Pacific Northwest
NATIONAL LABORATORY

PNNL-38833

AI for PFAS: Feasibility Scoping Study

Final Report

November 2025

Deborah K. Fagan
Fanny Chu
Charlotte N. Roiger
Kendall D. Hughey
Eva Brayfindley
Timothy J. Johnson
Thomas Danielson
Amie McElroy
Austin Coleman
Eugene Yan
Hassan Harb
Rajeev Surendran Assary
Jeremy Feinstein

U.S. DEPARTMENT of ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-38833

AI for PFAS: Feasibility Scoping Study

Final Report

November 2025

Deborah K. Fagan
Fanny Chu
Charlotte N. Roiger
Kendall D. Hughey
Eva Brayfindley
Timothy J. Johnson
Thomas Danielson
Amie McElroy
Austin Coleman
Eugene Yan
Hassan Harb
Rajeev Surendran Assary
Jeremy Feinstein

Prepared for
the U.S. Department of Energy
under Contract DEAC0576RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Per- and polyfluoroalkyl substances (PFAS) were originally of interest and of use because the fluorocarbons do not readily oxidize (burn) nor undergo other chemical changes such as hydrolysis. As such, their use in industrial processes became widespread, finding utility from fire suppression systems to non-stick coating applications and surface treatments. Due to the chemical stability of the carbon-fluorine (C-F) bonds, however, these “forever chemicals” can also persist in both soil and water for decades, often leaching into the groundwater adjacent to sites where they were used, including Department of Energy (DOE) and Department of Defense (DoD) sites. Concerns regarding their toxicity have led to subsequent efforts to limit exposure, as this class of chemicals has been correlated with certain forms of cancer, with onset often arising decades after original exposure. Materials laden with PFAS now require management and remediation. Whether it be monitoring releases to the environment from existing PFAS sources, identifying PFAS in an environment, or verifying PFAS destruction products, it is crucial to be able to quickly understand the PFAS signatures that result from various sources for several reasons, among them: (1) identifying and discriminating among PFAS sources to ensure responsible environmental management (EM) decision making, (2) determining the baseline condition that will be used to determine ecological and human health effects attributed to on-site sources, and (3) efficient verification of environmental removal or remediation of PFAS.

Given the environmental challenges associated with PFAS, a multi-organization team of experts from the Network of National Laboratories for Environmental Management and Stewardship (NNLEMS) was assembled to assess whether existing Artificial Intelligence/Machine Learning (AIML) approaches developed for other small molecule classes (as well as functional response data) can be successfully transferred to characterize PFAS. The team at Pacific Northwest National Laboratory (PNNL) (1) performed a survey of PNNL-developed AIML methods for small organic molecules identification, characterization and quantification from mass spectrometry (MS) data; (2) performed a survey of other relevant workflows that utilize multivariate signatures and features; (3) curated National Institute of Standards and Technology (NIST) PFAS MS datasets for AIML use; (4) retrained and tested existing AIML on curated NIST PFAS data. Proof-of-concept models were developed and optimized to predict high-resolution tandem mass spectra (MS2) detected from curated PFAS samples as belonging to either aqueous film forming foam (AFFF) and other commercial formulation (CF). The collection of models we applied range between 70% and 99% accuracy for these two classes of interest, but generalizability to environmental samples is yet to be studied. Our analysis demonstrated that non-linear dimensionality reduction techniques significantly outperform linear methods for PFAS class prediction, while systematic hyperparameter optimization and MLFlow experiment tracking established a robust framework for testing and evaluation. Though future work should continue to improve the AIML model, including investigation of generalizability, additional chemical complexity, and application to environmental samples; here, we advance toward a capability for automated PFAS contamination source identification in environmental samples.

Along with the PNNL tasking to use AIML for PFAS identification and attribution, both Argonne National Laboratory (ANL) and Savannah River National Laboratory (SRNL) contributed significant results in data science PFAS research. The goals of the ANL study were to (1) develop AI-assisted predictive models to estimate PFAS degradation potentials, using oxidation potential as an example, and (2) investigate their relationships with AFFF-sourced compounds and spectral data, specifically estimating reduction-oxidation (redox) potentials for individual PFAS compounds, and investigating the relationship of redox and oxidation potentials. The workflow included steps to (a) filter and clean 8214 PFAS spectra, (b) generate Morgan

fingerprint and chemical descriptors for the AI model, (c) estimate the redox potentials of PFAS molecules, and (d) perform a discrete Fourier transform (DFT) calculation for a random subset of 150 PFAS molecules. While this work and PNNL's effort both saw success using a Random Forest model, different configurations of this AIML model architecture were used in the two efforts (Random Forest and RandomForestUMAP, for ANL and PNNL, respectively). This framework facilitated ANL's ability to investigate the relationship of redox potential with various elements (e.g., N, O, S, Cl) that are contained in PFAS compounds, which can inform the prioritization and effectiveness with which these compounds can be degraded, thus facilitating remediation efforts from PFAS contamination.

For SRNL, the project scope included: (1) Identifying mass spectrometry datasets specifically within SRNL and at the Savannah River Site (SRS) related to PFAS in the environment, (2) performing data "cleaning" and standardization as needed such that ML algorithms can be applied, and (3) performing exploratory data analysis that offers comparison to NIST datasets to be used by PNNL during Year 1 of the AI for PFAS project (henceforth, Year 1) to identify substantial data differences. SRNL succeeded in this and organized a dataset of environmental PFAS mass spectra (analyzed using a DART-AccuTOF MS instrument) from environmental samples that were collected at various locations around the SRS. The samples were collected from rivers or lakes in multiple different areas of the SRS, including the General Separations Area, near former reprocessing facilities, and the SRS D-Area, where there was a firefighting training facility.

These capabilities represent advances in key elements in the environmental management of PFAS contamination.

Acknowledgments

This work is funded by the Department of Energy's Office of Environmental Management (DOE-EM). Follow up work may include AIML tools that demonstrate initial success in the scoping study which can be further fine-tuned and packaged for DOE-EM use. The authors gratefully acknowledge the financial support provided by the Department of Energy Environmental Management's Laboratory Policy Office (EM-3.2), its Senior Advisor for Laboratory Policy, Ming Zhu, and the technical guidance from April Kluever. Pacific Northwest National Laboratory is operated by Battelle for the DOE under contract DE-AC05-76RL01830.

The authors thank Jessie Yaros and Danish Hussain, Pacific Northwest National Laboratory, for their initial work on this project. Additional thanks to Amoret Bunn for her insightful advice in the preparation of this report.

Acronyms and Abbreviations

3M	Minnesota Mining and Manufacturing Company
AFFF	Aqueous film forming foam
AIML	Artificial Intelligence and Machine Learning
ANL	Argonne National Laboratory
CA	California
CF	Commercial formulation
C-F	Carbon-fluorine bond
$–COO^-$	Carboxylate
d3 MEFOSAA	N-Methyl-d3-perfluoro-1-octanesulfonamidoacetic acid
DART	Direct analysis in real time
DFT	density functional theory
DIMSpec	Database Infrastructure for Mass Spectrometry
DoD	Department of Defense
DOE	U.S. Department of Energy
ECF	Electrochemical fluorination
EM	Environmental Management
E_{ox}	Oxidation potential
GFN2-xTB	Geometry, Frequency, Non-covalent interactions, second-generation Extended Tight Binding
GSA	General Separations Area
HDPE	High density polyethylene
ID	Identification number
IR	Infrared
IS	Internal standard
ITRC	Interstate Technology Regulatory Council
IUPAC	International Union of Pure and Applied Chemistry
LC	Liquid chromatography
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
LDA	Linear discriminant analysis
LDA	Linear discriminant analysis
MPFOS	Perfluorooctanesulfonic acid 13C4
m/z	Mass-to-charge ratio
M2 PFOA	Perfluorooctanoic acid 13C2
MA	Massachusetts
MAE	mean absolute error
MEFOSAA	N-Methylperfluorooctanesulfonamidoacetic acid

mL	milliliter
Mm	millimeter
MO	Missouri
MS	Mass spectrum, or mass spectrometry
MS2	Tandem mass spectra, or fragmentation spectra
NC	North Carolina
NCI	National Cancer Institute
ng/L	Nanograms per liter
NIST	National Institute of Standards and Technology
NNLEMS	Network of National Laboratories for Environmental Management Stewardship
–OH	Hydroxyl
PB-X	Pen Branch
PCA	Principal components analysis
PCA	Principal components analysis
PFAA	Perfluoroalkyl acids
PFAS	Per- and polyfluoroalkyl substances
PFHxS	Perfluorohexane sulfonic acid
PFNA	Perfluoro-n-nonanoic acid
PFOA	Perfluorooctanoic acid
PFOS	Perfluorooctane sulfonate
PFPeA	Perfluoro-n-pentanoic acid
PFSA	perfluoroalkyl sulfonic acids
PNNL	Pacific Northwest National Laboratory
ppt	Parts per trillion
R^2	coefficient of determination
RF	Random forest
rpm	Revolutions per minute
SMILES	Simplified Molecular Input Line Entry System
SMOTE	Synthetic Minority Over-sampling TEchnique
S–O	Sulfur-Oxygen bond
SRNL	Savannah River National Laboratory
SRS	Savannah River Site
TOF	Time of flight
UMAP	Uniform manifold approximation and projection
USEPA	U.S. Environmental Protection Agency
UTR	Upper Three Runs

Table of Contents

Figures

Tables

1.0 Introduction

Per- and polyfluoroalkyl substances (PFAS) production was first established in the 1940s due to the chemicals having unique characteristics such as chemical stability, water and oil repellence, heat and oxidation resistance, and certain surfactant characteristics (Longendyke et al., 2022; Hughey et al., 2024). One of the largest direct uses of PFAS was by the military, as aqueous film-forming foams (AFFFs) — which contain PFAS — were found effective at extinguishing hydrocarbon fuel-based fires while also preventing reignition (Place and Field, 2012). There is widespread use in other mainstream consumer products such as food packaging, cosmetics, pesticides, paints, and cleaning products, to name just a few (Mahinroosta and Senevirathna, 2020; Cahuas et al., 2022); this rapid onset of industrial and manufacturing processes subsequently led to major environmental PFAS contamination (air, soil, water sources, as well as bioaccumulation in animals/livestock) (Mahinroosta and Senevirathna, 2020).

In recent years, the health effects of this previously unrestricted class of chemicals have become apparent. PFAS enter the human body through the digestive and respiratory systems as well as the skin (primarily through drinking water and food) (D'Hollander et al., 2010); they are not metabolized due to their strong carbon-fluorine and carbon-carbon bonds, and they demonstrate high absorption rates and low elimination rates, leading to accumulation in the body (Sznajder-Katarzynska et al., 2019). Blood serum samples taken from residents of all developed countries contain ppb levels of long chain perfluoroalkyl acids PFAAs, primarily perfluorooctanoic acid (PFOA) and perfluorooctane sulfonic acid (PFOS); the half-lives of these compounds in the human body are 2.1-8.5 y and 3.1-7.4 y, respectively (Goralczyk et al., 2015; Barzen-Hanson et al., 2017; Winquist et al., 2023). Historic use of PFAS and subsequent concerns regarding their carcinogenicity and toxicity have led to many efforts to limit exposure to these compounds (USEPA, 2024a; ITRC, 2023). The class of chemicals has been correlated to many forms of cancer, with onset often arising decades after original exposure (NCI, 2024). Some studies reveal an observed association between PFOA and kidney cancer (Winquist et al., 2023). Additional human adverse effects include hypertension (Pitter et al., 2020), hypercholesterolemia (Winquist and Steenland, 2014), developmental effects in children (Ames et al., 2025), decreased immune response to infection and vaccine (Bline, et al., 2024), and endocrine system interference (Ernst et al., 2019; Coperchini et al., 2021).

PFAS were originally of interest and of use because the fluorocarbons do not oxidize (burn), nor readily undergo other chemical changes such as hydrolysis. Due to the chemical stability of the carbon-fluorine (C-F) bonds, however, these "forever chemicals" can persist in both soil and water for decades, often leaching into the groundwater adjacent to sites where they were used, including U.S. Department of Energy (DOE) sites. Materials laden with PFAS now require management and remediation. Despite extensive research into alternatives, the approaches most widely implemented for their removal include thermal treatment in soils (i.e., incineration) and sorption mechanisms (e.g., granular activated carbon filtering, anion exchange, reverse osmosis, and nanofiltration) for liquids. These approaches have yet to be specifically optimized for PFAS-laden materials, however, and there remain many uncertainties as to the identity and fate of many of the F-laden products after treatment (Singh et al., 2019; Qin et al., 2024). While many byproducts are known, many are not, and these must also be identified to improve understanding of PFAS impact. Common analytical methods to assess fluorocarbon generation or detection both "in the wild" and during processes include liquid chromatography-mass spectrometry (LC-MS) and infrared (IR) spectroscopy (USEPA, 2024b; Hughey et al., 2024; Baker et al., 2024; Cui et al., 2024; Nahar et al., 2023). The IR and MS approaches, in

particular, are useful as they can detect PFAS with high sensitivity at high sampling rates. PFAS contamination is thus an acute challenge for environmental stewards.

Among the estimated 12,000 PFAS compounds that were in use and may still be in use today, a subset of less than 100 can be identified via analytical chemistry (ITRC, 2023; USEPA, 2024b). The remaining are considered potential emerging contaminants, and their impacts on ecosystems and human health are not yet fully known. Algorithms capable of detecting emerging contaminants with limited historical data or site knowledge are therefore of keen interest to the Environmental Management (EM) community.

Whether it be monitoring releases to the environment from existing PFAS sources, identifying PFAS in an environment, or verifying PFAS destruction products, it is in all cases crucial to be able to quickly understand the PFAS signatures that result from various sources for several reasons, among them: (1) discriminating among PFAS sources to ensure responsible EM decision making, (2) determining the baseline condition that will be used to determine ecological and human health effects attributed to on-site sources, and (3) efficient verification of environmental removal or remediation of PFAS.

With these considerations borne in mind in their responses to the PFAS problem, EM site managers at many facilities now seek to understand PFAS and its precursor chemical signatures, both to determine source attribution as well as to guide EM decision-making. To achieve such a goal, this report details how artificial intelligence and machine learning (AIML) techniques are utilized to successfully link chemical signatures of PFAS-containing compounds to source type via liquid chromatography with tandem mass spectrometry (LC-MS/MS) data using the National Institute of Standards and Technology (NIST) curated PFAS database within the Database Infrastructure for Mass Spectrometry (DIMSpec) Toolkit (Ragland and Place, 2023) (henceforth, the NIST PFAS database). The NIST PFAS database contains “clean” spectra, that is, well-annotated data for which detected signatures have been assigned and verified as belonging to known PFAS from analysis of standards and certified reference material samples. Next steps to enable site managers’ decision-making are to account for PFAS sample complexity in the AIML model, investigate generalizability of such a model to similar datasets, and then determine how and how well AIML methods can successfully be applied to actual environmental samples.

Of the various PFAS, PFOS and PFOA are the two most studied and toxic compounds contributing to PFAS contamination identified so far. They are a major source of PFAS contamination in the environment because of their widespread use in Class B AFFFs (aqueous film forming foams) used typically as fire or heat suppressants for flammable liquid fuel fires (ITRC, 2023). The 3M Company was the sole producer, via electrochemical fluorination (ECF), of PFOS-containing AFFF in use in the US between the mid-1960s and 1973 (ITRC, 2023). Commercial production ceased in 2002, but stockpile amounts may still be present at sites and fire departments across the country. Other AFFFs were produced via fluorotelomerization by different manufacturers (e.g., Angus, Buckeye, etc.) between the 1970s and 2016 (ITRC, 2023) and may contain PFOA. The 3M Company formulations are considered legacy AFFF, while other formulations may or may not be long chain legacy AFFF. In any case, AFFF is considered the primary source of environmental PFAS contamination and as such, the research efforts highlighted in this report focus predominantly on discriminating between 3M AFFF and other formulations. Future work will focus on PFAS source attribution in environmental samples, other PFAS source signatures, and algorithm development to detect emerging PFAS contaminants.

The larger effort was divided into separate research tasks executed by different DOE laboratories. Section 2 focuses on PNNL's effort to develop an AIML capability to predict NIST-curated mass spectra to relevant PFAS classes. Section 3 describes an effort by Argonne National Laboratory (ANL) to use AIML tools to help classify PFAS based on properties such as oxidation potentials. Section 4 summarizes Savannah River's efforts to capture and analyze real mass spectral data from the SRNL site. And finally, Section 5 provides key learnings across all three efforts and recommendations to guide environmental management and decision-making by continuing to advance capabilities for detection and characterization of PFAS contamination using AIML tools.

2.0 AIML Model for PFAS Class Prediction from Mass Spectrometry Data

To facilitate environmental management and guide decision-making where there is a need to identify sources of PFAS contamination, rapid methods of source determination (or attribution) are key. Adding to the complexity of the source attribution mission is that the formulations are often unknown, such that the individual chemicals—and combinations of chemicals—in these samples may not be fully elucidated. As such, it is critical to develop methods to determine PFAS source from the measured signatures of samples. To advance these aims, PNNL employed AIML techniques to demonstrate that chemical signatures from mass spectrometry data detected from these unknown formulations can be used to successfully distinguish PFAS-containing samples by source type.

Mass spectrometry provides rich analytical data representing chemical-specific signatures and its ability to detect broad classes of analytes with high sensitivity has led to widespread use in many chemical and biological applications, including the analysis of PFAS. AIML has also proven highly effective at learning and recognizing complex patterns in data without relying on human-enforced or conventionally applied rules for classification in numerous domains. As such, the application of AIML to better exploit the chemical signatures that we can obtain from mass spectrometry data of PFAS is ideal to tackle the complex problem of source determination of PFAS contamination.

As a proof-of-concept, we trained and evaluated models to predict PFAS source or class (i.e., AFFF, commercial formulations (CF)) from curated PFAS samples using mass spectrometry features, which we describe in detail below. While AFFF was used for firefighting activities, CF are formulations containing PFAS or PFOAs used in other areas of manufacturing or industry Evaluation of different classification algorithms and parameter optimization showed that non-linear dimension reduction using Uniform Manifold Approximation and Projection (UMAP) followed by Random Forest (RF) classifier was the most performant, demonstrating good accuracy to predicting mass spectra as originating from samples labeled as AFFF or CF. Further, the framework allows any misclassifications to be interrogated, thus providing a path for interpretability. Finally, the MLFlow framework established in this effort enables accessible, repeatable, and reproducible model development. This initial effort demonstrates the feasibility for AIML to begin to meet needs in environmental management of PFAS. Continued efforts to develop and improve upon AIML tools for PFAS class prediction will position us to better and more quickly address challenges with source determination of PFAS contamination.

