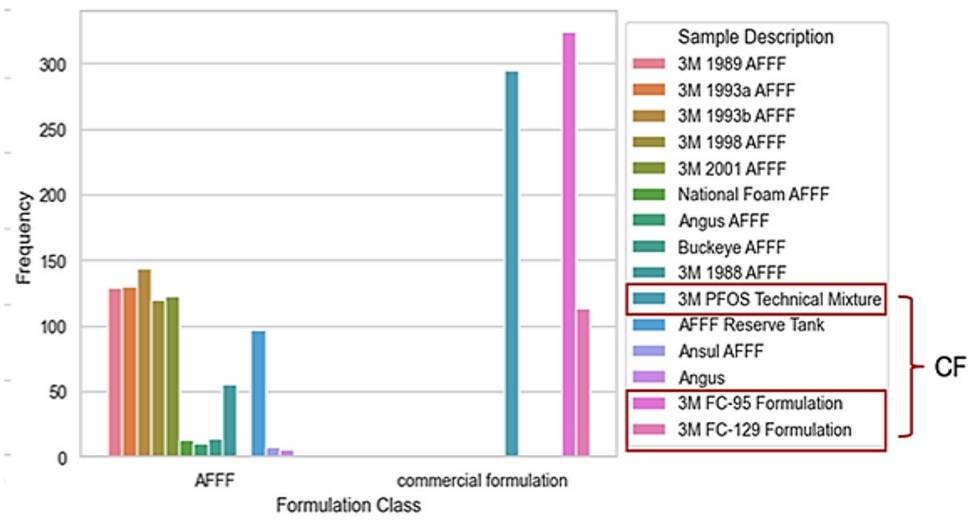

Figure 1. Number of MS/MS Spectra across AFFF and Commercial Formulation Subtype

Methods

1. Dimensionality Reduction

2. Classification

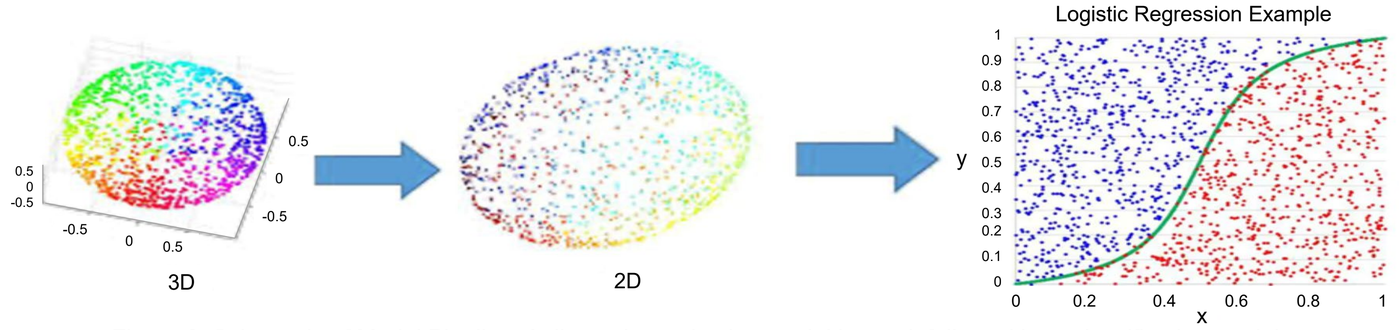

Figure 2. Schematic of Model Pipeline. A dimension reduction model is used, followed by a classification model

A series of three model pipelines were implemented with each consisting of a dimension reduction model followed by a classifier (Figure 2). Dimension reduction techniques include principal component analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). Three classifiers were employed: Linear Discriminant Analysis (LDA), Logistic Regression (LR), and Random Forests (RF). After testing multiple combinations of dimension reduction and classifiers, a combination of UMAP and RF models had the best performance for classifying AFFF and CF.

Results

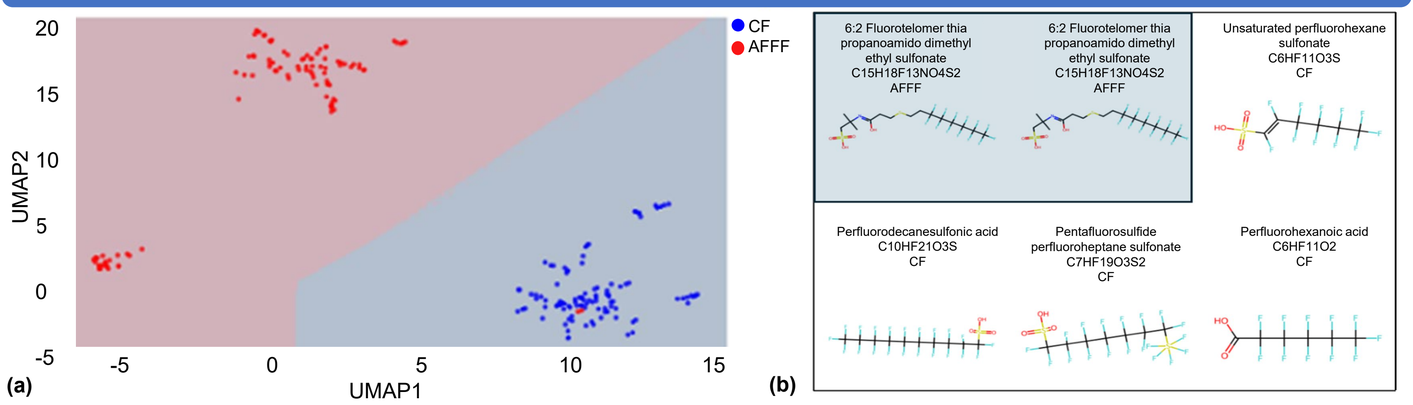

Figure 3. (a) Visualization of reduced dimension components of the classification model. (b) Parent compounds for the two misclassified AFFF MS2 fragments (boxed top left in grey) compared to other commercial formulations (CF).

After hyperparameter optimization, the model using UMAP and RF correctly distinguished AFFF from CF with 98% accuracy. Two AFFF samples were misclassified as CF, as shown in Figure 3(a). However, deeper investigation of sample compositions Figure 3(b) shows high similarity to other commercial formulations. Next steps will evaluate environmental samples with these models.

References: (1) USEPA, 2024a. PFAS Strategic Roadmap: EPA's Commitments to Action 2021-2024, United States Environmental Protection Agency, 2024. (2) USEPA, 2024b. Method 1633, Revision A, Analysis of Per- and Polyfluoroalkyl Substances (PFAS) in Aqueous, Solid, Biosolids, and Tissue Samples by LC-MS/MS. United States Environmental Protection Agency, 2024.

10/17/2025 | PNNL-SA-217163

U.S. DEPARTMENT of ENERGY BATTELLE
PNNL is operated by Battelle for the U.S. Department of Energy

For additional information, contact:
Debbie Fagan | (509)-371-7784 | Deborah.fagan@pnnl.gov

Pacific Northwest NATIONAL LABORATORY
www.pnnl.gov