2.1 Approach

We leveraged curated PFAS mass spectrometry data from the NIST PFAS database for model development, optimization, and evaluation. This section describes the rationale underlying the development and final selection of model architecture and parameters for classifying mass spectra as belonging to samples labeled as AFFF or CF.

2.1.1 Curated PFAS Data for Model Development

Data for this study come from the NIST DIMSpec Toolkit (Ragland and Place, 2023), which contains the NIST PFAS database—a SQLite database containing LC-MS/MS (also called MS2) spectra and corresponding metadata of PFAS. The database contains 104 samples, some of which are mixtures, resulting in representation of 131 unique chemical compounds with a total

of 7,194 high-resolution MS2 spectra. These spectra were collected using 12 different analytical methods, each employing fixed collision energies ranging from 15 to 60 volts, and include data acquired in both positive and negative ionization polarities. This database was developed to address critical challenges in identifying and categorizing PFAS in environmental samples, particularly using untargeted analysis approaches to characterize compounds without prior knowledge of which specific analytes might be present.

Data within the NIST PFAS database are segmented into data tables interconnected by a series of data IDs that are grouped together into different data nodes. Data tables within the analyte node contain identifying information both at the compound level and at the fragments level of the data. Identification information typically includes connecting different types of IDs with one another and relevant metadata such as formulation, machine-readable chemical structure information in the form of Simplified Molecular Input Line Entry System (SMILES) strings, and chemical names utilizing the International Union of Pure and Applied Chemistry (IUPAC) naming convention. The data node contains the majority of the mass spectral data, software used to generate data, data collection parameters, and other measures of quality control.

2.1.2 Data Extraction and Exploratory Data Analysis

Tables pertaining to peaks, samples, compounds, compound fragments, and other metadata are extracted from the database and joined using a series of unique identifiers linking each table as shown in Figure 1. Data are then filtered down to only include second stage mass spectrometry information (i.e., MS2 spectra as opposed to the first stage mass spectrometry information) since this type of information provides the level of specificity to identify PFAS chemicals, similar to a fingerprint.

sample_class_name	sample_description	compound_name	compound_formula	precursor_mz	measured_mz	measured_intensity
aqueous film-forming foam (AFFF)	3M 1998 AFFF	Perfluorooctanesulfonic acid	C8HF17O3S	498.9316	[59.9345, 63.9666, 63.9726, 63.9918, 69.0561, ...	[3.49499861983986, 7.05184967277006, 4.5135022...
analytical standard	PFAS150 - DTXSID7060332	(Perfluorobutyryl)-2-thenoylmethane	C10H5F7O2S	320.9822	[59.0125, 82.9948, 85.3011, 86.9409, 87.5749, ...	[5519.77734375, 85761.6953125, 4692.6743164062...
commercial formulation	3M FC-95 Formulation	Perfluorooctanesulfonic acid	C8HF17O3S	498.9311	[63.9673, 68.9989, 69.0148, 69.0276, 79.7099, ...	[1.91816624656411, 12.2460759498445, 12.481794...
analytical standard	PFAS150 - DTXSID90382620	3-Perfluoroheptylpropanoic acid	C10H5F15O2	440.9979	[57.033, 59.0124, 60.8866, 62.9874, 67.7245, 6...	[6098.1328125, 21247.34765625, 6214.5678710937...
analytical standard	Reference Standard for PFAS	Perfluorohexanesulfonic acid	C6HF13O3S	398.9361	[54.5155, 57.1525, 58.0998, 60.3781, 62.6216, ...	[4270.2607421875, 4434.08642578125, 5135.24023...

Figure 1. A subset of data extracted from multiple tables in the NIST PFAS database and combined into a single dataset, where each row represents information on a single MS2 spectrum. The MS2 spectral information that represents measured signatures of PFAS compounds is contained within the columns “measured_mz” and “measured_intensity”.

There are three classes of samples available in the NIST PFAS database as shown below in Figure 2: (i) analytical standards are single compound PFAS references, (ii) the AFFF class refers to Aqueous Film Forming Foam mixtures, mostly from the 3M company, and (iii) commercial formulations (CF). We note that AFFF and CF are complex mixtures sharing many underlying fluorinated substances.

On a more granular scale, the spectra in the database are derived from 18 discrete AFFF samples, and 3 CF samples. Figure 2 displays the frequency of MS2 spectra occurring in each specific formulation subtype, split across the two formulation categories. While there are fewer CF sub types (boxed in red in the legend below), these constitute more MS2 spectra overall,

leading to a balanced formulation dataset. Note, not all 18 AFFF subtypes are represented in the plot. This is due to the automated preprocessing pipeline which removes spectra which do not satisfy bare minimum requirements (i.e., enough samples per class for model training and sufficient number of peaks per spectra). Based on this exploration, we focused on the formulation classification (AFFF vs CF) problem in Year 1, refining models for formulation class attribution (see Section 2.0), though we discuss exploratory data analysis on compound overlap between manufacturers below.

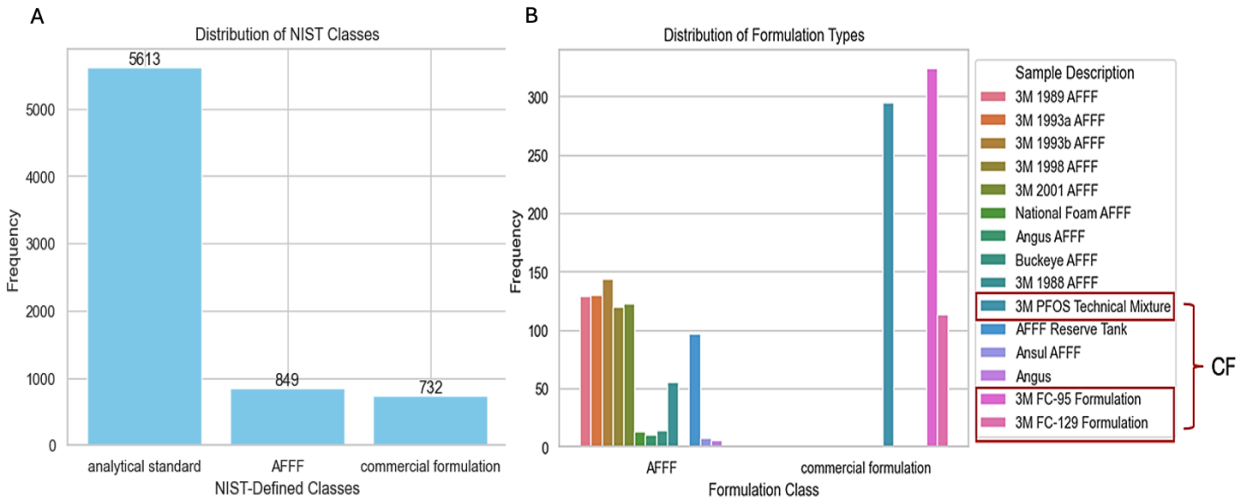

Figure 2. (A) Frequency of MS2 spectra in the NIST PFAS Database across each main class. (B) Frequency of MS2 Spectra across each AFFF and CF subtype.

Given that the formulations in the database all contain PFAS, we expect there to be a high level of overlap in the compounds represented in the dataset, which, in turn, raises the difficulty in accurate source attribution. To understand the level of similarity between compounds in the dataset, and where classifications might be particularly difficult, we computed a metric to capture the percentage of overlap in compounds between each pair of formulations (Figure 3). The overlap proportion is computed by finding the number of unique shared compounds across the set of a pair of formulations and normalizing by the size of the more infrequently occurring formulation type. This provides a metric that captures the proportion of compounds found in the smaller class of compounds that are also found in the larger class. This metric is therefore biased towards visualizing the potential difficulty in classifying formulations with less data, when they share chemical similarities with overrepresented formulation types. Figure 3 visualizes this compound overlap between each pair of the top 10 formulations in the database. Our assumption is that spectra belonging to classes with higher compound overlap will be more difficult to discriminate from one another with machine learning models. For instance, 3M 1993a and 1993b are expected to be difficult to disentangle, whereas 3M 1988 has very little overlap with other formulations, and will likely be classified separately more readily.

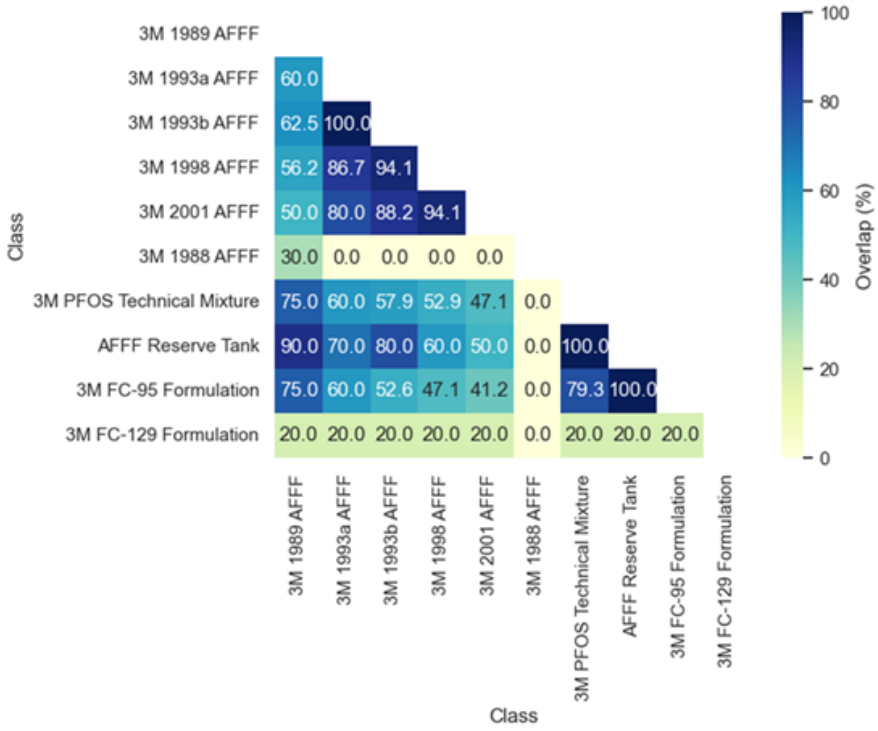

Figure 3. Percentage of compound overlap between each formulation type, where values are normalized by the minimum class size in each overlap pair.

2.1.3 Data Pre-Processing

After merging all relevant data tables by their appropriate unique identifiers, a series of preprocessing steps are applied to standardize spectral information and prepare the dataset for modeling. High-resolution mass spectrometry data such as that curated in the NIST PFAS database are inherently variable in length and high in dimensionality, necessitating transformation into fixed-length numeric encodings that can be ingested by machine learning algorithms. The software developed on this project as part of the AIML framework (AI4PFAS) in Python implements three such encoding strategies: intensity binning, Gaussian random projection, and feature hashing.

In mass spectrometry, intensity binning typically partitions the continuous mass-to-charge (m/z) axis into moderately sized intervals, summing measured intensities within each bin. This produces a compact, fixed-length representation of each spectrum, with bin width chosen to balance mass resolution and noise reduction. With the other two encoding strategies, Gaussian random projection and feature hashing, the process begins with fine-grained intensity binning, in which very small bin widths are applied to capture precise fragment m/z values. This produces sparse, high-dimensional vectors that serve as input for dimension reduction. In the Gaussian random projection approach, these high-dimensional vectors are multiplied by a randomly generated Gaussian matrix to obtain lower-dimensional numeric vector representations of the original data (Johnson and Lindenstrauss, 1984 and Li, et al., 2006). This method approximately preserves pairwise distances between spectra. In the feature hashing approach, each fine-grained bin index is mapped to a target bin using a deterministic hash function (e.g., MurmurHash3 (Appleby, 2015)). Intensities from bins mapping to the same target bin are aggregated, producing a fixed-size representation that retains much of the spectra's similarity structure (Moody, 1988).

Following application of one of these three encoding methods, each feature (i.e., column) in the resulting matrix is scaled via normalization or standardization. This ensures that all features contribute equally to the model, regardless of their original magnitude or units, while preserving within-feature variance.

2.1.4 Model Pipelines

There are four model pipelines currently implemented. Each is made up of a dimension reduction model followed by a classifier (Figure 4), meaning the workflow first transforms the high-dimensional input data into a lower-dimensional representation that preserves essential structure, and then applies a classification algorithm to assign labels based on that reduced representation. The dimension reduction techniques implemented are principal components analysis (PCA) and UMAP, while the classifiers include linear discriminant analysis (LDA), logistic regression (Log), and RF. We tested several combinations of dimension reduction + classifiers, resulting in the final set of 4 models: 1) PCA+Log; 2) UMAP+Log; 3) UMAP+LDA; 4) UMAP+RF.

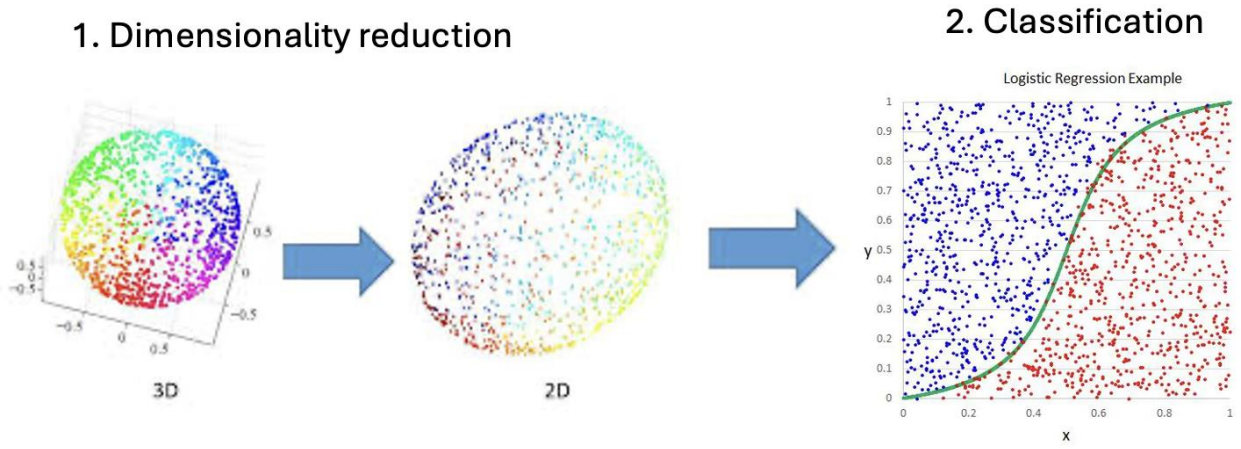

Figure 4. Schematic of model pipeline. A dimension reduction model is followed by a classification model.

PCA is a linear dimension reduction technique that transforms the data into a new, orthogonal coordinate system where each axis is a different linear combination of the original data. The first principal component captures the largest proportion of variability in the original data, with additional principal components capturing subsequently less. The number of components selected for classification is determined by the desired proportion of total variance preserved. In contrast, UMAP is a non-linear dimension reduction technique that constructs topographical representations of the data such that the loops, voids, and connectedness are preserved (McInnes, et al., 2018). From a high-dimension graphical representation of the data, a simplified, low-dimension version of the graph is created such that relationships between features of the data are preserved while minimizing the total number of variables in the data.

For classifiers, LDA finds a linear combination of features to separate between two or more classes is optimized to increase separability between classes while minimizing the variability within classes. Logistic regression uses linear combinations of data to model the log odds of an event. The response variable is typically binary, in this case whether the PFAS compound is considered AFFF or CF. An RF model calculates a collection of decision trees and votes on the

best subset or combination of data allow for classification. Under this algorithm, decision trees refer to a classification method where random subsets or features of the data are iteratively divided to find the optimal split of the data by using metrics such as the mean square error to quantify the amount of information retained after each data split.

2.1.5 Hyperparameter Optimization

During the initial phase of the study, we manually selected key settings—known as hyperparameters—that govern both data preprocessing and model training. These choices were informed by early experimental results. Based on these initial experiments, we observed that the model that combines UMAP and RF outperformed the others in terms of accuracy and recall. We therefore chose to focus on this model moving forward and employed a Bayesian hyperparameter optimization strategy that helped us select the most effective combination of hyperparameters for each classification problem. Table 1 includes two sets of optimized hyperparameters found for each classification problem.

2.2 Results & Discussion

Here, we examine model performance to predict mass spectra detected from PFAS samples as belonging to PFAS sources of interest at various levels of granularity. In this framework, models developed to predict these classes are a step towards providing a capability to rapidly determine PFAS contamination source from measured chemical signatures.

In the exploratory phase of the study, we evaluated a range of model pipelines combining different dimension reduction techniques and classifiers. Pipelines using UMAP consistently and significantly outperformed those using PCA. While PCA is a linear method that captures variance across features, it often fails to preserve complex, non-linear structures in the data. In contrast, UMAP is a non-linear technique that better maintains the local and global structure of the data in a lower-dimensional space. This capability made UMAP particularly well-suited to the spectra dataset, which likely contains non-linear relationships among features relevant to class separation.

Given these observations, we chose to focus exclusively on UMAP-based pipelines for all subsequent modeling. Further experimentation revealed that, among various classifiers, RF and LDA models achieved the highest performance for formulation class attribution, though RF demonstrated a minimal edge. As a result, we selected the UMAP + RF (henceforth, RandomForestUMAP) pipeline for more in-depth Bayesian hyperparameter optimization. Table 1 includes parameter combinations for three models examined in this effort that demonstrate the range of performance levels that could be achieved for the spectrum to formulation class problem; two using RandomForestUMAP and one LinearDiscriminantAnalysisUMAP. We discuss the performance of these three models in further detail in Section 2.2.1.

Table 1.Optimized Hyperparameter Values for Formulation Class Attribution

Hyperparameter	Description	RandomForestUMAP (with sample metadata)	RandomForestUMAP (spectral signature only)	LinearDiscriminant AnalysisUMAP
Add precursor mz feature	Whether to include the precursor *m/z* value as an additional feature in the encoded spectra representation	False	False	False
Add fluorination feature	Whether to include a categorical feature indicating the fluorination process used for AFFF spectra (electrochemical fluorination, or fluorotelomerization	True	False	False
Top Peak Count	Number of highest-intensity peaks selected from each spectrum before applying the encoding method	10	10	10
Encoding Type	Method used to encode MS2 spectra into numerical features. Options: Intensity binning, Gaussian Random Projections, Feature Hashing	Intensity binning	Intensity binning	Intensity binning
Bin Size	Width of each bin in m/z units, used only when using intensity binning	22	22	22
Encoding Dim	Dimensionality of encoded features – only used when not encoding type is not intensity binning	None	None	None
Scaler Type	Feature scaling method applied before modeling. Options: minmax scaling or standard scaling	minmax	minmax	minmax
Class balancing	Class balancing technique applied to the training data. Options: SMOTE, SMOTE Tomek, or down sampling	SMOTE	SMOTE	SMOTE
N components	Number of dimensions in the UMAP embedding.	2	2	2
min_dist	UMAP hyperparameter controlling how closely points are packed together in the embedding space in [0.1, 1] range (smaller	0.1	0.1	0.1

Hyperparameter	Description	RandomForestUMAP (with sample metadata)	RandomForestUMAP (spectral signature only)	LinearDiscriminant AnalysisUMAP
	values preserve more local structure)			
model	Type of model applied to the encoded spectra	RandomForestUMAP	RandomForestUMAP	LinearDiscriminant AnalysisUMAP
n_estimators	Number of decision trees in the Random Forest model	371	371	--
max_depth	Maximum depth of the decision trees.	None	None	--
max_features	Number of features considered when looking for the best split in each tree	log2	log2	--
min_samples_leaf	Minimum number of samples required to be at a leaf node	2	2	--
min_samples_split	Minimum number of samples required to split an internal node	5	5	--

One of the hyperparameters that we investigated during the model optimization stage was class balancing (see Table 1 for optimized setting). Class balancing addresses one of the key challenges with dataset readiness for AIML and is one of the methods that can help mitigate model bias in class prediction—by ensuring that the dataset has equal or close to equal representation among the class labels on which the model is trained to predict. Class imbalance within the dataset (i.e., when a class is more highly represented compared to the others or a class is only represented by few examples) can create bias within the model to favor prediction to the class that is most represented, given the large number of examples the model has for learning. Additionally, for the class that is least represented, the model may learn the underrepresented class extremely well (i.e., as if the model has memorized the examples), but that can result in the model's inability to generalize outside the dataset used for its training. Additionally, application of appropriate class balancing strategies can be challenging when data for AIML model development are scarce. We applied and examined relevant class balancing methods here, including Synthetic Minority Over-sampling TEchnique (SMOTE) (Chawla et al., 2002; Blagus and Lusa, 2013), SMOTE with Tomek links (SMOTE-Tomek) (Batista et al., 2004; Zeng et al., 2016), and down-sampling (Lemaître et al., 2017). These approaches represent different ways to address class imbalance in a dataset.

SMOTE provides a way to oversample the class that is least represented by generating synthetic data via interpolation to achieve class balance (Chawla et al., 2002; Blagus and Lusa, 2013). By utilizing interpolation and creating new examples rather than oversampling the existing examples in the smaller, underrepresented class, SMOTE overcomes the minority class overfitting issue (i.e., the model learns the few examples in the minority class extremely well and cannot generalize) that is common to the more conventional random oversampling method. Alternatively, SMOTE can be combined with Tomek links, a data cleaning method that reduces noise and creates better distinction between two classes by removing examples (down-sampling) from both classes when these examples lie at the boundary between the two classes (Batista et al., 2004; Zeng et al., 2016). The SMOTE-Tomek approach thus mitigates potential overfitting via oversampling while enhancing the distinction between two classes for model prediction. And finally, class balance with random down-sampling using the Imbalanced Learn

Python package selects examples from the most-represented class without replacement to include in the model's training (Lemaître et al., 2017) such that the number of examples among the classes are equal. However, the random sampling is performed without accounting for any underlying structure within the dataset, such as similarity among examples, and random elimination of samples can potentially remove useful information for prediction to the majority class distinct from the smaller classes.

In the AFFF versus CF classification problem, we have 526 spectra measured from samples that belong to the AFFF class and 550 spectra from samples belonging to the CF class, that meet the top peak count parameter criteria. While the numbers of spectra between the two classes do not differ drastically, class balance should still be implemented to mitigate any potential bias that could arise to prefer prediction to the majority CF class. With the RandomForestUMAP configuration in general, the SMOTE over-sampling approach yielded some of the more performant models.

2.2.1 Formulation Class Attribution

We compare the performance of three AIML models that represent the range of prediction accuracy for the formulation class problem in Sections 2.2.1.1 – 2.2.1.3 below. All utilize UMAP as the dimension reduction technique and demonstrate the effects of using a different ML algorithm (e.g., LDA, RF) and selection of different features (spectral signatures vs. spectra and sample metadata). Section 2.2.1.4 discusses the benefits and disadvantages of each and comments on why we may choose one or the other, depending on the intended use case. Figure 5 below provides an overview of each model's performance (for a select set of hyperparameters) that will be discussed in greater detail in Sections 2.2.1.1 – 2.2.1.3.

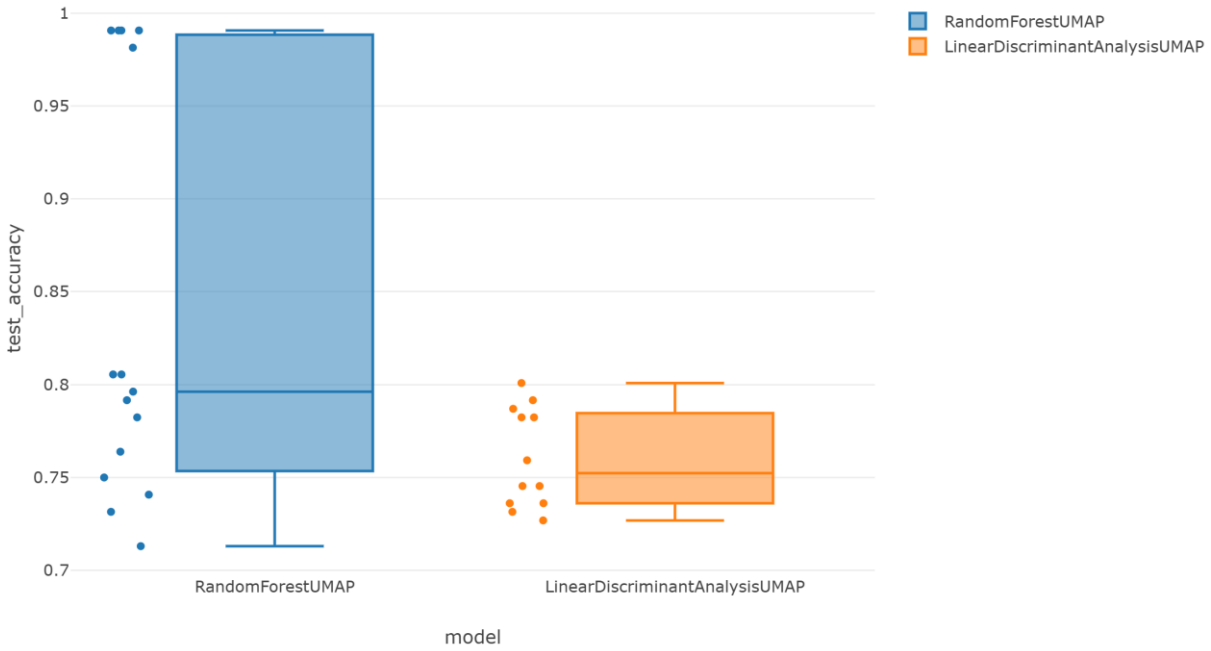

Figure 5. Boxplot comparing performance of models with different architectures and hyperparameters. Note the two distributions with the RandomForestUMAP model, where the higher test set accuracy values correspond to inclusion of sample metadata as a feature in the model and the lower test set accuracy range utilizes spectral information only.

Though not the most performant model, we selected the RandomForestUMAP architecture that uses spectral signatures only (80.6% accuracy representing better than modest performance) on which to continue development and improve upon, and on which to investigate generalizability to other datasets. We discuss our rationale for selecting this model in Section 2.2.1.4.

2.2.1.1 Linear Discriminant Analysis as Classifier

One of the classifiers that we investigated in Year 1 was a simple LDA. Figure 6 displays the performance of a representative model using the LinearDiscriminantAnalysisUMAP configuration as a confusion matrix, where we achieve a test set accuracy of 80.1%, one of the more performant models utilizing LDA as the classifier. A confusion matrix is a table that compares the model's predictions (columns) against the true labels (rows). Correct predictions appear along the diagonal from the top left to the bottom right—these are cases where the predicted and actual categories match. Off-diagonal entries represent misclassifications, where the model assigned a spectrum to the wrong category. The closer all values are to the diagonal and the fewer entries there are elsewhere, the better the model's performance.

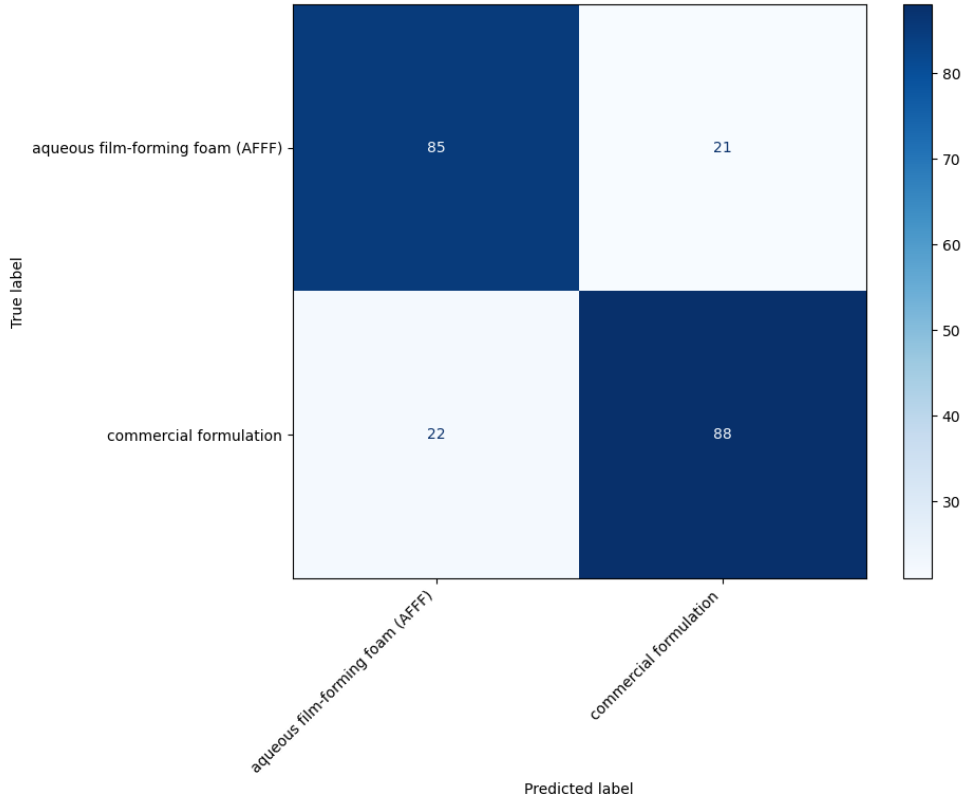

Figure 6. Confusion matrix for one of the most performant LinearDiscriminantAnalysisUMAP models using spectral signatures only as features for model training.

Here, we observe that there are misclassifications for both classes. While 85 of the 216 mass spectra from the test set were correctly predicted as detected in samples from the AFFF class, 21 were misclassified as CF, and 22 spectra from CF samples were misclassified as detected in AFFF samples. This indicates that while the model has learned some mass spectral features that can distinguish between the AFFF and CF classes, the model is still making mistakes and the set of mass spectral features used to train the model does not enable complete distinction

between the two PFAS classes of interest. We examine approaches to allow us to gain more insight into misclassifications in the next two sections that discuss predictions with the RandomForestUMAP model.

2.2.1.2 RandomForestUMAP: Using Spectral Signatures Only

When we move from the LinearDiscriminantAnalysisUMAP model to RandomForestUMAP, we observe a slight increase in prediction accuracy for the test set (average of 76.8% accuracy compared to 76.4%). The confusion matrix in Figure 7 below indicates that misclassifications still occur in both classes with a different model architecture, though where misclassifications lie differs.

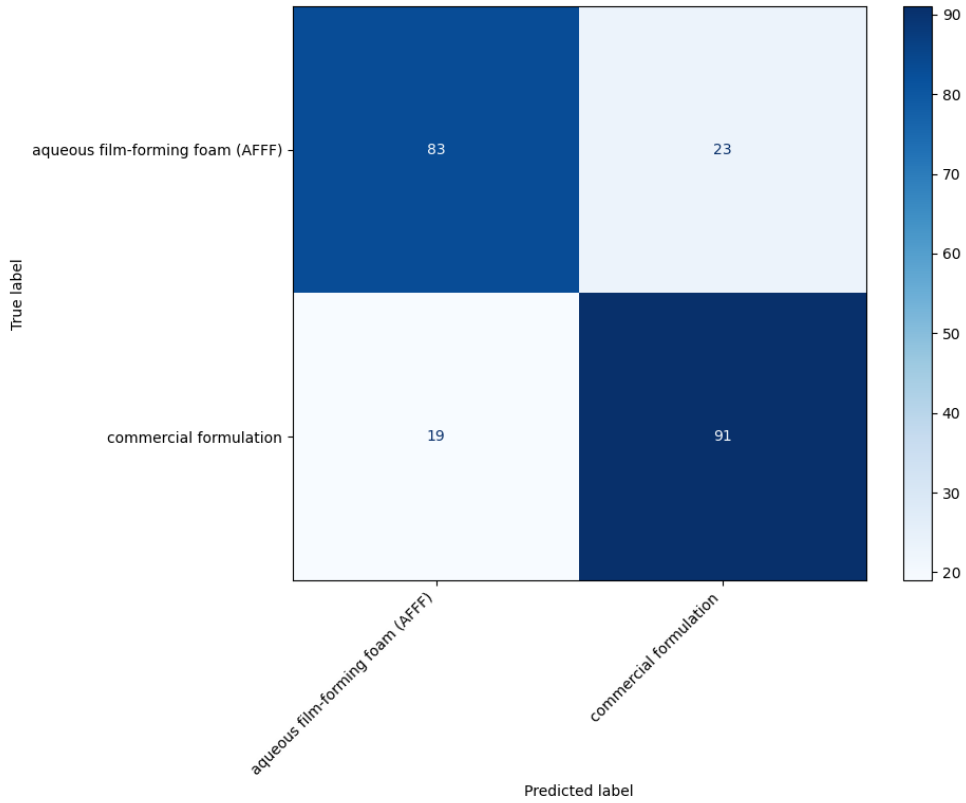

Figure 7. Confusion matrix for the RandomForestUMAP model using spectral signatures only as features for model training.

Interestingly, when we compare the confusion matrices between the RandomForestUMAP model to the LinearDiscriminantAnalysis model in Section 2.2.1.1, we see that fewer misclassifications occur for the CF class (91 as opposed to 88 correct predictions; Figure 7 and Figure 6, respectively). We can further identify and investigate the spectra that were misclassified to gain a better understanding of the model's decisions in predicting the two PFAS classes.

Because each model includes a dimension reduction step before classification, we can visualize how the originally high-dimensional spectral data are transformed and arranged in a lower-dimensional space prior to applying the classifier. This visualization reveals clusters of similar spectra and can provide insight into why a model might misclassify certain samples. Figure 8 below shows the reduced components from the same RandomForestUMAP model featured in

Figure 7, visualized as a scatter plot. Each data point, representing an MS2 spectrum in the test set, is labeled with its true PFAS class label in the figure. Further, we display the model's decision boundary in predicting to AFFF or CF with the light red/blue contours along with the black line. In this case, all spectra appearing below the decision boundary on UMAP2 space and to the bottom right on UMAP1 space classified as CF, while all above and top left were classified as AFFF.

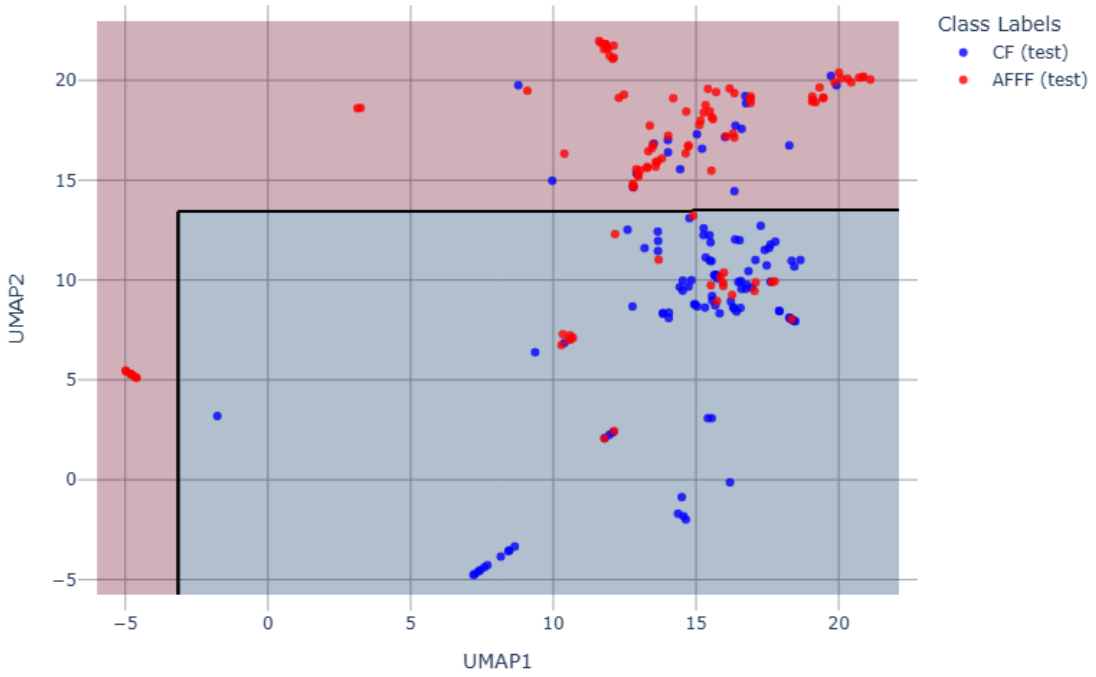

Figure 8. UMAP embedding visualizing the transformation of each MS2 spectrum in the test set in two dimensional UMAP space before classification. The light red/blue contours along with the black line represent the RandomForestUMAP's decision boundary for class prediction to the AFFF and CF classes.

Most notable when looking at Figure 8 with respect to misclassifications is that the misclassified points often lie near, if not on top of, correctly predicted points, meaning the transformed spectral features of misclassified spectra appear extremely similar to those of the correctly classified spectra. While the source of these misclassifications can be difficult to determine with complete certainty, we can consider what UMAP does and the nature of mass spectra. UMAP groups samples based on the similarity of their spectral features, patterns of peaks and intensities that correspond to different chemical components. If two spectra have patterns which more closely resemble those of another class, UMAP will place them nearer to that class in the reduced space.

For example, in the center cluster in Figure 8 (coordinates UMAP1 10.3-10.7, UMAP2 6.8-7.3), we see 6 data points—corresponding to spectra misclassified as CF—that lie on top of a single correctly classified spectrum belonging to the CF class. In examining the compounds that are associated with these spectra, we observe that all (including the lone correctly classified spectrum) represent the PFAS compound PFOS, detected in the following AFFF samples: 3M 1993a, 3M 1993b, 3M 2001, and 3M 1998; whereas the correctly predicted spectrum was detected in the 3M FC-129 Formulation sample. PFOS is a well-characterized PFAS from the perfluoroalkyl sulfonate chemical family.

Figure 9 displays most of the spectra in question, with the top left spectrum correctly predicted as originating from a CF sample. Given that these mass spectra are measurements of the same compound PFOS, albeit in different samples, the high degree of similarity in fragment peak patterns is not surprising. Of note among the mass spectra associated with the AFFF samples is the range in fragment intensities, particularly in the lower m/z region between m/z 100 and m/z 250, relative to the base peak at m/z 498.932, where we observe slightly higher intensities in this region for the 3M 2001 AFFF sample but lower abundance of these fragments are typical among the other AFFF samples. The spectrum belonging to the CF sample contains few peaks compared to the other AFFF spectra for PFOS. However, the fragment peaks that are present in the CF spectrum match those of the highly abundant fragment peaks in the spectra from AFFF samples. From comparing fragmentation pattern similarity and visualizing the clustering in UMAP embedding space, we can infer that one or more of the peaks at m/z 79.96, m/z 98.96, and m/z 498.93 may be important for the model to predict these spectra as belonging to AFFF samples.

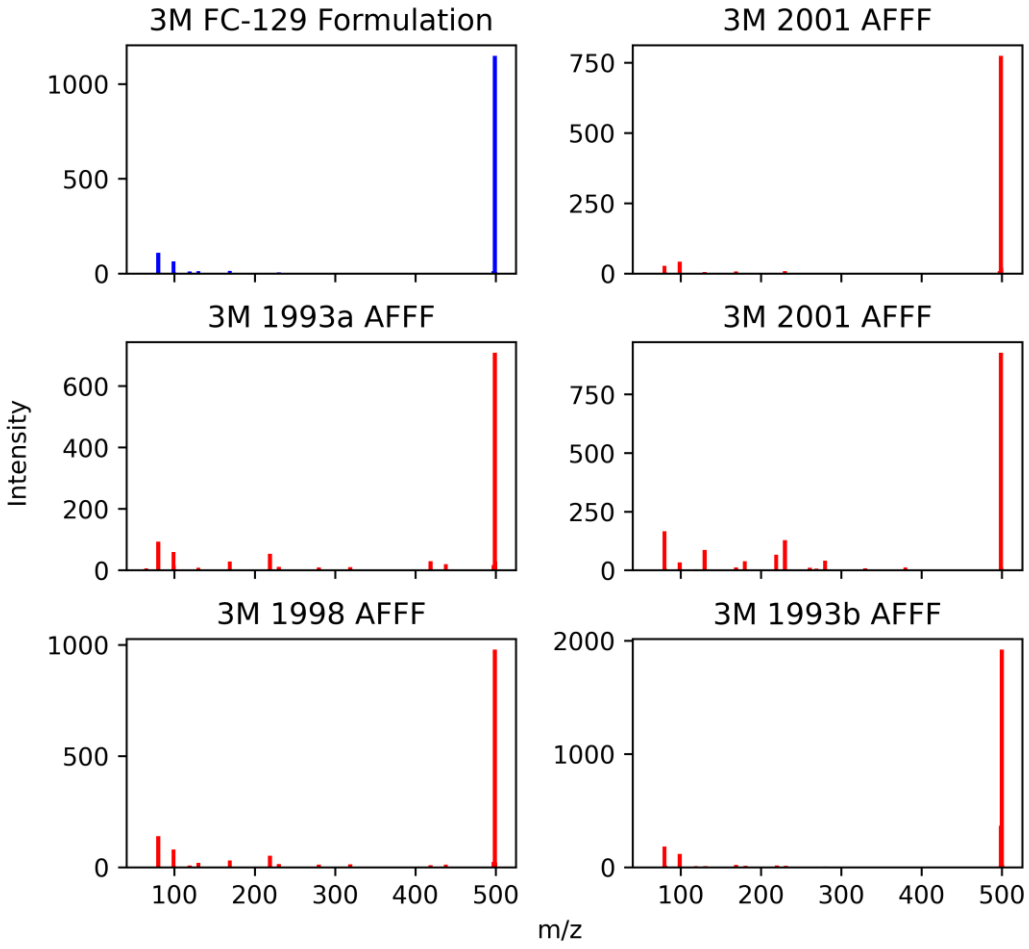

Figure 9. Mass spectra corresponding to a cluster of data points that were predicted to belong to CF samples, when most of these spectra were detected in AFFF samples. All spectra are measurements of PFOS. Note the spectra in red were misclassified as belonging to CF samples.

When examining misclassifications to the AFFF class, we similarly observe spectra of another perfluoroalkyl sulfonate—perfluorohexane sulfonic acid (PFHxS)—detected from both AFFF and CF samples. Here, this cluster of data points lies just above the decision boundary (Figure 8; coordinates UMAP1 12.8-13.3, UMAP2 14.7-15.7). As a perfluoroalkyl sulfonate, PFHxS shares core substructural components to PFOS, but with two fewer carbons to create the perfluoroalkyl backbone.

Two CF spectra belong to the 3M PFOS Technical Mixture, while the correctly predicted spectra for the AFFF class belong to 3M 1989, 3M 1993a, 3M 1993b, 3M 1998, and AFFF Reserve Tank samples. The mass spectra for this cluster of data points are shown in Figure 10. Note the few fragment ions in spectra belonging to CF samples (m/z 79.96, 98.96, and 399.13) are also present in AFFF samples and represent the most abundant peaks in these spectra. We also note that two of the more abundant fragment ions at m/z 79.96 and m/z 98.96 are shared between PFHxS and PFOS, and spectra for these two compounds are detected in both AFFF and CF samples, which complicates the distinction between the two PFAS classes when solely considering spectra representing singular compounds.

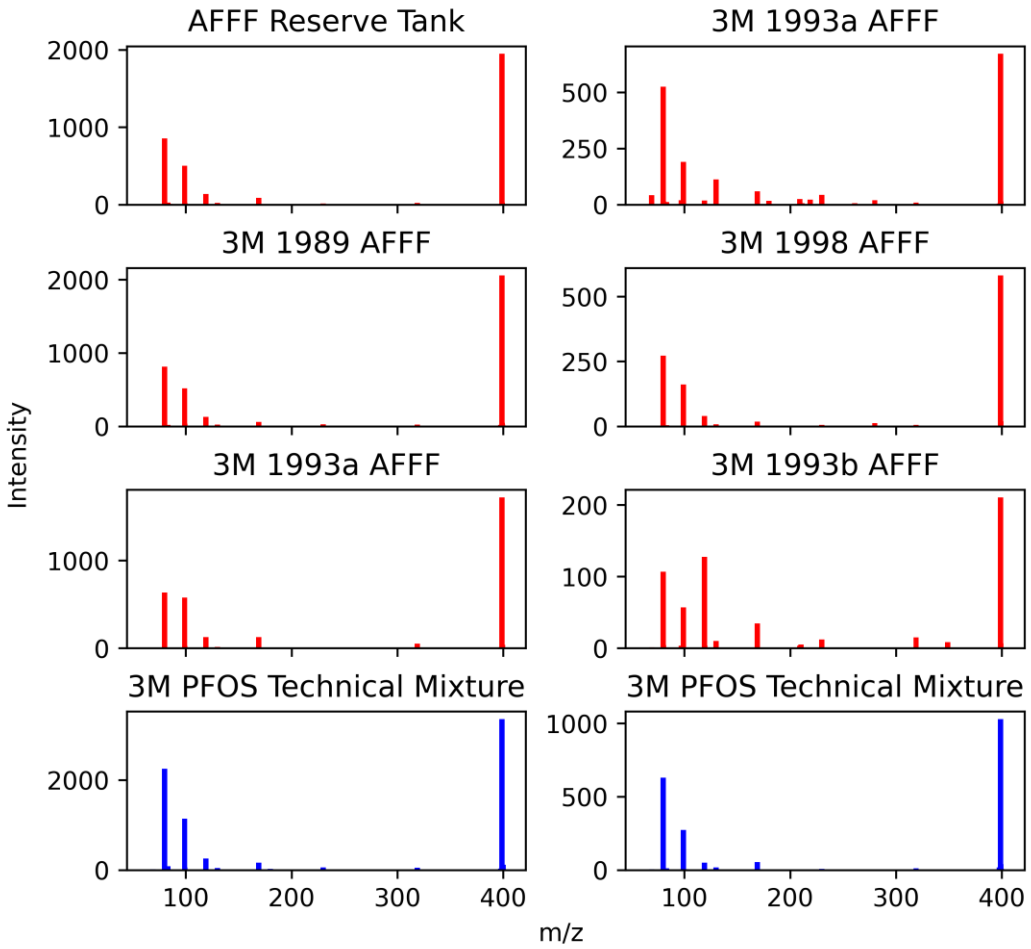

Figure 10. Mass spectra corresponding to a cluster of data points that were predicted to belong to AFFF samples, but two spectra were detected in CF samples (spectra in blue). All spectra are measurements of PFHxS. Note the spectra in blue were misclassified as belonging to AFFF samples.

In probing the misclassifications to the two PFAS classes, we observe that the UMAP clusters spectra with very similar fragmentation patterns (very likely belonging to the same compound) to the same embedding space, giving us confidence that the UMAP is utilizing and honing in on certain spectral features. This suggests that the misclassification may not necessarily be a classification model failure but rather a reflection of inherent signature similarities between the chemicals detected in these particular AFFF formulations and those detected in the CFs in this dataset. PFAS class prediction is complicated by different mixtures of PFAS compounds such that the same compound may be found in different samples, whether they be AFFF or CF. It is the unique formulations of these mixtures (chemical identities and abundance), that can be indicative of the various PFAS sample classes as opposed to individual chemicals alone. As such, we need to account for the chemical complexity and thus, signature complexity, that mixtures bring to the PFAS class prediction problem.

Despite the misclassifications, the ability to predict PFAS class with an accuracy of 80% using spectral signatures alone demonstrates the power of and art-of-the-possible using AIML tools, and applicability to the PFAS source attribution mission space.

Overall, these observations demonstrate some of the challenges that come with characterizing and differentiating among PFAS source types. Misclassifications, which we observed above owing to shared chemical components, as well as within class clustering, provide an opportunity to identify model enhancements—such as the need to account for additional chemical composition complexity—and discover emerging contaminant signatures.

2.2.1.3 Most Performant Model: Inclusion of Sample Metadata

Finally, we examine model performance when we include a feature derived from sample metadata in training the model. While we demonstrate 80% accuracy as the most performant, representing better than modest performance, with the RandomForestUMAP trained solely on spectral information, we investigated whether model performance could be further improved with additional features, such as incorporating different features derived from sample metadata. Here, we describe a configuration where the model learns the same spectral features as the model described in Section 2.2.1.2, but additionally includes a feature describing the fluorination process of the PFAS sample. Feature selection is a key part of model optimization and can substantially affect model performance.

On average, the RandomForestUMAP model that includes fluorination process as a feature boasts excellent performance, achieving near 100% accuracy. Figure 11 displays the confusion matrix for one of the most performant models in such a configuration. In this run, only two spectra were misclassified out of a total of 880 MS2 spectra. Nearly every spectrum lies on the diagonal, indicating that the model is making highly accurate predictions, with only two spectra detected from AFFF samples misclassified as originating from commercial formulations.

This high level of accuracy indicates that the inclusion of the fluorination process metadata value as a feature in the model substantially boosts model performance, jumping from 80% to 99% accuracy on the test set data.

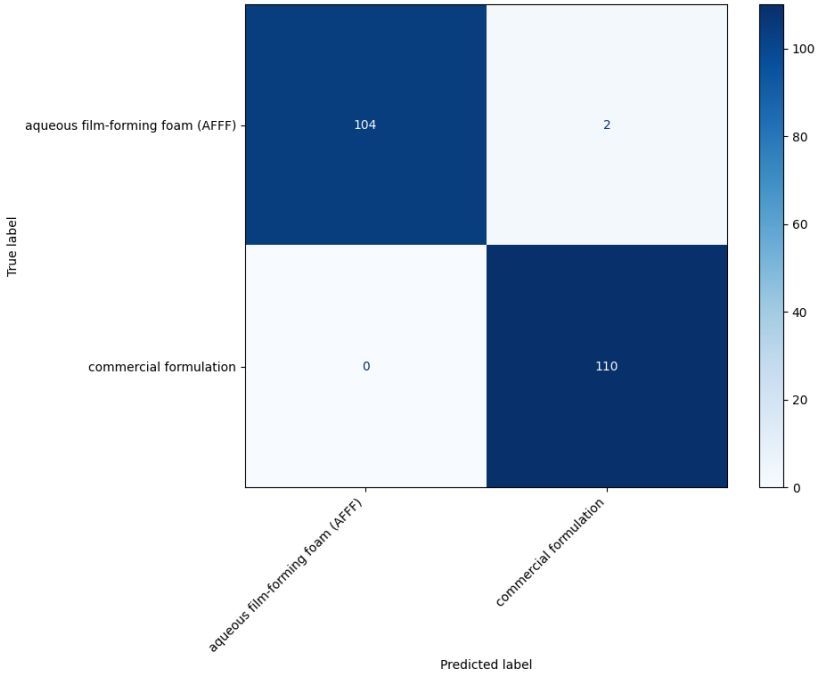

Figure 11. Confusion matrix for one of the highest performing formulation classification models. Here, only two out of a total 880 spectra were misclassified.

Figure 12 shows the reduced components from the same RandomForestUMAP model featured in the Figure 11, visualized as a scatter plot. Each data point, representing an MS2 spectrum in the test set, is labeled with its true PFAS class label in the figure. In this case, all spectra appearing on the right classified as CF, while all to the left were classified as AFFF.

In the UMAP projection in Figure 12, we note some within-class clustering of the spectra of compounds detected in CF samples into three major clusters. In particular, the larger cluster of spectra from CF samples belongs to perfluoroalkyl sulfonates (e.g., perfluorodecane sulfonic acid, perfluoroheptane sulfonic acid) detected in FC-95 and PFOS technical mixture, which are consistent with chemicals listed in products that include these specific formulations. For example, potassium perfluorooctane sulfonate (a chemical that belongs to the perfluoroalkyl sulfonate family) is listed as the main ingredient in FC-95 (brand name FLUORAD Brand Fluorochemical Surfactant) and is used in products like Scotchgard (3M Canada Company 2015; Renner 2006).

Two misclassifications within the large CF cluster can also be observed (Figure 12), where two red data points from spectra detected in AFFF samples are centered within the blue CF cluster. Investigation into the chemical identities of these misclassified AFFF-detected spectra provided critical insights into the model's behavior and the underlying chemical realities. These two spectra (belonging to the same compound, 6:2 fluorotelomer thia propanoamido dimethyl ethyl sulfonate, from the Angus AFFF sample) exhibited high similarity in UMAP-encoded space to those belonging to compounds detected in CF samples—specifically to the perfluoroalkyl sulfonates. The compound in question shares many substructural features with compounds from its spectral neighbors detected in CF samples within the UMAP space.

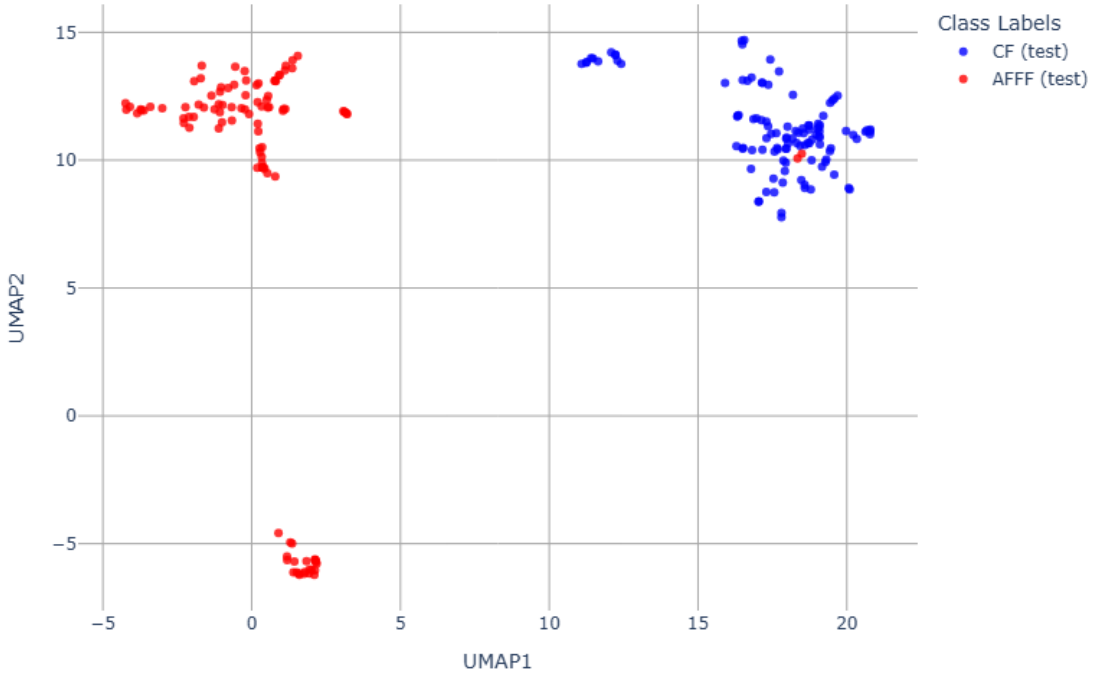

Figure 12. Visualization of reduced dimensionality components of a high performing formulation classification model, with true PFAS class labels indicated (red = AFFF; blue = CF). It is visually apparent why the classifier misclassified two AFFFs as CFs, given that the UMAP dimensionality reduction placed these spectra closer to other CF spectra.

Interestingly, we also observed that the vast majority of spectra in the lower left cluster—correctly classified as detected in AFFF samples (Figure 12)—also belong to perfluoroalkyl sulfonates (e.g., PFOS, perfluoroheptane sulfonic acid (PFHpS)), similar to the compounds in the large CF cluster that are correctly predicted as detected in CF samples. While these spectra belong to compounds that are highly structurally similar, the dimension reduction step via UMAP has focused on key features (i.e., spectral fragments and/or the metadata feature) that distinguish the spectra detected in AFFF samples from those detected in CF samples. Further investigation in identifying these spectral fragments and relationship with the metadata feature will yield insight into the important features used by the model to differentiate between AFFF and CF classifications and aid in development of a more complex model that includes signatures from mixtures of chemicals.

When we compare this RandomForestUMAP configuration to the one using spectral signatures only, it behooves us to understand the influence of the fluorination process on the model's decisions, which manifests in the substantial increase in prediction accuracy. As such, we examine the relationship between this feature and the PFAS classes. Table 2 below tabulates the number of MS2 spectra detected in samples that belong to either the AFFF or CF classes (total spectra and test set spectra) and further delineated by fluorination process. Note that the two fluorination processes, electrochemical fluorination and fluorotelomerization, are only associated with AFFF samples, and remain unknown for all the CF samples, and for a handful of spectra belonging to AFFF samples. This means that all the spectra detected in CF samples are associated with a single, unique encoding of fluorination process (in this case, NA is as

important as an actual fluorination process), which is different from the two fluorination processes that are only associated with AFFF samples. This clear distinction between the AFFF and CF classes is very likely being exploited by the model, resulting in near-perfect prediction accuracy.

Table 2. Count of MS2 spectra belonging to samples by PFAS class and delineated by fluorination process

PFAS class	Fluorination process	Total spectra	Test set spectra
AFFF	Electrochemical fluorination	415	80
AFFF	Fluorotelomerization	105	24
AFFF	--	6	2
CF	--	550	110

2.2.1.4 Model Assessment

While it is obvious that we can achieve superb prediction accuracy (99%) when sample metadata such as fluorination process is included as a feature the model trains on, we prefer the model described in Section 2.2.1.2 that demonstrates better than modest performance at 80% accuracy. This is because the model that only trains on spectral signatures is likely to be more generalizable in deployment to predict other unknown samples that may or may not contain PFAS.

In the spectral-signature only model, LC-MS/MS spectral signatures, when processed and analyzed by AIML algorithms, contain distinctive information to begin to differentiate between the PFAS source types of interest when we consider the spectrum to PFAS class problem. This type of capability is vital for environmental management as it provides a data-driven method for source attribution, which is often a complex and resource-intensive task, especially in the face of unknown chemical compositions in these samples.

We recognize the value of including sample metadata and see the clear benefits of augmenting spectral features with metadata features in AIML models, but caution against including features that may not be able to be derived from unknown samples to be predicted in model deployment. In our test case with including fluorination process, many elements of how each PFAS sample was produced and the chemical compositions in each sample were known and accessible in the NIST PFAS database. However, much of this information will not be available for environmental samples, which is the intended use case for our fully developed model. Alternatively, metadata that could be derived from environmental samples, such as sampling location, topologies, etc., could be useful for model predictions and should be investigated for their utility during model refinement.

This proof-of-concept effort demonstrates the promise of AIML in exploiting chemical signatures—in particular mass spectral signatures—for distinguishing among PFAS sources when only signature information is available. To offer a more generalizable model that can distinguish among other PFAS sources and formulations in the face of novel contaminant signatures that continue to emerge, further model enhancements are needed to consider the signatures of chemical combinations that contribute to the various formulations. While our developed model demonstrated that individual chemical signatures can be associated with certain PFAS source types, we also recognize that different PFAS source types may share some of the same chemicals (and contain some of the same signatures), but combinations of

individual signatures representing their full chemical compositions can provide enhanced differentiation. As such, to capture the diversity of chemical compositions that could be found and continue to emerge among PFAS sources, and enable their differentiation, we are currently developing a more robust model that will account for the co-occurrence of combinations of these chemicals and their larger families and will investigate the generalizability of such a model to different datasets.

2.2.2 Model Tracking and Comparison with MLFlow

Model development for source attribution requires systematic exploration of many hyperparameters and modeling approaches, often across distinct classification problems. Without a structured tracking system, comparing results and drawing conclusions about model performance quickly becomes cumbersome. To address this, we integrated the project's modeling framework with MLFlow, an open-source platform designed for managing the end-to-end machine learning lifecycle. During the R&D phase, MLFlow has enabled the team to record every model training run—including data inputs, hyperparameter settings, performance metrics, and artifacts—within a unified interface. This functionality allows efficient comparison across experiments, facilitates reproducibility, and helps identify the most promising model configurations for further development.

MLFlow will provide a foundation for long-term model management. By maintaining a transparent record of model provenance and decision history, the client will be able to track how the best-performing models were derived, revisit earlier approaches if needed, and deploy final models with confidence. In this way, the integration of MLFlow not only accelerates research progress but also ensures that the modeling products delivered are transparent, reproducible, and maintainable beyond the initial project phase. Refer to Figure 13, Figure 14, and Figure 15 for images of experiment and results tracking using the MLFlow dashboard.

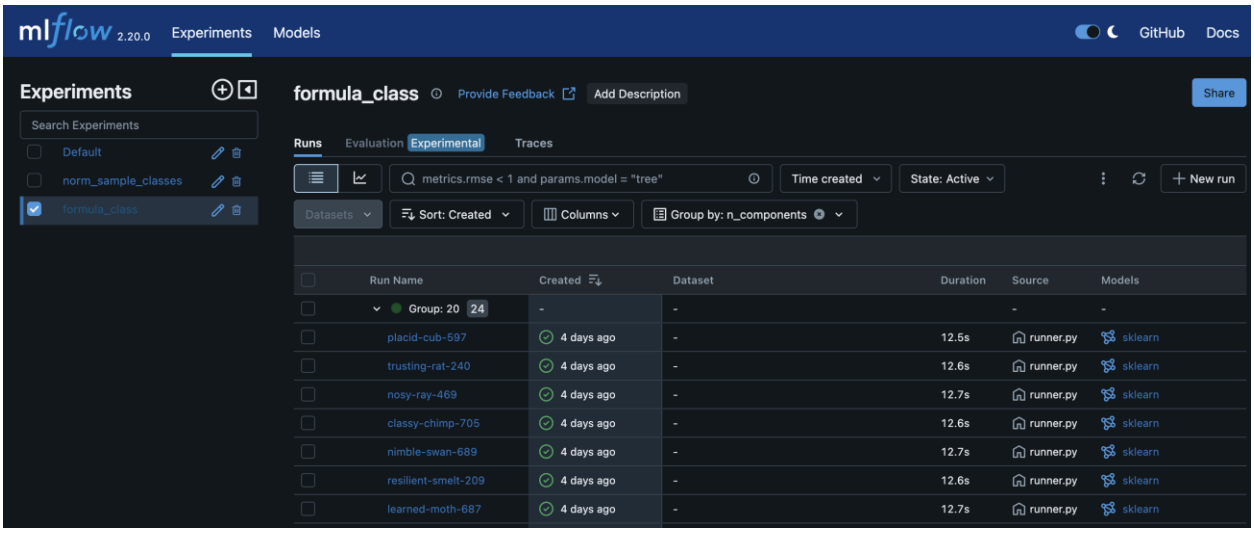

Figure 13. MLFlow Experiments Landing Page, used to track and organize all models run on a specific classification problem type.

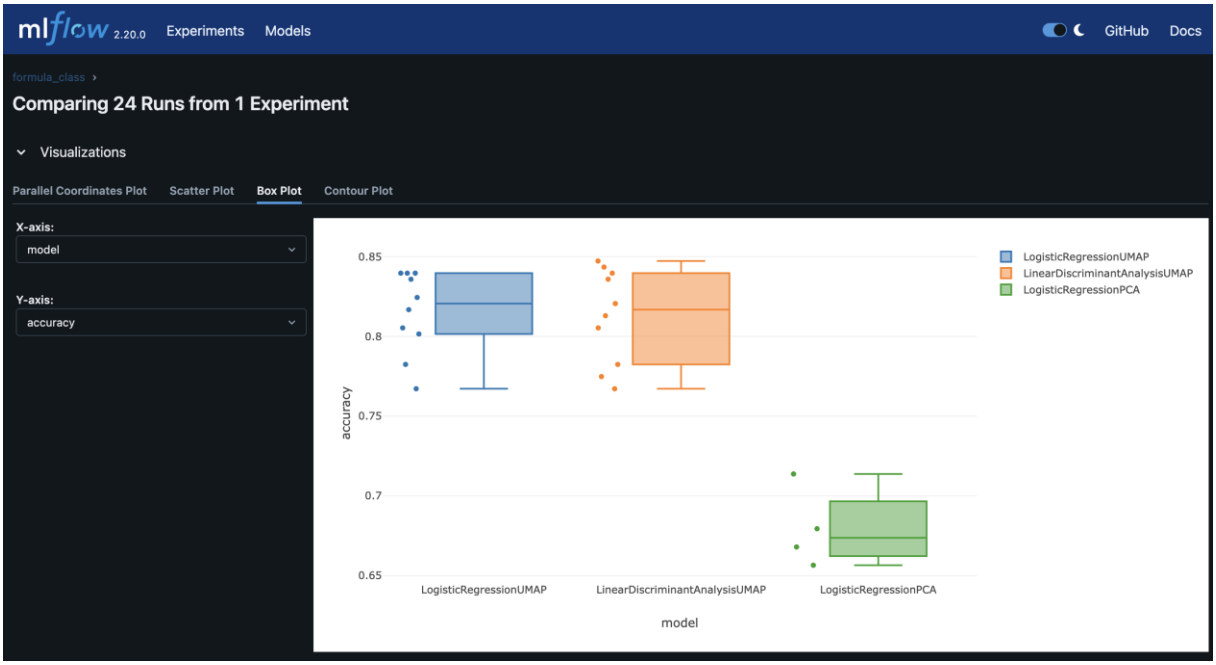

Figure 14. Box plots can be visualized across hyperparameters. The dashboard can also display scatterplots and can be used to explore the relationship between any hyperparameter selection and performance metric.

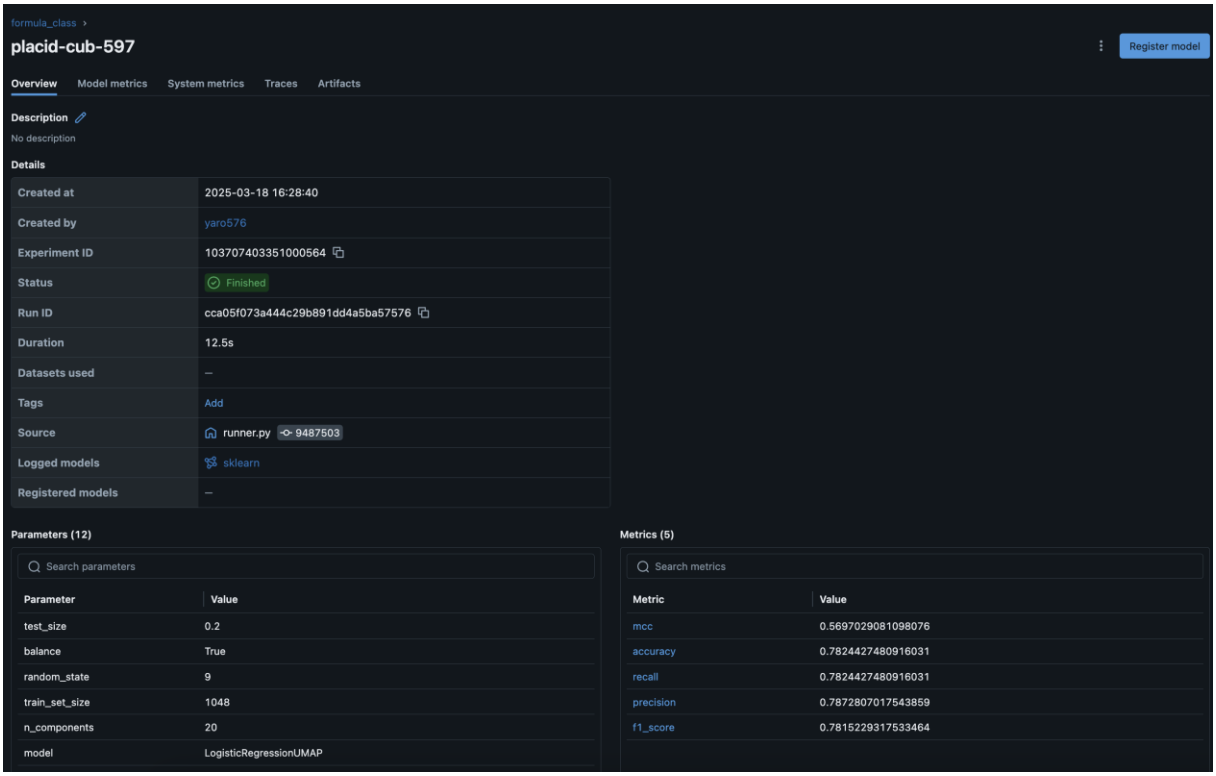

Figure 15. Hyperparameters and results are logged for each model. This includes metadata that can be used to easily locate where the model is stored on your system, to facilitate follow-on research with the best performing models.

2.3 Conclusions for PFAS Source Attribution

This study successfully demonstrates the feasibility of AIML techniques to link chemical signatures to PFAS sources and provide attribution indications without the need for *a priori* chemical identification. We investigated a variety of different model architectures and configurations that range in performance to develop and optimize a model for PFAS class prediction using mass spectral information. By applying a pipeline involving UMAP for dimension reduction and an RF classifier using solely LC-MS/MS data from the NIST PFAS database, we achieved 80% accuracy in distinguishing signatures detected in AFFF samples from those detected in other commercial PFAS-containing formulations.

The good classification performance, coupled with the ability to analyze misclassified samples that revealed underlying chemical similarities to CFs, is indicative of the model's ability to discern distinct spectral patterns that can begin to be linked to source origin, as well as the model's inherent interpretability. The integration of MLFlow for experiment tracking established a robust framework for systematic model development and comparison, ensuring accessibility and reproducibility in AIML model development. Overall, this capability is critical to facilitate responsible environmental management.

This work provides a robust proof-of-concept for AIML as a tool in the ongoing effort to manage PFAS contamination. Future work will involve applying and further refining these AIML algorithms to account for challenges with and generalizability to chemical mixtures, as well as complex environmental samples, moving from the controlled environment of clean spectra to the challenging realities of contaminated sites. Ultimately, we aim to develop AIML-driven approaches that enhance our ability to discern emerging contaminant signatures and rapidly and accurately identify PFAS sources. Additional work is being done to determine model generalizability, including modeling with environmental samples, which will lead to more informed and effective environmental protection decisions.

3.0 Large-Scale Computational Screening and Machine Learning Prediction of PFAS Degradation Potentials

Conventional remediation relies predominantly on sorptive technologies (granular activated carbon, ion-exchange resins) that concentrate rather than destroy contaminants. Destructive approaches – high-temperature incineration (>1000 °C), plasma-based methods, photochemical processes, sonolysis, and electrochemical oxidation – offer pathways to mineralization but remain energy-intensive, costly, or sensitive to water-matrix effects (Cleston and Charles, 2024; Li et al. 2025; Sidnell et al., 2022; Kim et al., 2024). Importantly, more than 12,000 PFAS structures are now registered (USEPA, 2023) displaying enormous structural diversity (chain length, functional groups, branching, and charge state). This heterogeneity translates into widely varying thermodynamic stability and degradation behavior, making universal treatment protocols elusive.

The feasibility of many destructive methods (plasma, photochemical, sonolysis, UV, electrochemical oxidation) hinges on the oxidation potential of each PFAS congener; compounds with lower potentials are thermodynamically easier to degrade. However, experimental measurement of oxidation potentials for thousands of PFAS is impractical, and high-level quantum chemical calculations are computationally prohibitive at this scale.

In this study, we combine semiempirical quantum chemistry, density functional theory (DFT), and machine learning to enable rapid, accurate prediction of PFAS oxidation potentials across chemical space. Starting from the EPA PFAS Master List, we curate a dataset of 8214 unique PFAS molecules (PFAS-8k), compute their adiabatic oxidation potentials in aqueous solution using a computationally efficient two-tier protocol, analyze structure–property trends, and train a high-performing Random Forest model capable of instant predictions for new structures. This workflow provides the first broad thermodynamic map of PFAS oxidative degradability and identifies structural motifs that render certain congeners significantly more vulnerable to electrochemical treatment.

3.1 Methods

Our approach integrates three hierarchical levels to balance accuracy and throughput: (1) Data preprocessing through filtering to generate 8214 unique PFAS molecules; (2) Large-scale screening of 8,214 PFAS using the semiempirical GFN2-xTB method to estimate oxidation potential; (3) High-accuracy benchmark calculations on a representative 600-molecule subset using full DFT geometry optimization and frequency calculations; and (4) a Machine-learning surrogate model trained on Morgan fingerprints to deliver oxidation-potential predictions for any PFAS structures.

3.1.1 Data Curation and Chemical Space

PFAS structures and SMILES strings were retrieved from the EPA CompTox Chemicals Dashboard PFAS Master List (USEPA, 2023). Inorganic species, fragments, and molecules containing heavy metals were removed using RDKit filters. After deduplication and canonicalization, 8,214 unique organic PFAS remained (denoted PFAS-8k). The dataset spans perfluoroalkyl carboxylic acids (PFCAs), perfluoroalkyl sulfonic acids (PFSAs), fluorotelomers, sulfonamides, ether-PFAS (e.g., GenX), and numerous emerging structures with aromatic, hydroxyl, or charged functional groups.

3.1.2 Quantum Chemical Calculations

All quantum chemical calculations were performed using xTB and Gaussian 16. For the full PFAS-8k dataset, molecular geometries were first optimized at the GFN2-xTB level (Bannwarth et al., 2019), followed by DFT single-point energy calculations in implicit water dielectric using the CPCM model at the ωB97X-D/6-31+G(d) level of theory (Chai and Head-Gordon, 2008). The oxidation potentials were calculated as the energy difference between the neutral and cationic states and referenced to the Li/Li^+ electrode potential taken from literature (Eq. 1). We benchmarked the xTB protocol against full ωB97X-D/6-31+G(d) optimizations for a 600-molecule subset and found that xTB reproduces DFT oxidation potentials with small systematic errors, supporting its use as a surrogate method for large-scale PFAS screening.

$$E_{ox} = -\frac{\Delta G_{ox}}{nF} - 1.24\,V \qquad (1)$$

Benchmarking showed that the xTB→DFT single-point protocol reproduces full-DFT oxidation potentials with a mean absolute error of only ~0.2 V, justifying its use for large-scale screening (Figure 16).

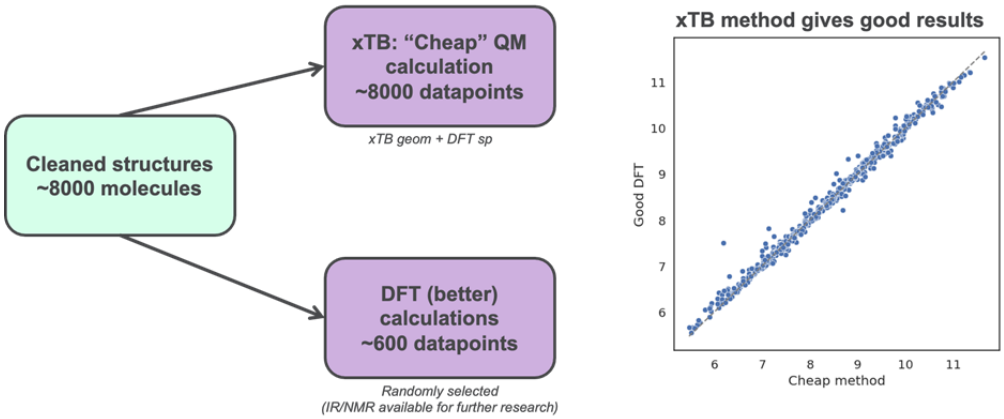

Figure 16. Two-tier quantum-chemical workflow for PFAS oxidation potentials. The left panel shows the use of xTB to screen ~8,000 PFAS structures and the selection of a 600-molecule subset for full ωB97X-D/6-31+G(d) optimizations. The right panel compares oxidation potentials from xTB and DFT, showing that the low-cost protocol closely reproduces the high-level results.

3.1.3 Machine Learning

Molecular structures were represented by Morgan fingerprints, which were generated from SMILES strings (radius = 2, nBits = 2048). A Random Forest regression model was trained to predict oxidation potentials using these fingerprints as input features. The model consisted of 100 estimators and default hyperparameters. The dataset was randomly split into 80 % training and 20 % testing subsets. Model performance was evaluated using the mean absolute error (MAE) and coefficient of determination (R^2).

3.2 Results and Discussion

3.2.1 Distribution of Oxidation Potentials

Computed oxidation potentials for the PFAS-8k dataset range from approximately −2 V to +10 V vs Li/Li^+, with the majority of neutral and cationic species falling between 5 and 9 V (Figure 17). This pattern reflects the high oxidative stability of typical PFAS compounds, driven by strong C–F bonds and the electron-withdrawing nature of fluorine. A small but significant tail at lower potentials (<4 V) contains anionic species and molecules bearing polar/charged headgroups that are thermodynamically far more susceptible to electrochemical oxidation.

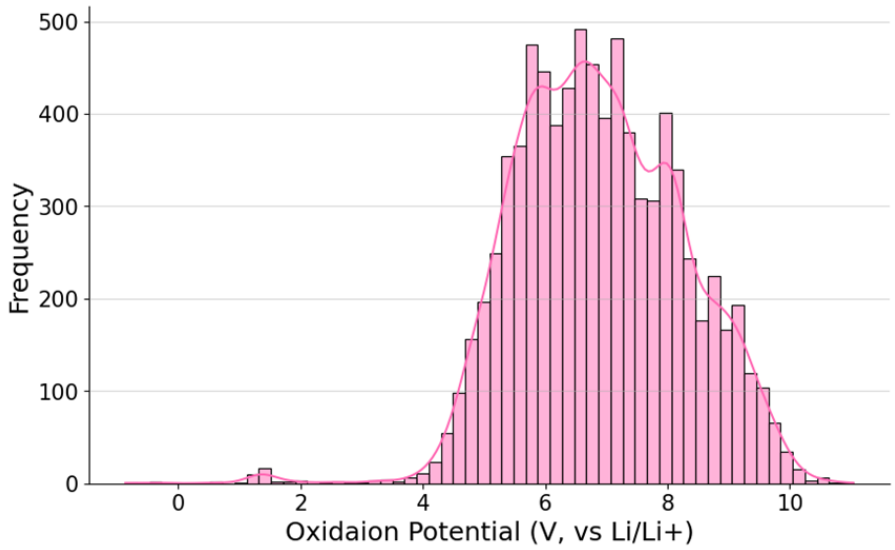

Figure 17. The computed oxidation potentials for the PFAS-8k dataset span a broad range from 4 V to about 10 V vs Li/Li^+. The distribution is unimodal, with most molecules clustered between 5 V and 9 V, and a maximum around 6–7 V.

3.2.2 Influence of Molecular Charge

Neutral PFAS molecules (n = 7730) and cationic species (n = 333) exhibit similar distributions centered around 7 V over the range of 5-9 V. In sharp contrast, anionic species (n = 132: −1 charged; n = 3: −2 charged) are shifted to dramatically lower potentials with most values between 0 and 7 V, and the rare −2 species lie almost entirely below 0 V (Figure 15). This charge-dependent shift indicates that anionic PFAS are thermodynamically easier to oxidize than their neutral or cationic counterparts.

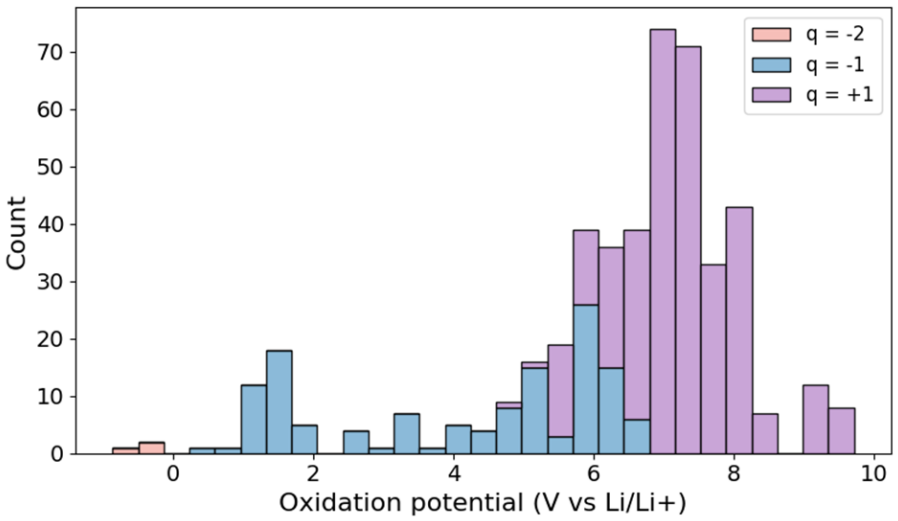

Figure 18. Oxidation potential distribution by molecular charge (non-neutral only).

3.2.3 Functional-Group and Compositional Effects

Analysis of functional group effects (Figure 19) reveals clear links between local chemistry and oxidation potential. Sulfonic and sulfonamide groups show a modest but systematic decrease in E_{ox}. These motifs often carry or stabilize negative charge through S–O bonding, which delocalizes the hole after oxidation and lowers the required potential. Carboxylate ($–COO^-$) and hydroxyl (–OH) groups show a similar trend. Their polarity and ability to form hydrogen bonds with the solvent stabilize oxidized states and shift E_{ox} to lower values. In our dataset these motifs cluster toward the low end of the redox distribution, indicating that oxidation tends to occur in the polar headgroup rather than along the perfluorinated tail. Aromatic rings also reduce E_{ox}, consistent with π-electron delocalization that spreads positive charge over a larger conjugated framework.

In contrast, increasing the number of fluorine atoms leads to higher oxidation potentials. Molecules with many C–F bonds are more electron poor and have very strong C–F interactions, so removing an electron becomes energetically expensive. This trend tracks closely with molecular weight, since heavier PFAS usually have longer perfluoroalkyl chains and higher F content. Together, these patterns show how electron-withdrawing fluorinated backbones raise E_{ox}, while charge-stabilizing headgroups and conjugated units pull it down. They provide a chemically intuitive map between PFAS composition and oxidative stability that we later use as interpretable descriptors in our machine learning models.

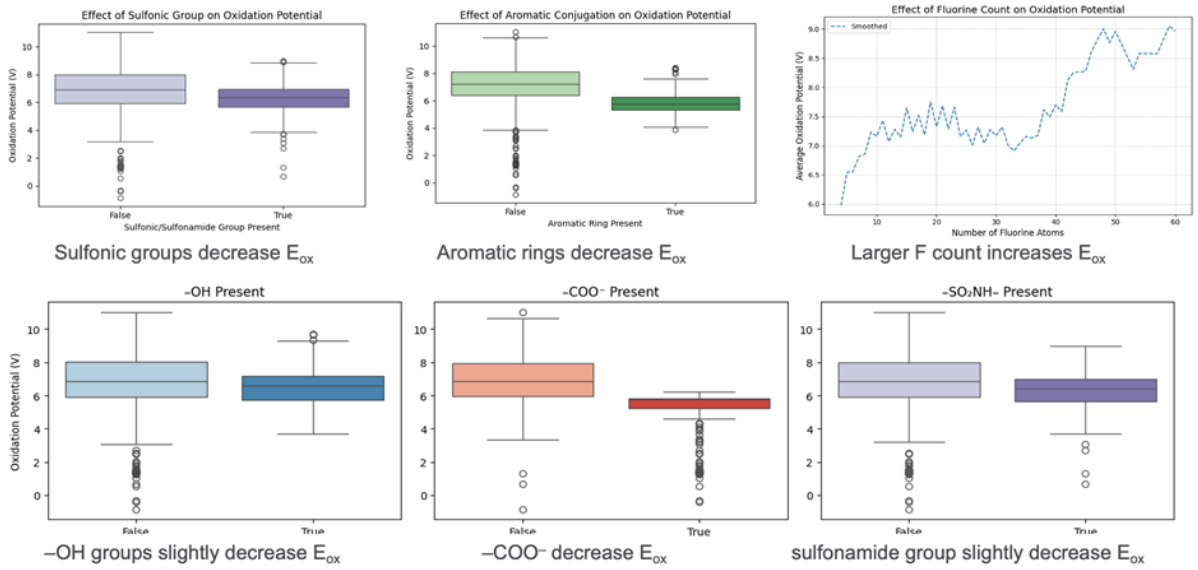

Figure 19. Functional group and composition effects on PFAS oxidation potentials. Box plots show how sulfonic, sulfonamide, carboxylate, hydroxyl, and aromatic motifs shift the E_{ox} distribution relative to molecules that lack these groups. The line plot on the right summarizes the monotonic increase of average oxidation potential with fluorine atom count.

3.2.4 Machine Learning Performance

The Random Forest model trained on Morgan fingerprints (radius 2, 2048 bits) achieved high predictive accuracy, with MAE = 0.23 V and R^2 = 0.91 on a 20% test set. The parity plot (Figure 20) shows that most points lie close to the 1:1 line, which indicates that the model reproduces oxidation potentials across the full range of values. Performance is stable from low to high potentials, suggesting that the fingerprint representation captures the structural patterns that control redox stability. A small number of outliers occur at the lowest potentials (around 2 V), where the data are sparse and include anionic species. Overall, these results show that a relatively simple Random Forest model can learn PFAS redox trends from molecular fingerprints without more complex feature engineering.

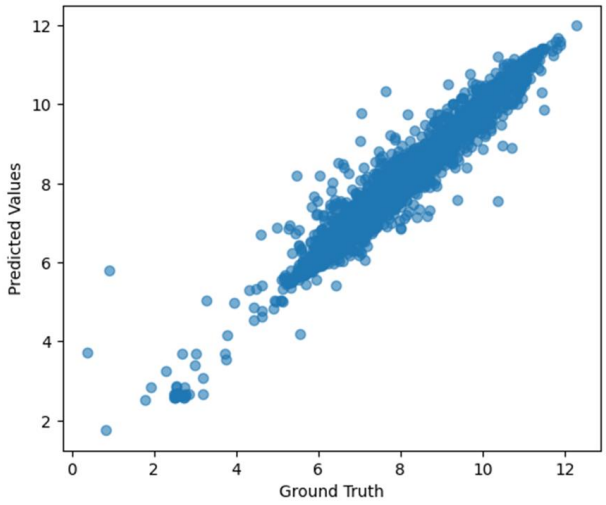

Figure 20. Random Forest regression Parity plot showing strong agreement between predicted and DFT-calculated oxidation potentials (MAE = 0.23 V, R^2 = 0.91).

3.3 Conclusions for PFAS Degradation Potentials

We have established the first comprehensive thermodynamic map of electrochemical oxidation potentials for over 8,200 environmentally relevant PFAS using a hierarchical quantum-chemical workflow validated against high-level DFT. Key findings include:

1. Most legacy PFAS require oxidation potentials >6 V vs Li/Li^+, explaining their resistance to conventional electrochemical treatment.

2. Anionic and polar-headgroup-containing PFAS exhibit lower potentials, identifying them as priority targets for advanced destructive technologies.

3. A simple Random Forest model trained on Morgan fingerprints predicts oxidation potentials with minimal errors, enabling instantaneous screening of new or untested PFAS structures.

The framework presented here is general and readily extendable to other degradation metrics (e.g., C–F bond dissociation energies, adsorption free energies on electrode surfaces) and emerging contaminants. The results may also provide some insights for evaluating transformation of PFAS precursors in subsurface environment.

4.0 SRNL Sampling on the Savannah River Site

The ability to detect emerging PFAS contaminants algorithmically from mass spectra is driven by the breadth of data that is available for training ML algorithms. That is to say, robustness in the algorithmic approach is gained when algorithms are exposed to different sites, with different levels of exposure, and different compounds. Furthermore, variations in the analytical methods enable different resolutions or capabilities. While there is an existing standard approach for PFAS sampling and analysis in EPA 1633, the method can be expensive and laborious to execute, potentially serving as a limiter in the realm of data collection. Hence, the ability to leverage information provided from other data sources that include different analytical techniques is further enabling for ML algorithms. As part of this effort, the team thus sought to identify existing environmental sampling datasets that begin to capture these variations. SRNL internally funded a seedling effort in FY23 titled “Rapid Screening for PFAS”, which produced one such data set that the team has repurposed in the AI for PFAS (henceforth, AI4PFAS) effort for preliminary exploratory data analysis that may guide future efforts. In particular, the team sought to use knowledge of site history, coupled with historical groundwater monitoring, to guide targeted PFAS sampling that could help characterize “high”, “medium”, and “low/no” exposure signals that would be expected. The sampling approach and analysis will be presented in the following subsections.

4.1 Sampling

Samples for this project were collected on the DOE’s Savannah River Site (SRS). The current level of PFAS contamination is not yet well characterized for the SRS, though knowledge of the site’s activities, as well as historical groundwater monitoring and sampling programs, can inform the extent to which various regions of the site have been exposed to PFAS. In the “Rapid Screening for PFAS” effort, this historical knowledge guided the selection of locations for additional, more targeted PFAS sampling. Approximate regions of the SRS having locations with presumed “high”, “medium”, or “low/no” historical exposure to PFAS were selected and are shown in Figure 21. Regarding presumed “high” exposure areas, D-area is home to a firefighter training facility and has a known existing plume of legacy PFAS contamination. For sampling locations not in D-Area, previous exposures were estimated based on the positioning in their drainage basin and proximity to potential PFAS sources. Referencing Figure 21, the General Separations Area (GSA) on the SRS is a centralized location of facilities where many of the site’s legacy/ongoing DOE operations occurred. While there were firefighter training activities that occurred in GSA, they were not to the extent of those that occurred in D-Area. Hence, downstream locations (Pen Branch and L-Lake) are presumed to have moderate to low levels of exposure based both on distance from potential sources and the level of activities that occurred. Finally, the Upper Three Runs region of sampling is situated upstream of both D-Area and the GSA and was therefore presumed to have “low/no” prior exposure, serving as background for the site.

Coupling knowledge of the site history with site sampling, two sites were selected with no presumed previous exposure to PFAS (Upper Three Runs [UTR] and PB-X [Pen Branch]), four sites were selected with presumed low or moderate exposure (PBr [Pen Branch], L-Lake, DSWM12 [D-Area], DSWM4A [D-Area]), and one site was selected which was known to have high exposure (DSWM11 [D-Area]). Samples were collected using an extendable pole arm sampler designed to hold single use 1-liter high density polyethylene (HDPE) bottles. Unlined HDPE bottles were used to avoid a PFAS background from sampling equipment. Prior to

sampling, all HDPE bottles were rinsed three times with water from the sampling location. Samples were stored refrigerated until spiking and analysis were completed.

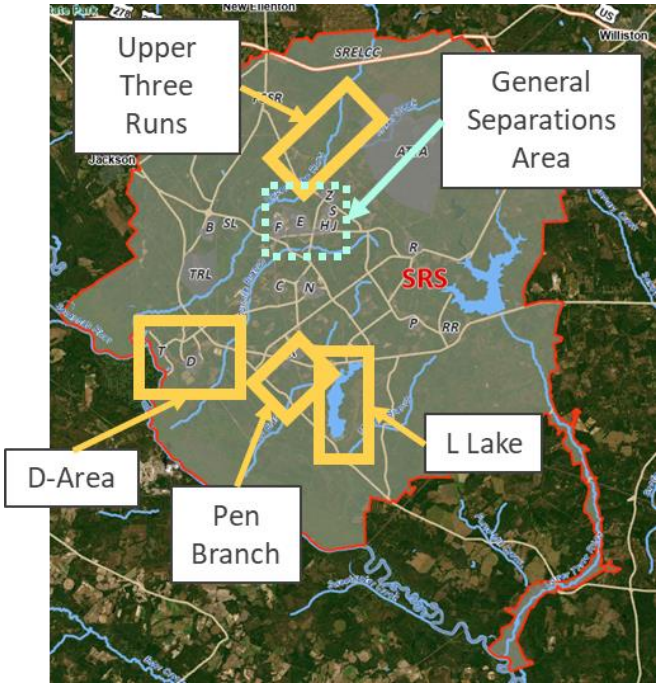

Figure 21. Approximate regions on the Savannah River Site targeted for sampling during the "Rapid Screening for PFAS" effort.

4.2 Sample Preparation and Instrument Analysis

Samples were analyzed with and without the addition of PFAS standards as a means to identify the PFAS signatures that are present in the mass spectra. To perform spiking, per- and polyfluoroalkyl substance (PFAS) standards were purchased as methanolic solutions from Wellington Laboratories (Ontario, CA), shown in Figure 22. M2 PFOA, d3 MEFOSAA, and MPFOS were added to all samples at the 100 ng/L level. Additional native PFAS were added in varying concentrations of 0 ng/L, 10 ng/L, 30 ng/L, 100 ng/L, 300 ng/L, and 1000 ng/L. Additional dilutions of these standards were completed in methanol (Chromosolv, Honeywell International, Charlotte, NC). Octadecylamine (Sigma Aldrich, >99%, Burlington, MA) was used as the adsorbent and was applied to borosilicate glass melting point tubes (Kimble, Vineland, NJ) by first dissolving in hexane (Chromosolv, Honeywell International, Charlotte, NC). Plastic products used for this work including 2 mL vials (Thermo Scientific, Waltham, MA), pipette tips (Eppendorf EP tips, Hamburg, Germany), 15 mL polypropylene centrifuge tubes (Falcon, Irvin, CA), and one-liter Nalgene bottles (Thermo Scientific, Waltham, MA) were selected based on their known low PFAS background. Deionized water was from an ultrapure water system (Supelco, Bellefonte, PA). FomblinY (HVAC 16/6, Sigma-Aldrich, St. Louise, MO) was used for mass calibration of the mass spectrometer (MS).

	A	B	C	D	E
1	**Negative Ionization**				Indicative m/z
2		formula for acid		MW of compound	
3	Perfluoro-n-butanoic acid	C4HF7O2	PFBA	213.986476	212.9792
4	Perfluoro-n-pentanoic acid	C5HF9O2	PFPeA	263.983282	262.976
5	Perfluoro-n-hexanoic acid	C6HF11O2	PFHxA	313.980088	312.9728
6	Perfluoro-n-heptanoic acid	C7HF13O2	PFHpA	363.976894	362.9696
7	Perfluoro-n-octanoic acid	C8HF15O2	PFOA	413.9737	412.9664
8	Perfluoro-n-nonanoic acid	C9HF17O2	PFNA	463.970506	462.9632
9	Perfluoro-n-decanoic acid	C10HF19O2	PFDA	513.967312	512.96
10					
11	Potassium perfluoro-1-butanesulfonate	C4HF9O3S	L-PFBS	299.950269	298.943
12	Sodium perfluoro-1-hexanesulfonate	C6HF13O3S	L-PFHxS	399.943881	398.9366
13	Sodium perfluoro-1-octanesulfonate	C8HF17O3S	L-PFOS	499.937493	498.9302
14					
15	2,3,3,3-Tetrafluoro-2-(1,1,2,2,3,3,3-heptafluoroproproxy)propanoic acid	C6HF11O3	HFPO-DA	329.975003	284.977348
16	Sodium dodecafluoro-3H-4,8-dioxanonanoate	C7H2F12O4	NaDONA	377.976146	376.9689
17					
18	Potassium 9-chlorohexadecafluoro-3-oxanonane-1-sulfonate	C8ClF16HO4S	9Cl-PF3ONS	531.902858	530.8956
19	Potassium 11-chloroeicosafluoro-3-oxaundecane-1-sulfonate	C10ClF20HO4S	11Cl-PF3OUdS	631.89647	630.8892
20					
21					
22	M2 PFOA	[13C]2C6HF15O2	M2 PFOA	415.9737	414.9731
23	d3 MEFOSAA	C11H3D3F17NO4S	d3 MEFOSAA	573.974607	572.9862
24	M PFOS	[13C]4C4HF17O3S	M PFOS	503.937493	502.9436
25					
26		formula for alcohol		MW	
27	2-Perfluorooctyl ethanol	CF3(CF2)7CH2CH2OH	8:2 FTOH	464.006891	495.9973

Figure 22. Native PFAS standards and mass labeled internal standards added to the environmental samples from the SRS, along with the indicative mass to charge ratio (m/z) that would be seen in the mass spectra.

To prepare adsorbent probes for concentrating samples, octadecylamine was first dissolved to its solubility limit in hexane. After dissolving octadecylamine, glass melting point tubes were submerged closed-side down into the octadecylamine-hexane solution for 30-seconds. After this the adsorbent probe was lifted, and the probe was dried. This was repeated for a total of two layers of adsorbent on the probe.

Samples were prepared including spikes of native PFAS compounds and internal standards as needed in 10 mL volumes in 15 mL polypropylene centrifuge tubes. Methanol in the mixture was < 0.5% by volume. Each sample was vortexed after spiking to ensure homogeneity. One adsorbent probe was added per sample and then the system was gently shaken (300 rpm) for 1-hour at room temperature to adsorb PFAS from the sample. To avoid temperature impacts, unknown and calibration samples were prepared and analyzed together in the same batch.

Samples were analyzed using direct analysis in real time (DART) coupled with high-resolution time-of-flight (TOF) MS. Helium was used as the ionization gas and the sample was analyzed in negative ionization mode. The DART source was set to 550 ˚C and was positioned 14 mm from the inlet to the AccuTOF. The AccuTOF was set to scan from 100-1500 *m/z*. The MS was mass calibrated at the end of each analytical run using FomblinY. In total, 64 mass spectra from the "Rapid Screening for PFAS" effort were provided for preliminary exploratory data analysis in the AI4PFAS effort.

4.3 Exploratory Data Analysis

Figure 23 shows the mass spectra of sampled water from all locations. There are notable similarities between the spectra for DSWM4A (D-Area), DSWM11 (D-Area), L-Lake, Pen Branch (PBr) and Upper Three Runs (UTR). In contrast, the spectra for DSWM12 and Pen Branch (PBx) show greater differences in the general profile, while still having visually similar groupings of high intensity peaks. DSWM4A, DSWM11, and DSWM12 all reside in the same area of the site (D-Area). However, DSWM4A and DSWM11 were sampled from a drainage ditch, whereas DSWM12 was sampled from a pond. Hence, the differences in the spectra likely arise from the differences in the overall content of organic matter present in the samples at different locations, illustrating the complexity of real environmental samples. UTR, being upstream of most site facilities, is the likely best representative of an environmental background for samples from the region without impacts from the site.

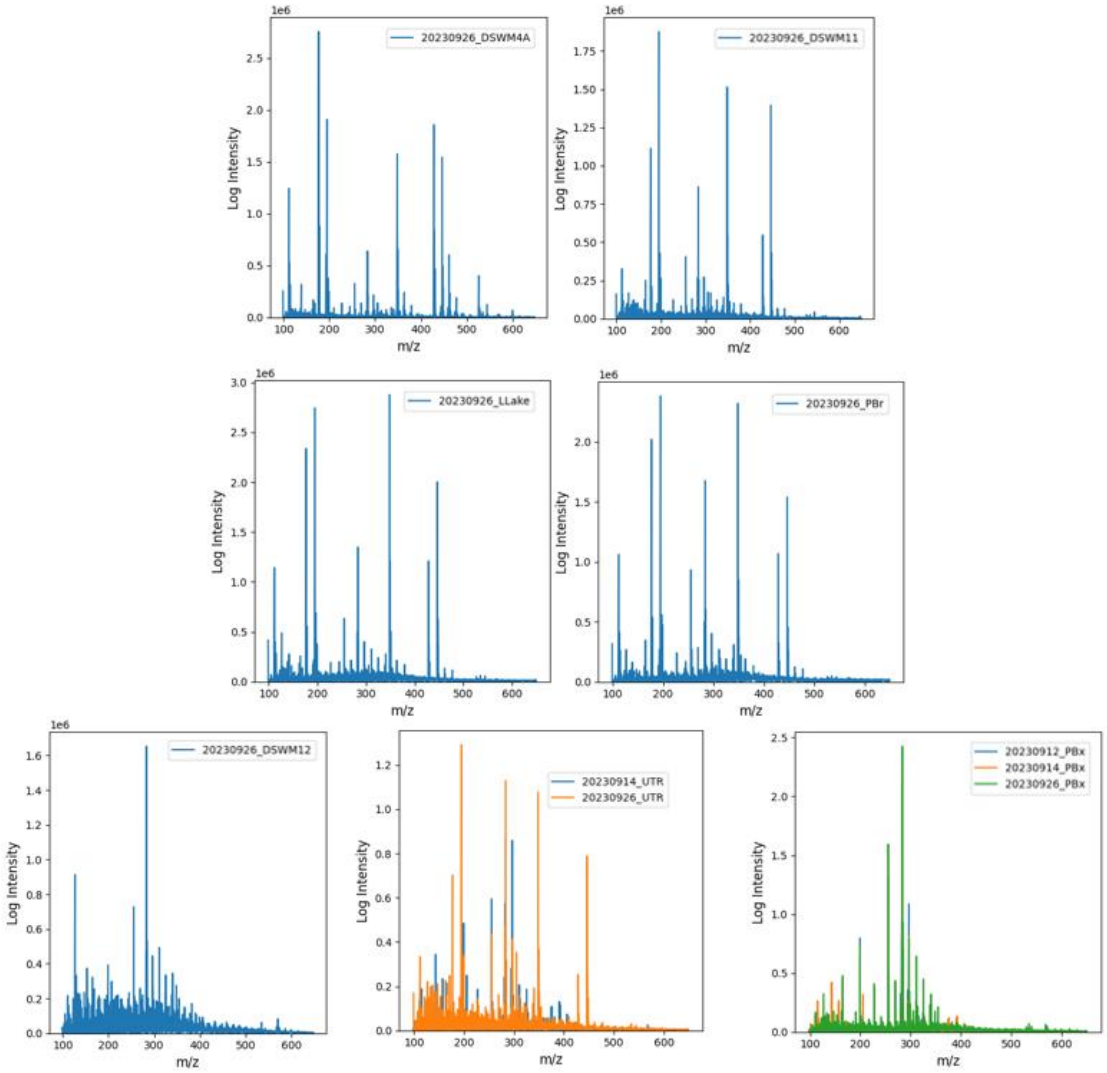

Figure 23. Mass spectra of sampled water without any spiking from all locations. Multiple measurements on the same plot indicates instrumental analysis of the same aqueous sample.

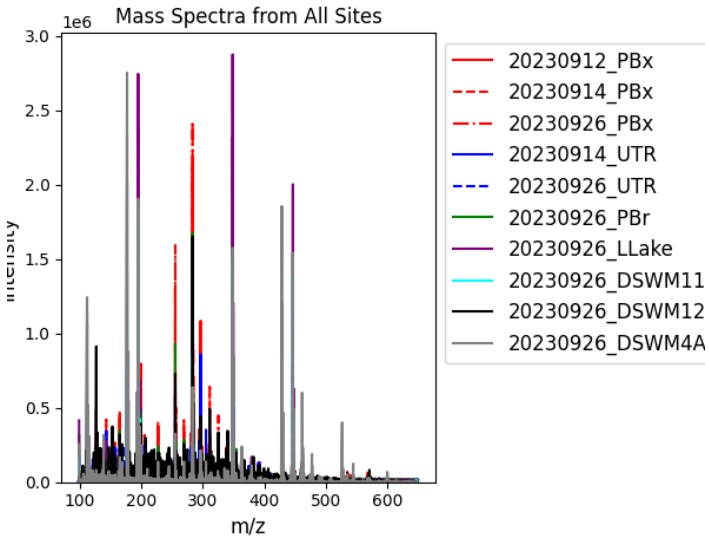

Figure 24. Overlapping mass spectra without spiking from all sites. Multiple measurements on the same plot indicates instrumental analysis of the same sample across different days.

Spiking with the native PFAS standards from Figure 22 was performed in different concentrations to further identify and quantify the peaks of interest in the environmental samples. In general, the higher the concentration added, the higher intensity of those peaks.(Note: the current scope in AI4PFAS is not to be quantitative but rather characterize the challenges with real environmental samples – a separate publication is in progress with quantitative analysis). Figure 25 shows the mass spectra for DSWM11 with and without the addition of native PFAS, while Figure 26 focused on the *m/z* for PFNA. While the *m/z* for PFNA cannot be clearly detected against nearby peaks, in the sample with only mass labeled internal standards added, the *m/z* for PFNA is clearly seen after the native PFAS are spiked into the sample, as would be expected to be seen in a sample collected from an area with PFNA contamination. In Figure 27, it is shown that the natural site water does not have a background of the mass labeled internal standards d3 MEFOSAA. This is as anticipated. These zoomed-in views demonstrate the importance of looking closely at the specific mass spectral signatures for these PFAS, as similar *m/z* can be present at higher abundances, as well as the potential quantitative capabilities enabled by spiking the samples at different concentrations. Additional work is in progress to fully review samples for these PFAS signatures.

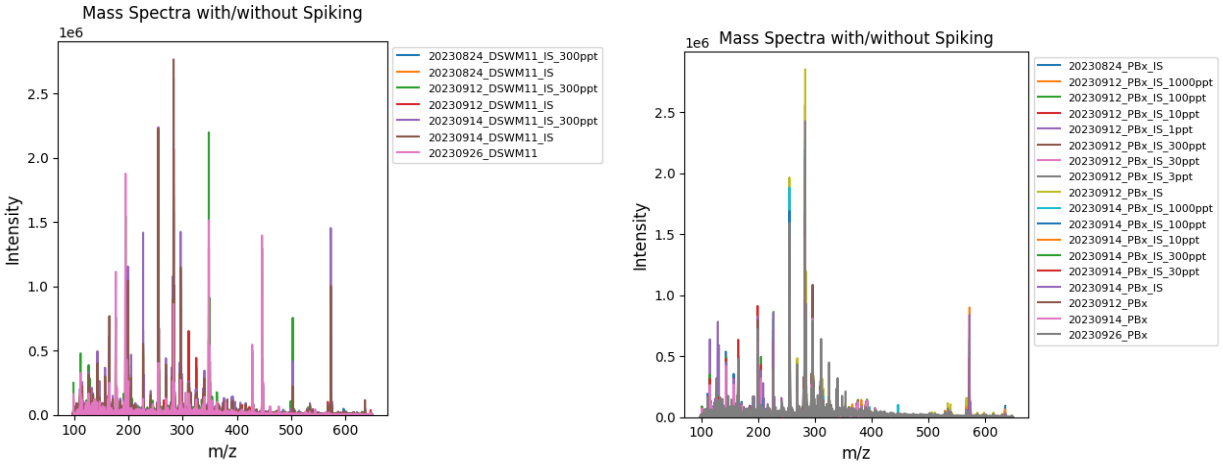

Figure 25. PFAS mass spectra for DSWM11 and PBx locations with and without spiking. “IS” appended to the name indicates only the mass labeled internal standards (M2PFOA, d3 MEFOSAA, and M PFOS) are added to the sample at a concentration of 100 ng/L, and “Xppt” indicates the full suite of native PFAS standards from Figure 22 have been added.

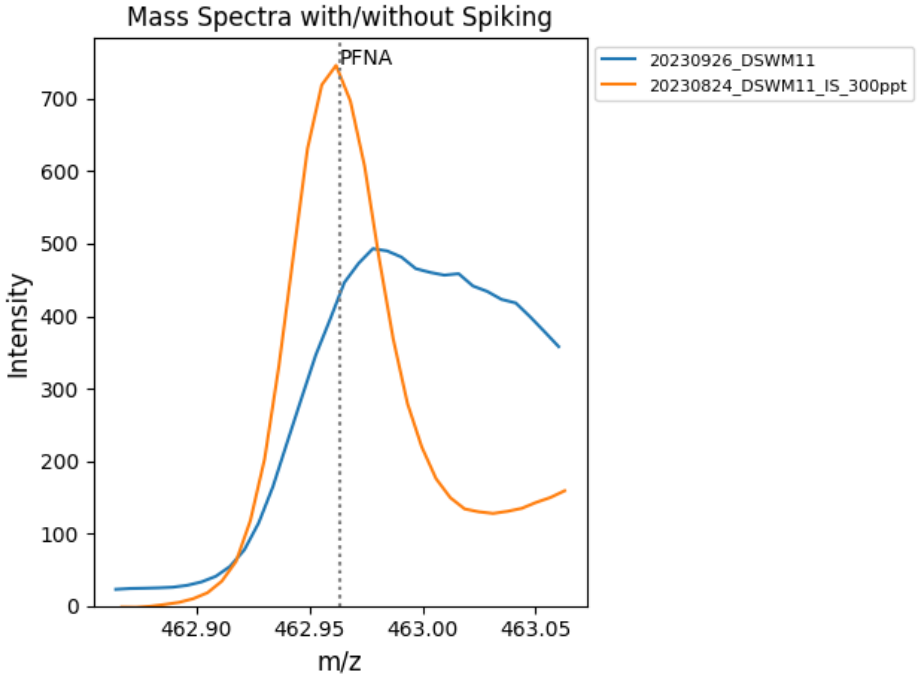

Figure 26. Zoomed in mass spectra for DSWM11 with and without spiking showing PFNA. “IS” appended to the name indicated only the mass labeled internal standards were added while 300 ppt indicates the full suite of native PFAS standards were added.

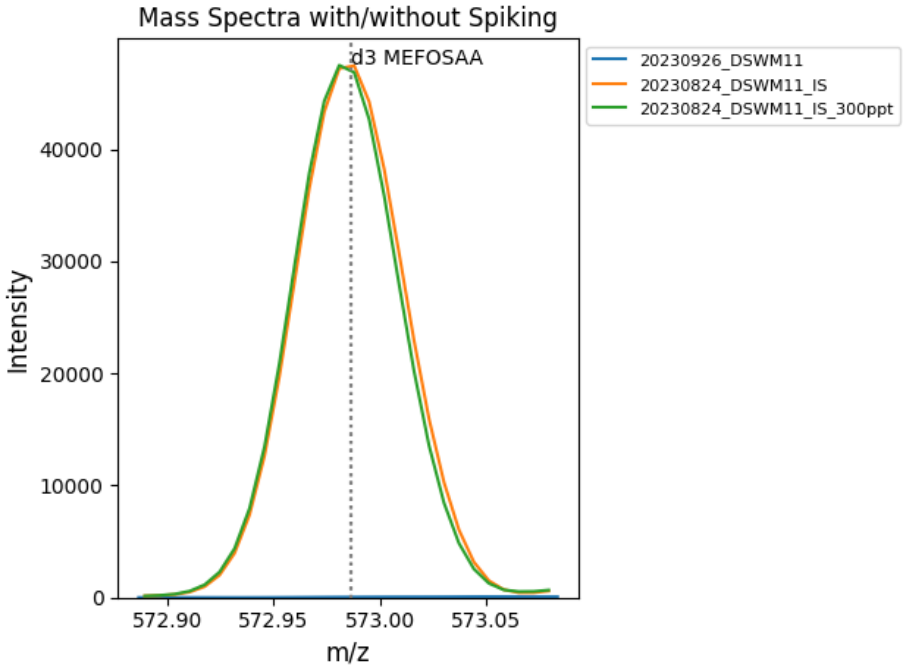

Figure 27. Zoomed in mass spectra for DSWM11 with and without spiking showing d3 MEFOSAA.

Figure 28 and Figure 29 show the zoomed in mass spectra for deionized water and the PBx location, respectively, with the different levels of spiking. Zooming in around the m/z for PFPeA shows that the compound was readily detected in de-ionized water but proved difficult to disambiguate in natural waters due to the presence of other ionic compounds. Notably, even higher concentration spikes did not aid in disambiguating the presence of PFPeA. Hence, such challenges with the detection of PFAS compounds must be accounted for both in the characterization approaches and the data analytics pipelines for real environmental samples.

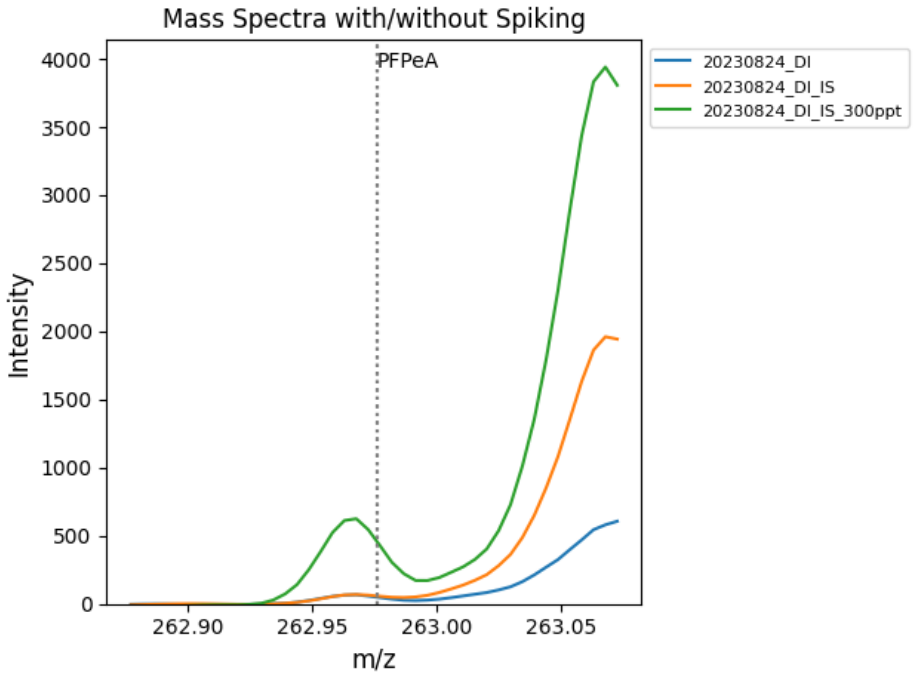

Figure 28. Zoomed in mass spectra for de-ionized water with and without spiking showing PFPeA. “IS” appended to the name indicates only M2PFOA, d3 MEFOSAA, and MPFOS are added to the sample at a concentration of 100 ng/L, and “300 ppt” indicates the full suite of PFAS standards from Figure 22 have been added.

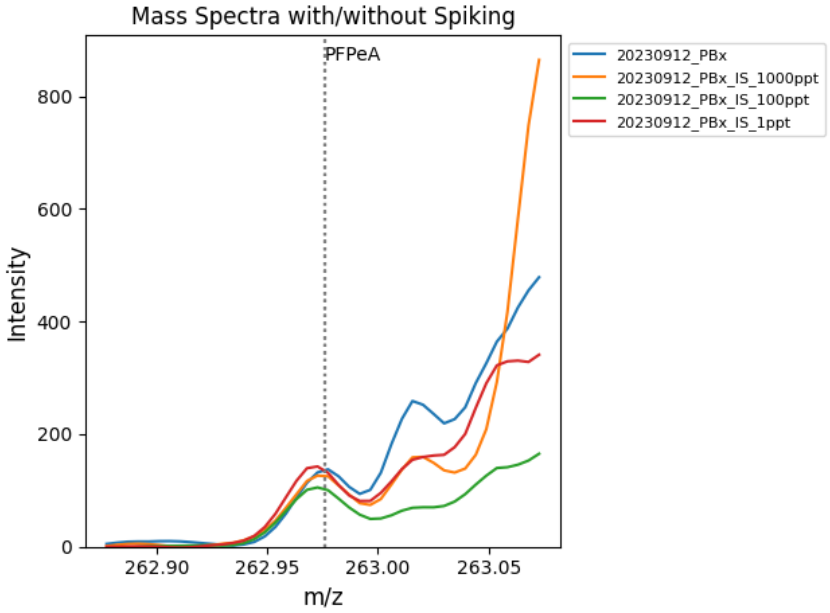

Figure 29. Zoomed in mass spectra for PBx with and without spiking showing PFPeA. “IS” appended to the name indicates only M2PFOA, d3 MEFOSAA, and M PFOS are added to the sample at a concentration of 100 ng/L, and “300 ppt” indicates the full suite of PFAS standards from Figure 22 have been added.

5.0 Concluding Remarks and Future Work

The three efforts undertaken by the individual laboratories represent capabilities that advance key elements of environmental management for PFAS contamination, from optimal sampling (SRNL); to rapid detection, characterization, and source attribution (PNNL); and finally, strategies for effective remediation (ANL). Leveraging analytical capabilities and AIML tools, together, they have the potential to serve as a basis for the next-generation strategy in monitoring and verification of PFAS contamination.

Understanding the challenges that different complex matrices present to the detection of PFAS can guide environmental sampling and inform sample preparation strategies to maximize the detectability of potential PFAS. The ability to rapidly detect and characterize environmental samples to various PFAS classes will enable us to attribute PFAS contamination source. And once PFAS class from environmental samples is determined and unknown chemicals that are potentially PFAS are characterized, having a predictive capability to better understand their physicochemical properties with an eye towards identifying the most effective methods for degrading and transforming these chemicals would facilitate remediation efforts.

Continued efforts to advance these nascent capabilities will enable us to be informed and strategic to not only address current EM and remediation efforts, but to also remain prepared for the challenges that emerging contaminants and other unknown chemicals bring. While these capabilities were developed and evaluated for PFAS, the methodologies can generalize to other chemical classes that may be of interest to environmental management.

6.0 References

3M Canada Company. Material Safety Data Sheet. February 27, 2015. https://multimedia.3m.com/mws/mediawebserver?mwsId=SSSSSuUn_zu8lZNU4xtxoY_BPv70kDVFNvu9lxtD7SSSSSS--.

Ames, J.L., V. Sharma, and K. Lyall. 2025. “Effects of early-life PFAS exposure on child neurodevelopment: a review of the evidence and research gaps.” *Current Environmental Health Reports*, Vol 12, No. 9. https://doi.org/10.1007/s40572-024-00464-5

Appleby, A. MurmurHash3. GitHub, 2015. https://github.com/aappleby/smhasher/blob/master/src/MurmurHash3.cpp

Baker, T. J.; Tonkyn, R. G.; Thompson, C. J.; Dunlap, M. K.; Koster van Groos, P. G.; Thakur, N. A.; Wilhelm, M. J.; Myers, T. L.; Johnson, T. J. An Infrared Spectral Database for Gas-Phase Quantitation of Volatile Per- and Polyfluoroalkyl Substances (PFAS). J. Quant. Spectrosc. Radiat. Transfer. 2023, 295, 108420. DOI 10.1016/j.jqsrt.2022.108420.

Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB - An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. J. Chem. Theory Comput. 2019, 15 (3), 1652–1671. doi: 10.1021/acs.jctc.8b01176.

Barzen-Hanson, K. A.; Roberts, S. C.; Choyke, S.; Oetjen, K.; McAlees, A.; Riddell, N.; McCrindle, R.; Ferguson, P. L.; Higgins, C. P.; Field, J. A. Discovery of 40 Classes of Per- and Polyfluoroalkyl Substances in Historical Aqueous Film-Forming Foams (AFFFs) and AFFF-Impacted Groundwater. Environ. Sci. Technol. 2017, 51, 2047-2057. DOI: 10.1021/acs.est.6b05843.

Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 2004, 6(1), 20-29. DOI:10.1145/1007730.1007735.

Blagus, R.; Lusa, L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics, 2013, 14:106. DOI:10.1186/1471-2105-14-106.

Bline, A.P., J.C. DeWitt, C.F. Kwiatkowski, K.E. Pelch, A. Reade, and J.R. Varshavsky. 2024. “Public health risks of PFAS-related immunotoxicity are real.” *Current Environmental Health Reports*, Vol. 11, No. 2, ppg. 118 – 127. https://doi.org/10.1007/s40572-024-00441-y.

Cahuas, L.; Muensterman, D.J.; Kim-Fu, M.L.; Reardon, P.N.; Titaley, I.A.; Field, J.A. Paints: a source of volatile PFAS in air – potential implications for inhalation exposure. Environ. Sci. Technol. 2022, 56, 17070–17079. DOI: 10.1021/acs. Est.2c04864.

Chai, J.D; Head-Gordon, M. Long-Range Corrected Hybrid Density Functionals with Damped Atom-Atom Dispersion Corrections. Phys. Chem. Chem. Phys. 2008, 10, 6615-6620.

Chawla, N.V.; Bowyer, K.W.; Hall, L.O., Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. JAIR, 2002, 16, 321-357. DOI:10.1613/jair.953.

Cleston, L.; Charles, C., Degradation of Poly- and Perfluoroalkyl Substances (PFAS) in Water via High Power, Energy-Efficient Electron Beam Accelerator. 2024, DOE-3M-9132. DOI: https://doi.org/10.2172/2349585.

Coperchini, F., L. Croce, G. Ricci, F. Magri, M. Rotondi, M. Imbriani, and L. Chiovato. 2021. "Thyroid disrupting effects of old and new generation PFAS." *Frontiers in Endocrinology*, Vol. 11. doi: 10.3389/fendo.2020.612320.

Cui, Y.; Wang, S.; Han, D.; Yan, H. Advancements in detection techniques for per- and polyfluoroalkyl substances: A comprehensive review. Trends Analyt. Chem. 2024, 176, 117754. DOI 10.1016/j.trac.2024.117754.

D'Hollander, W. D.; de Voogt, P.; De Coen, W.; Bervoets, L. Perfluorinated Substances in Human Food and Other Sources of Human Exposure. Reviews of Environmental Contamination and Toxicology, 2010, 208, 179-215.

Ernst, A., N. Brix, L.L.B. Lauridsen, J. Olsen, E.T. Parner, Z. Liew, L.H. Olsen, and C.H. Ramlau-Hansen. 2019. "Exposure to perfluoroalkyl substances during fetal life and pubertal development in boys and girls from the Danish National Birth Cohort.", *Environmental Health Perspectives*, Vol. 127, No. 1, pp. 017004-1 – 15. https://doi.org/10.1289/EHP3567

G. Pitter, M.Z. Jeddi, G. Barbieri, M. Gion, A.S.C. Fabricio, F. Dapra, F. Russo, T. Fletcher, and C. Canova. 2020. "Perfluoroalkyl substances are associated with elevated blood pressure and hypertension in highly exposed young adults." *Environmental Health*, Vol. 19. No. 102. https://doi.org/10.1186/s12940-020-00656-0

Goralczyk, K.; Pachocki, K/ A/; Hernik, A.; Strucinski, P.; Czaja, K.; Lindh, C. H.; Jonsson, B. A. G.; Lenters, V.; Korcz, W.; Minorczyk, M.; Matuszak, M.; Ludwicki, J. Perfluorinated chemicals in blood serum of inhabitants in central Poland in relation to gender and age. Sci. Total Environ. 2015, 532, 548-555. DOI: 10.1016/j.scitotenv.2015.06.050.

Hughey, K. D.; Gallagher, N. B.; Zhao, Y.; Thakur, N.; Bradley, A. M.; Koster van Groos, P. G.; Johnson, T. J. PFAS remediation: Evaluating the infrared spectra of complex gaseous mixtures to determine the efficacy of thermal decomposition of PFAS. Chemosphere. 2024, 362, 142631. DOI 10.1016/j.chemosphere.2024.142631.

ITRC (Interstate Technology & Regulatory Council). 2023. *PFAS Technical and Regulatory Guidance Document and Fact Sheets* PFAS-1. Washington, D.C.: Interstate Technology & Regulatory Council, PFAS Team. https://pfas-1.itrcweb.org/

Kim, N.; Elbert, J.; Shchukina, E.; Su, X. Integrating redox-electrodialysis and electrosorption for the removal of ultra-short- to long-chain PFAS. Nat. Commun. 2024, 15, 8321. https://doi.org/10.1038/s41467-024-52630-w

Lemaître, G., Nogueira, F., Aridas, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. JMLR, 2017, 18(17), 1-5. https://jmlr.org/papers/v18/16-365.html

Li, H.; Song, X.; Shan, X., Liu, F.; Liu, D.; Feng, M.; Xu, X.; Zhang, Q.; Yin, Y.; Cai, Y. Photochemical Degradation of PFAS: Mechanistic Insights and Design Strategies for

Homogeneous and Heterogeneous Systems. Small 2025, 21(40), e06040. https://doi.org/10.1002/smll.202506040

Li, Ping, Trevor Hastie, and Kenneth Church (2006). "Very sparse random projections". *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 287–296. doi:10.1145/1150402.1150436. ISBN 1-59593-339-5. S2CID 7995734.

Longendyke, G. K.; Katel, S.; Wang, Y. PFAS fate and destruction mechanisms during thermal treatment: a comprehensive review. Environ. Sci.: Processes Impacts. 2022, 24, 196. DOI: 10.1039/d1em00465d.

Mahinroosta, R.; Senevirathna, L. A review of the emerging treatment technologies for PFAS contaminated soils. J. Environ. Manag. 2020, 255, 109896. DOI: 10.1016/j.jenvman.2019.109896.

McInnes, L., J. Healy, and J. Melville (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. DOI: 10.48550/arXiv.1802.03426.

Moody, J. (1988). "Fast learning in multi-resolution hierarchies", In D. Touretzky (Ed.), *Advances in Information Processing Systems 1 (NIPS 1988)*, (pp. 29 - 39). MIT Press.

Nahar, K.; Zulkarnain, N. A.; Niven, R. K. A Review of Analytical Methods and Technologies for Monitoring Per-and Polyfluoroalkyl Substances (PFAS) in Water. Water 2023, 15(20), 3577. DOI 10.3390/w15203577.

NCI, 2024. PFAS Exposure and Risk of Cancer, Division of Cancer Epidemiology and Genetics, National Cancer Institute, https://dceg.cancer.gov/research/what-we-study/pfas

Place, B. J.; Field, J. A. Identification of Novel Fluorochemicals in Aqueous Film-Forming Foams Used by the US Military. Environ. Sci. Technol. 2012, 46, 7120-7127. DOI: 10.1021/es301465n.

Qin, X., Y. Zhuang, and B. Shi. 2024. “PFAS promotes disinfection byproduct formation through triggering particle-bound organic matter release in drinking water pipes.” *Water Research*, Vol. 254. 2024. https://doi.org/10.1016/j.watres.2024.121339

Ragland, J.M. and Place, B.J., 2024. A Portable and Reusable Database Infrastructure for Mass Spectrometry, and Its Associated Toolkit (The DIMSpec Project). Journal of the American Society for Mass Spectrometry, 35(6), pp.1282-1291.

Renner, R. The long and short of perfluorinated replacements. Environ. Sci. Technol. 2006, 40, 1, 12–13 https://doi.org/10.1021/es062612a

Sidnell, T.; Wood, R.J.; Hurst, J.; Lee, J.; Bussemaker, M.J. Sonolysis of per- and poly fluoroalkyl substances (PFAS): A meta-analysis. Ultrason. Sonochem. 2022, 8, 105944. https://doi.org/10.1016/j.ultsonch.2022.105944

Singh, R.K., S. Fernando, S.F. Baygi, N. Multari, S.M. Thagard, and T.M. Holsen. 2019. “Breakdown products from perfluorinated alkyl substances (PFAS) degradation in a plasma-based water treatment process.” *Environmental Science & Technology*, Vol. 43, ppg. 2731 – 2738, 2019. DOI: 10.1021/acs.est.8b07031.

Sznajder-Katarzynska, K.; Surma, M.; Cieslik, I., A review of perfluoroalkyl acids (PFASs) in terms of sources, applications, human exposure, dietary intake, toxicity, legal regulation, and methods of determination. J. Chem., 2019, 2717528. DOI: 10.1155/2019/2717528.

USEPA CompTox Chemicals Dashboard PFAS Master List, 2023.

USEPA, 2024a. PFAS Strategic Roadmap: EPA's Commitments to Action 2021-2024.

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, December 2024.

Johnson, W.B and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space. In *Conference in modern analysis and probability*, vol 26 of *Contemporary Mathematics*, pages 189-206. American Mathematical Society, 1984.

Winquist, A.; Hodge, J. M.; Diver, W. R.; Rodriguez, J. L.; Troeschel, A. N.; Daniel, J.; Teras, L. R. Case-Cohort Study of the Association between PFAS and Selected Cancers among Participants in the American Cancer Society's Cancer Prevention Study II LifeLink Cohort. EHP. 2023, 131(12), 127007. DOI: 10.1289/EHP13174.

Zeng, M.; Zou, B.; Wei, F.; Liu, X.; Wang, L. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), Chongqing, China, 2016, pp. 225-228, DOI:10.1109/ICOACS.2016.7563084.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov