Environmental Science & Technology

pubs.acs.org/est

Article

Target and Suspect Screening Integrated with Machine Learning to Discover Per- and Polyfluoroalkyl Substance Source Fingerprints

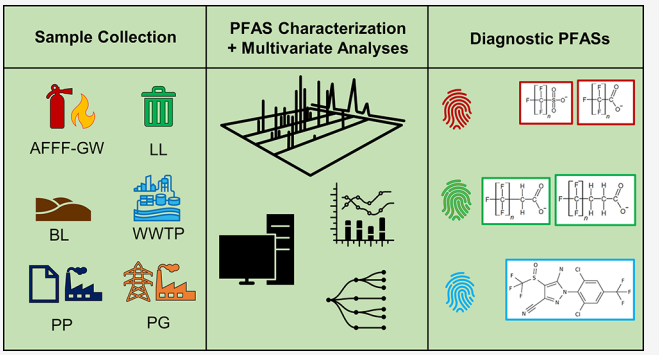
Nayantara T. Joseph,[#] Trever Schwichtenberg,[#] Dunping Cao, Gerrad D. Jones, Alix E. Rodowa, Morton A. Barlaz, Joseph A. Charbonnet, Christopher P. Higgins, Jennifer A. Field,* and Damian E. Helbling*

Cite This: https://doi.org/10.1021/acs.est.3c03770

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: This study elucidates per- and polyfluoroalkyl substance (PFAS) fingerprints for specific PFAS source types. Ninety-two samples were collected from aqueous film-forming foam impacted groundwater (AFFF-GW), landfill leachate, biosolids leachate, municipal wastewater treatment plant effluent (WWTP), and wastewater effluent from the pulp and paper and power generation industries. High-resolution mass spectrometry operated with electrospray ionization in negative mode was used to quantify up to 50 target PFASs and screen and semi-quantify up to 2,266 suspect PFASs in each sample. Machine learning classifiers were used to identify PFASs that were diagnostic of each source type. Four C5−C7 perfluoroalkyl acids and one suspect PFAS (trihydrogen-substituted fluoroethernonanoic acid) were diagnostic of AFFF-GW. Two target PFASs (5:3 and 6:2 fluorotelomer carboxylic acids) and two suspect PFASs (4:2 fluorotelomer-thia-acetic acid and *N*-methylperfluoropropane sulfonamido acetic acid) were diagnostic of landfill leachate. Biosolids leachates were best classified along with landfill leachates and *N*-methyl and *N*-ethyl perfluorooctane sulfonamido acetic acid assisted in that classification. WWTP, pulp and paper, and power generation samples contained few target PFASs, but fipronil (a fluorinated insecticide) was diagnostic of WWTP samples. Our results provide PFAS fingerprints for known sources and identify target and suspect PFASs that can be used for source allocation.



KEYWORDS: *PFAS forensics, source allocation, principal component analysis, hierarchical clustering, support vector classification, logistic regression, random forest, multivariate analyses*

INTRODUCTION

Per- and polyfluoroalkyl substances (PFASs) are a class of chemicals used in mixtures in a variety of commercial and industrial applications and are common environmental contaminants.[1] Environmental monitoring has revealed a variety of PFAS point sources including fire-fighter training sites impacted by aqueous film-forming foam (AFFF),[2,3] land-applied biosolids,[4−6] municipal landfill leachate,[7−12] municipal wastewater treatment plant (WWTP) effluent,[13−19] and the pulp and paper industry.[20,21] Industries including fluorochemical manufacturing,[22−26] electronics manufacturing,[27] paper manufacturing,[20] and electroplating[28] are also PFAS point sources, but access to effluent samples is often limited or restricted.

There is increasing interest in developing forensics tools to enable source allocation of PFASs in the environment.[29−32] Several features of PFASs make them ideal candidates for environmental forensics including their persistence in the environment and the expectation that specific PFASs are used for certain commercial and industrial applications: these PFASs could collectively define a PFAS "fingerprint" that is specific to known sources.[31,33] Such PFAS fingerprints, when viewed through the lens of a conceptual site model and hydrogeologic data for a particular site, could aid in identifying PFAS releases potentially responsible for PFAS contamination. Recent studies combined environmental sample analysis with multivariate statistics[34] and/or machine learning techniques[30,35] to identify PFAS sources across geographic scales. Despite the successes of these studies, there are at least three major factors that limit the potential application of these techniques for broad and comprehensive source allocation or source tracking of PFASs. First, most previous studies restricted the analysis to relatively

Received: May 18, 2023
Revised: August 25, 2023
Accepted: August 28, 2023

few perfluoroalkyl carboxylates (PFCAs), perfluoroalkyl sulfonates (PFSAs), and other PFASs included in widely implemented analytical methods.[30,34−37] However, differentiating sources based on so few target PFASs is challenging because these PFASs are common to many sources.[38] Second, analytical data reported from multiple laboratories is often aggregated in an effort to generate datasets that are large enough to be amenable to multivariate statistics and/or machine learning techniques for the purposes of source allocation.[30,35] This approach can introduce biases to downstream analyses from differences in sample handling and extraction, the PFASs measured, data quality, and reporting limits. Third, results of previous studies are typically restricted to *inferences* of sources based on the composition of PFASs exhibiting highly correlated spatiotemporal variability in occurrence and concentration.[34,36,37] For example, samples that contain high abundances of PFSAs are often inferred to be impacted by AFFF sources whereas samples that contain high abundances of PFCAs are often inferred to be impacted by municipal landfill leachate discharges.[39,40]

The central hypothesis of this study is that the PFAS composition of various source types is not random but is rather structured by the types of PFASs used for certain industrial or commercial applications and the biological, chemical, and physical processes associated with each source type. If this hypothesis is true, we predict that all source types are probabilistically distinguishable using some subset of PFASs and that a self-consistent and robust characterization of diverse PFASs in samples from multiple source types can be integrated with multivariate statistics and machine learning classifiers to define PFAS fingerprints that are unique to specific source types. To test this hypothesis, 92 samples were collected from six different source types including AFFF-impacted groundwater, landfill leachate, biosolids leachate, municipal WWTP effluent, and wastewater effluent from the pulp and paper and power generation industries. High-resolution mass spectrometry operated with electrospray ionization in negative mode was used to quantify up to 50 target PFASs, and the US National Institute of Standards and Technology (NIST) suspect PFAS list was used to screen and semi-quantify up to 2,266 suspect PFASs in each sample.[41] We included these 50 target PFASs to ensure that the study included a large proportion of PFASs for which authentic standards are commercially available as well as broad suspect screening to supplement fingerprint identification in cases where target PFASs alone might not be sufficient. The resulting data were used to address three research objectives: (1) evaluate whether the six source types exhibit characteristic PFAS fingerprints, (2) identify the specific PFASs that are most diagnostic of each source type, and (3) determine whether target PFASs alone are sufficient to define PFAS fingerprints or if the addition of suspect PFASs is needed to define a PFAS fingerprint for one or more source types. Our results provide the first comprehensive PFAS fingerprinting for multiple source types and represent a new foundation for PFAS source allocation.

MATERIALS AND METHODS

Reagents. Water (HPLC grade) (>99%, high purity, Burdick and Jackson brand), hydrochloric acid (BDH Chemicals), and ammonium acetate (regent grade, Macrom Chemicals) were purchased from VWR (Radnor, PA). Ethyl acetate (99.9%, reagent grade) and 2,2,2-trifluoroethanol (99%, Fluka Analytical) were purchased from Sigma-Aldrich (St. Louis, MO). Methanol (>99%, LC/MS grade) was purchased from Fisher Scientific (Hampton, NH). Sodium chloride was purchased from Mallinckrodt Chemical (>99%). Authentic standards for 50 target PFASs and 28 isotope-labeled surrogates and two additional stable isotope-labeled internal standards (M2PFOA and M8PFOS) were purchased from Wellington Laboratories (Guelph, ON, Canada). See Table S1 of the Supporting Information for list of target PFASs, their acronyms, and assigned surrogates.

Sample Collection. Ninety-two samples were collected from six different source types including 15 archived samples of AFFF-impacted groundwater (AFFF-GW) collected at the site of former firefighter training areas on 15 separate military bases, 19 archived samples of landfill leachate (LL), 15 samples of laboratory-derived leachate from biosolids-amended soils (BL), 19 samples of municipal WWTP effluent (WWTP), 16 samples of wastewater effluent from the pulp and paper industry (PP), and 8 samples of wastewater effluent from the power generation industry (PG). All sites were located throughout the United States, and samples were collected directly at the source. None of the sources are from the same connected hydrologic unit. Samples were collected to represent distinct source types, although it is possible that some sources may be minor contributors to other sources. For example, landfills may receive municipal biosolids and municipal WWTPs may receive landfill leachate. Details on sample dates, sampling, and handling procedures are provided in text in the Supporting Information and Table S2.

Sample Preparation. Samples with lower ionic strength and organic matter (i.e., those from AFFF-GW, WWTP, PP, and PG sources) were extracted using liquid−liquid microextraction as previously described[2] and detailed in the Supporting Information. Samples with higher ionic strength and organic matter (i.e., those from LL and BL sources) were treated separately with a separate method to minimize the extraction of organic matter as previously described[8] and detailed in the Supporting Information.

Liquid Chromatography Quadrupole Time-of-Flight Mass Spectrometry. Chromatographic separations were achieved using an Agilent 1260 HPLC (Santa Clara, CA). Aliquots of 100 μL were injected onto a Zorbax Eclipse XDB-C8 (Agilent, 4.6 × 20 mm, 3.5 μm) guard column fitted with a Zorbax Eclipse Plus analytical column (Agilent, 4.6 × 75 mm, 3.5 μm).[2] The aqueous mobile phase (A) was 20 mM ammonium acetate (Fisher Scientific) in 3% v/v HPLC-grade methanol in HPLC-grade water, and the organic mobile phase (B) was HPLC-grade methanol. A SCIEX X500R QTOF system (Framingham, MA) was operated in negative electrospray ionization (ESI) mode. Data was collected using SWATH data-independent acquisition for both TOF-MS and MS/MS modes (except for PFBA and MPFBA, which were analyzed in MRM HR mode to reduce background). Details on the acquisition parameters, calibration curves, use of third-party reference standards, and continuing calibration standards can be found in the Supporting Information.

Suspect Screening and Semi-Quantification. Candidate suspect PFASs were required to have at least 100 area counts and be at least three times that of the average of three field blanks for all source matrices except BL, which only had a process blank since they were generated under laboratory conditions. The NIST "Suspect List of Possible Per- and Polyfluoroalkyl Substances (PFAS)" version 1.511 was used for identifying suspects. After removing suspect PFASs with a mass

https://doi.org/10.1021/acs.est.3c03770
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

greater than 1250 Da, duplicates, and target PFASs, the list was further sorted using an RDKit function to filter for *NumHDonors* for PFASs detected in ESI$^-$ mode. The 2,266 imported molecular formula were screened as $[M-H]^-$. Mass spectral features were considered matches when associated with a compound on the NIST list[41] with <5 ppm mass error, <10% isotope ratio difference, and >70% spectral library match based on the SCIEX OS purity algorithm (or a higher library score with only the precursor ion matching upon visual inspection). Because MS/MS spectral matching gives higher confidence in identification, library matches were allowed greater tolerance for mass error and isotope ratio difference. Thus, mass spectral features were considered library matches when associated with a compound on the NIST list with <10 ppm mass error, <20% isotope ratio difference, and >70% spectral library match, as well as visual confirmation of at least one matching fragment. Details on assigning confidence to suspect matching is provided in the Supporting Information.[42] Suspect concentrations were estimated using an average PFAS calibration curve constructed from the area of target PFASs divided by the average area of the 22 surrogate areas versus target concentration in units of nmoles/L.[43] Suspect concentrations were reported in ng/L after converting from nmoles/L using the suspect molar mass. The LOQ for suspects was 5 ng/L, the minimum LOQ for target PFASs. In this manner, 1:1 matching was avoided given the number of suspects detected and treated by a single common calibration curve.[43]

Dimensional Reduction and Clustering Analyses. Principal component analysis (PCA) and hierarchical clustering analysis (HCA) were performed to determine whether the six source types exhibit characteristic PFAS fingerprints. The ranked u-score method was applied to normalize the concentration data and to address values reported as <LOQ and <LOD as is recommended for implementation of censored environmental data in multivariate association testing.[44] PCA and HCA were applied with the ranked u-score data using the *FactoMineR* and *hclust* and *pheatmap* packages, respectively, of the R working environment (R version 4.1.0) in R Studio (R Studio version 1.4.1717). For HCA, the agglomerative hierarchical clustering technique was applied using the Euclidean distance metric and the average linkage method to produce two-way HCA dendrograms coupled with a relative concentration heatmap.

Machine Learning Classifiers. Linear support vector classification (SVC), logistic regression (LR), and random forest classification (RF) were performed to identify specific PFASs that are diagnostic of each source type. These classifiers were selected because they provide weighted coefficients that can be used to define feature importance for the classification. Concentration data for target PFASs and semi-concentration data for suspect PFASs were used for classification because classification takes advantage of differences in magnitude between the different variables and normalization obstructs subtle differences between the different variables. Values reported as <LOQ and <LOD were substituted with one-half of the LOQ or 0, respectively, as has been previously described.[45−48] A parallel analysis performed following a log transformation of the concentration data resulted in nearly identical classification performance; therefore, we report the results of our classifiers for the untransformed dataset.

All three classifiers were run on Jupyter Notebook (v6.4.6) using Python (v3.10.2) and were run in a one-versus-all classification format (i.e., the classifiers were run to best differentiate samples of a source type of interest from samples of all other source types). SVC was run using the *sklearn.svm* function, RF was run using the *RandomForestClassifier* function from *sklearn.ensemble* package, and LR was run using the *LogisticRegression* function from the *sklearn.linear_model* package in scikit-learn.[49] For each classifier, the dataset was split into training, validation, and testing sets in a 70−20−10 ratio as recommended for hyperparameter tuning for each of the three classifiers.[50−52] Relevant hyperparameters were tuned using *Grid Search* coupled with stratified k-fold cross-validation with 10 folds and 1000 repeats, which maintains the proportion of samples of the source of interest from the original dataset while defining the training, validation, and test set of each fold. *Grid Search* evaluates every possible combination of relevant hyperparameters for each classifier on the training set to select the combination that gives the highest balanced accuracy. The relevant hyperparameters for each classifier were the SVC regularization parameter (*C*) for SVC; the solver (*newton-cg*, *lbfgs*, *liblinear*), penalty term (*l1*, *l2*), and *C* value for LR; and the number of estimators, *n_estimators* (10, 100, 1000), and maximum number of features, *max_features* (*sqrt*, *log2*), for RF. Including additional hyperparameters for each classifier did not increase performance meaningfully, and more information on hyperparameters and associated value selection is included in the Supporting Information. Stratified k-fold cross-validation was then used to apply the tuned classifiers to the validation and testing sets to check for overfitting. Our tuned classifiers were run with the combination of hyperparameters that yielded the best performance on the training, validation, and testing sets that minimized overfitting.

Once hyperparameters were selected, the tuned classifiers were run over the entire dataset. Classifier performance for each source type for all three classifiers was evaluated and visualized by means of confusion matrices and measured balanced accuracy. We used the population size of each source class to define a balanced accuracy threshold (one false positive and one false negative allowed on average across 100 iterations) that was used to identify well-performing classifiers. Because each source type had a different number of samples, this threshold definition leads to different balanced accuracy thresholds for each source type. The most diagnostic PFASs for each source type were defined for well-performing classifiers as those PFASs, which had the highest positive coefficient weights for SVC and LR and feature importance for RF.

Because of the large number of target and suspect PFASs included in our high-dimensional dataset, we used recursive feature elimination (RFE) on each well-performing classifier to overcome the curse of dimensionality. RFE is an iterative method that can be used to determine the minimum number of PFASs that must be included for successful classification (i.e., highest balanced accuracy) of a sample of a particular source type. RFE was run using the *RFE* function from the *sklearn.feature_selection* package in Python and incorporated the same workflow as described in the preceding paragraph for hyperparameter optimization and cross-validation.

RESULTS AND DISCUSSION

Target and Suspect Screening. To harmonize the approach to target quantification, surrogates of target PFASs that were significantly suppressed (>20%) by high concen-

https://doi.org/10.1021/acs.est.3c03770
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

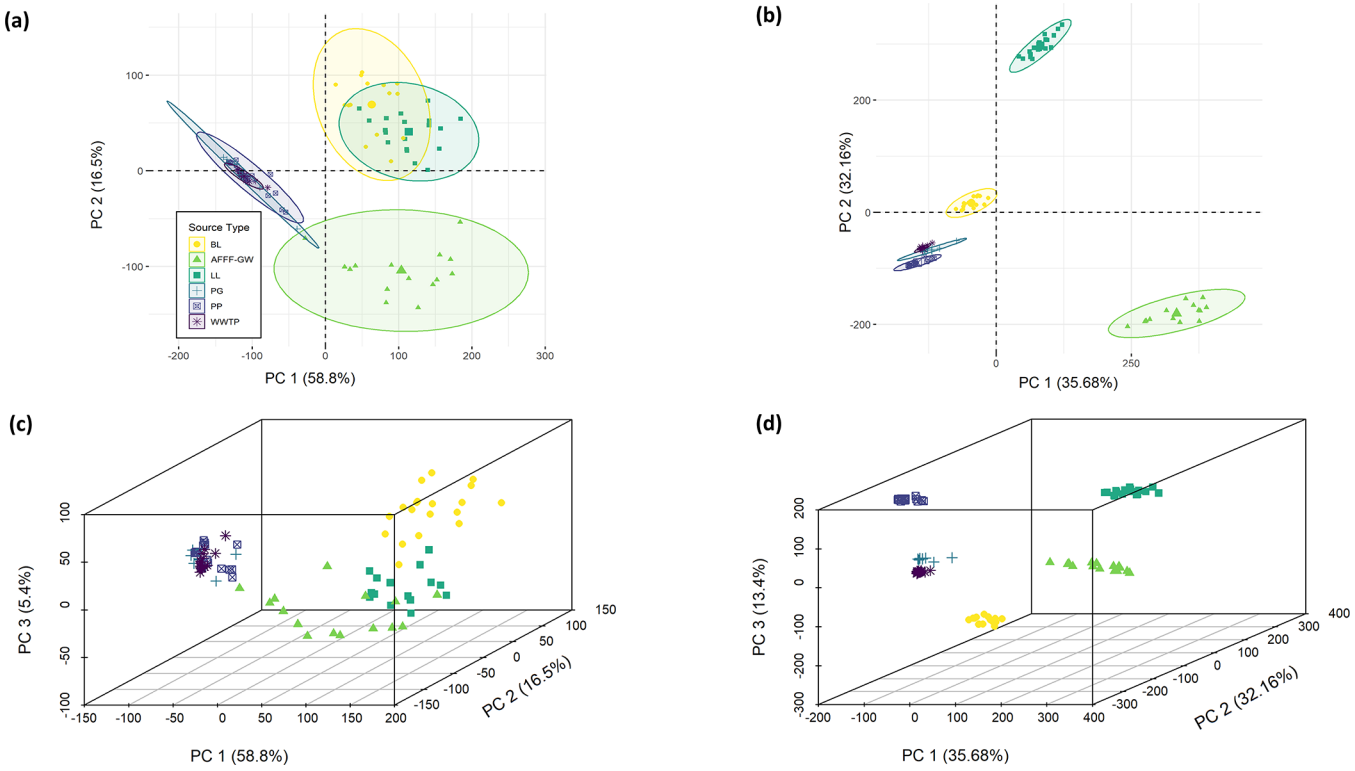


Figure 1. (a) PCA score plot along PC1 and PC2 among samples in the target dataset (n = 92 samples; n = 34 PFAS that were detected in at least one sample). The different source types are identified by the different symbols. The center point of each source type is identified by a larger datapoint than the surrounding points. The ellipses show the 95% confidence interval around the mean of each source type. The plot shows three major clusters: AFFF-GW, LL and BL, and the three WWTPs. (b) PCA score plot along PC1 and PC2 among samples in the target + suspect dataset (n = 92 samples; n = 222 PFASs). The plot shows six distinct clusters for each source type, (c) 3D-PCA score plot among samples in the target dataset. There is greater separation between BL and LL samples along PC3, (d) 3D-PCA score plot among samples in the target + suspect dataset. There is separation between PP, PG, WWTP, and BL samples along PC3.

trations of target PFASs in some sources (e.g., AFFF-GW and LL) were not used for any targets other than their matched target in all sources (Table S1). Thus, a constant quantification strategy for targets was applied across samples from all source types. Details on method accuracy and precision for all sample types are provided in the companion text in the Supporting Information and in Tables S3−S6.

Thirty-four of the 50 target PFASs were quantified above the LOQ in at least one of the 92 samples. A total of 188 suspect PFASs from the NIST list were identified with a confidence of Level 4 or higher[42] and semi-quantified[43] in at least one of the 92 samples. PFAS occurrence and concentration data from AFFF-GW, LL, BL, and WWTP samples have been previously reported from sites around the world. We therefore provide key details on the sample analysis from these source types here and refer the reader to the Supporting Information for a more complete discussion. A summary of the homologue range, highest frequency of detection, and concentration range for selected target and suspect PFASs is provided in Table S7, and the complete data set is provided as Table S12 and is appended to the back of the Supporting Information document. For suspect PFASs identified with a confidence of Level 3 or Level 4 that had alternate matches in the NIST list, we provide the structures of the alternate structural assignments in Table S12.

The AFFF-GW samples contained the greatest number of target and suspect PFASs at the highest concentrations (Table S12). All 15 AFFF-GW samples contained measurable concentrations of target PFASs. The total number of target and suspect PFASs in the AFFF-GW samples ranged from 19 to 77 depending on the source of the AFFF-GW sample. Although many suspect PFASs were detected, most were identified with a confidence of Level 4 and confirmation of the PFAS structures provided in Table S12 was outside the scope of this study.

The major classes of PFASs observed in landfill leachate (LL) samples were similar to those identified in previous studies including PFCAs (C4−C10), PFSAs (C3−C8), n:3 and n:2 saturated fluorotelomer acids, and *N*-methyl and *N*-ethyl perfluoroalkyl sulfonamides (Table S12).[7−12,53,54] Suspect PFASs in LL samples included homologues within classes containing targets (Level 2), four homologous series that contained two to three suspect PFASs that share a common residual (e.g., as defined by the NIST list as "the residual mass after removing the CF2 repeating units"),[41] and many single Level 4 suspect PFASs not in homologous series (e.g., do not share a common residual). These suspect classes included substituted and unsubstituted perfluoroalkyl sulfonamides, perfluorosulfonamido acetic acids, and perfluoroalkyl sulfinates, which have been previously reported in suspect screening[11,55] and extended target screening[8] of LL samples.

Across the biosolids-amended soil leachate (BL) samples, 22 of 50 target PFASs were quantified (Table S12). Every BL sample contained at least three target PFASs, with a maximum of 14 target PFASs present in one BL sample. Both perfluorooctanoic acid (PFOA) and perfluorooctanesulfonic acid (PFOS) were found in 100% of the BL samples. The six target PFAS classes detected included PFCAs, PFSAs, perfluoroalkyl sulfonamides, perfluorosulfonamido acetic acid

https://doi.org/10.1021/acs.est.3c03770
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

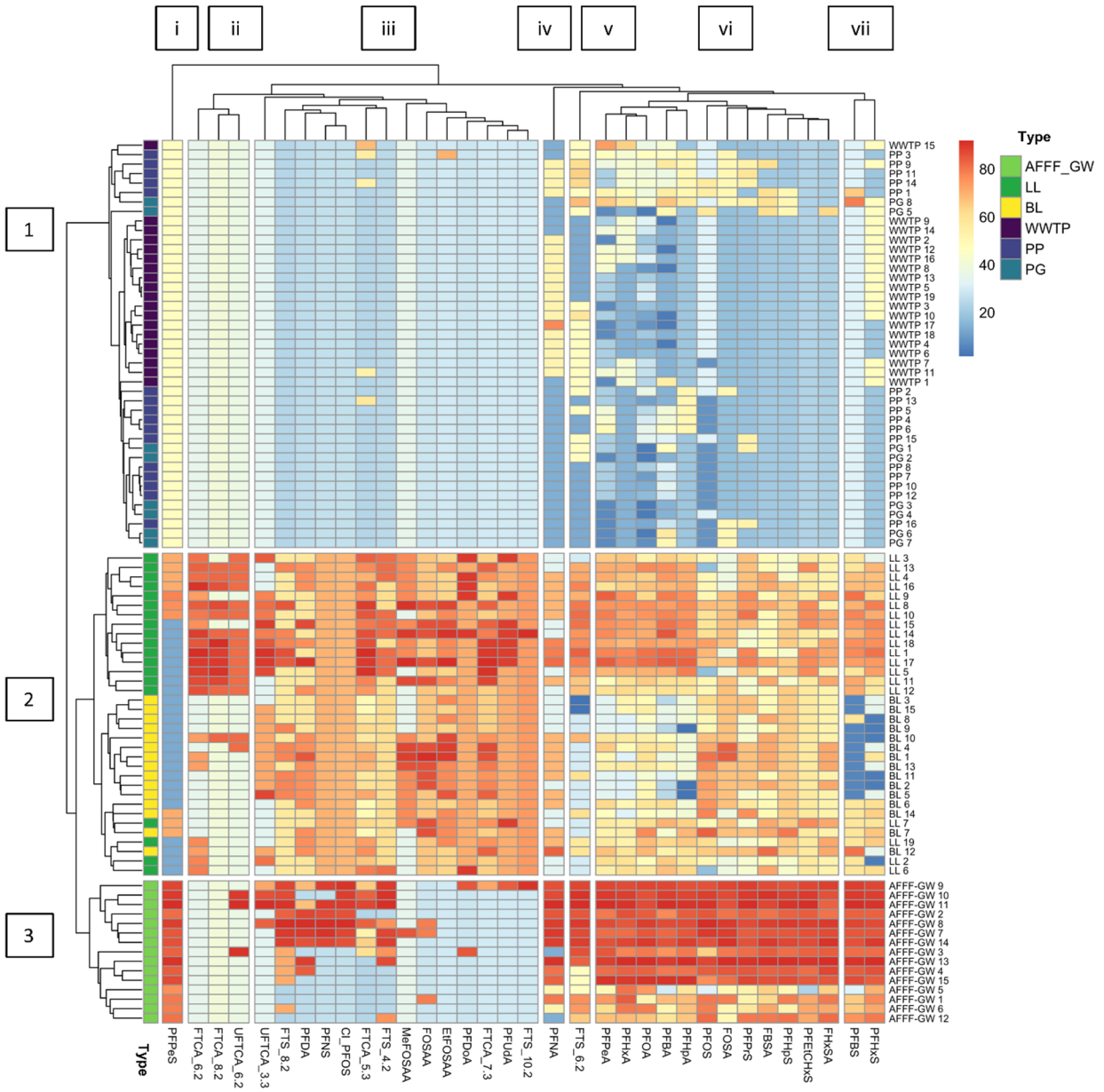


Figure 2. Two-way HCA dendrogram with heatmap. The top dendrogram shows PFAS groupings, and the left dendrogram shows sample groupings. The heatmap is color-coded based on ranked u-scores of the different samples for each PFAS. The dendrogram shows three major clusters along the left axis: AFFF-GW, LL and BL, and the three WWTPs, showing agreement with the PCA score plot. Note that the n:3 and n:2 fluorotelomers have their acronyms revered (e.g., FTCA_6:2 instead of 6:2 FTCA) because the package used to create the dendrogram requires labels to begin with a letter.

precursors, 6:2 fluorotelomer sulfonate (6:2 FTS), and both saturated and unsaturated fluorotelomer carboxylates. The PFCAs and PFSAs in BL samples were consistent with previous findings for biosolids-amended soil leachate[12] and water bodies impacted by nearby biosolids application.[56]

Among the 19 WWTP samples, only 10 target PFASs were observed with concentrations ranging from 8.0 to 1200 ng/L (Table S12). The quantified target PFASs included PFCAs, PFSAs, 5:3 FTCA, and 6:2 FTS. Further, all of the PFASs observed in WWTP samples were also observed in AFFF-GW samples but at lower concentrations in the WWTP samples.[13−16,18,19,57,58] Both 5:3 FTCA[59,60] and 6:2 FTS[15,18,61−63] have been previously reported in WWTP effluent.

To the best of our knowledge, this study marks the first report of PFASs in power generation (PG) effluent. This study also adds additional context to the limited investigation of PFASs in pulp and paper (PP) effluent.[20,64] A total of 10 target PFASs from four classes were detected in the PG or PP samples, including five PFCAs, three PFSAs, one FTS, and one FTCA (Table S12). Detection frequencies of PFASs and concentrations were generally higher in the PP effluents. Of the 10 PFASs, four PFCAs (C4−C6, C8), PFOS, and 6:2 FTS were common between both sample groups and PFOA was the most frequently detected PFAS (present in 100% of samples). 5:3 FTCA is reported for the first time in PP effluents. The suspect PFASs identified in PP and PG samples were identified with a confidence of Level 4, and confirmation of the PFAS

https://doi.org/10.1021/acs.est.3c03770
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

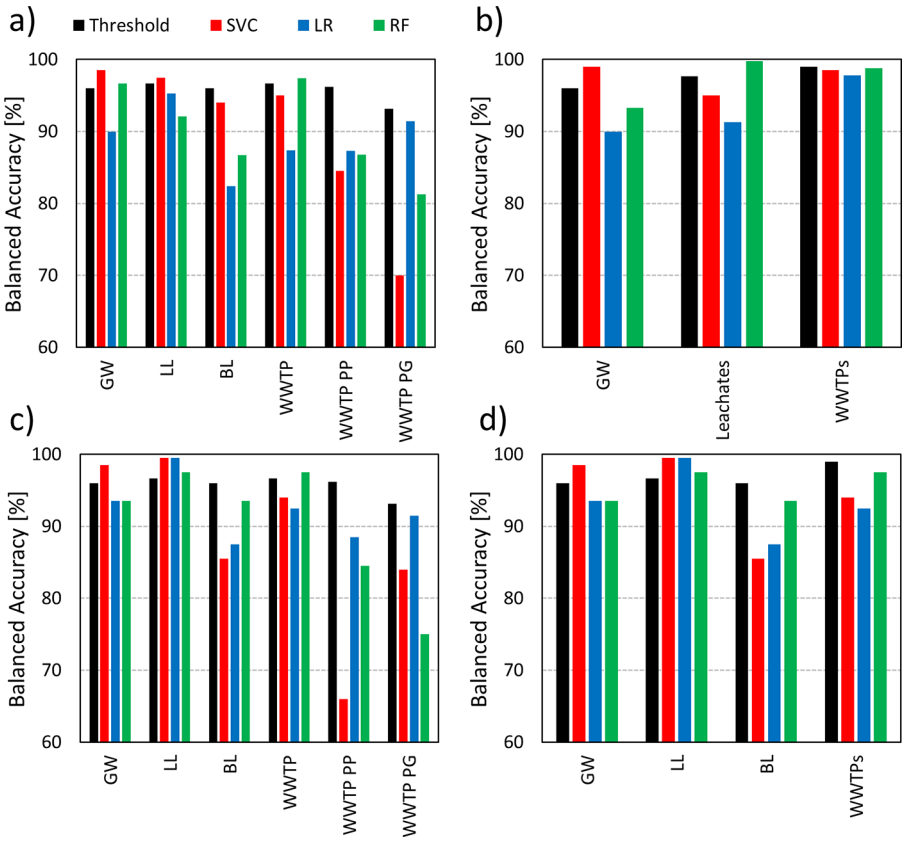


Figure 3. Performance of the tuned and cross-validated classification models for the target dataset for (a) case 1 and (b) case 2 and for the target + suspect dataset for (c) case 3 and (d) case 4. SVC = support vector classification, LR = logistic regression, RF = random forest. The threshold is defined as the balanced accuracy that results in one false negative and one false positive misclassification on average across the cross-validation.

structures provided in Table S12 was outside the scope of this study.

Dimensional Reduction and Clustering Analyses: Target PFASs. PCA and HCA were used to determine whether the six source types exhibit characteristic PFAS fingerprints. These analyses initially focused on only the target PFASs to identify characteristic fingerprints among the most commonly measured PFASs that could be quantified with high accuracy and sensitivity. The PCA score plot is provided as Figure 1a, and the 3D-PCA score plot is provided as Figure 1c. The PCA was primarily used for descriptive purposes so only the first three PCs (which contained 80.7% of the variation) were retained, and scree plot analysis shown in Figure S1a further supports the retention of only the first three PCs. Samples from the six source types clustered into three major clusters consisting of AFFF-GW samples, LL and BL samples, and all wastewater samples (Figure 1a). The clustering of samples suggests that the LL and BL and WWTP, PP, and PG samples contain similar types and relative abundances of target PFASs. The 3D-PCA score plot shows that there is separation between BL and LL samples along the PC3 direction, indicating that the PFASs that have high loadings in that PC must be driving the separation of these source types. Examination of the PCA loadings suggest that a group of PFASs dominated by PFCAs and PFSAs drive the separation of the GW samples and that a group of PFASs dominated by n:3 and n:2 saturated and unsaturated fluorotelomer acids and *N*-methyl and *N*-ethyl perfluorooctane sulfonamides drive the separation of the LL and BL samples. The separation of BL and LL samples along PC3 is driven by C10−C12 PFCAs and fluorotelomer sulfonates (BL) and n:2 saturated and unsaturated fluorotelomer acids (LL). No target PFASs have significant PCA loadings in the direction of the WWTP samples, and this is expected because only few PFASs at relatively low concentrations were measured in the WWTP samples. Together, the PCA analysis demonstrates that characteristic PFAS fingerprints exist for at least three distinct groups of the six source types.

The HCA dendrogram and heatmap are provided as Figure 2. The three clusters along the left side (labeled 1−3) describe relationships among the sources, the seven clusters along the top (labeled i−vii) describe relationships among the 34 target PFASs that were detected in at least one sample, and the heatmap colors describe the relative concentrations of the PFASs across the samples. In examining the labels along the left side of Figure 2, it is noted that the sources cluster together similar to the way they clustered in the PCA score plot (Figure 1a). All of the AFFF-GW samples are contained in cluster 3. The LL and BL samples cluster together, with most of the LL samples at the top of cluster 2 (the exceptions are LL2, LL6, LL7, and LL19) and all of the BL samples at the bottom of cluster 2. This supports the 3D-PCA score plot results that indicate that BL and LL are similar but do contain differences in the inherent composition of their PFASs driven by the PFASs that have high loadings along PC3. The WWTP samples cluster together in cluster 1, and there is some separation into subclusters based on the type of WWTP, but samples from all types of WWTPs are spread throughout cluster 1. This more refined observation from the HCA supports the clustering of the WWTPs in the PCA score plot

https://doi.org/10.1021/acs.est.3c03770
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

(Figure 1a). A broad view of the heatmap reveals that there are low relative concentrations of most PFASs in the WWTP samples, moderate relative concentrations of most PFASs in the LL and BL samples, and high relative concentrations of about half of the PFASs in the AFFF-GW samples.

In examining the labels along the bottom of Figure 2, relationships among the target PFASs can also be identified. First, clusters i (PFPeS), iv (PFNA), and v (6:2 FTS) each contain only a single PFAS, reflecting distinctive relative abundance patterns for these PFASs among the samples. PFPeS was measured in all AFFF-GW samples and sporadically in LL and BL samples. However, the different LOQs for PFPeS in the WWTP samples and the LL and BL samples result in different ranked u-scores for censored data in those source types leading to the unique clustering. PFNA and 6:2 FTS are both characterized as exhibiting high relative concentrations in AFFF-GW samples, lower relative concentrations in LL samples, and sporadic and lower relative concentrations in WWTP and PP samples. These unique relative abundance patterns drive their clustering in the HCA dendrogram. The remaining clusters are more informative with respect to assessing PFAS fingerprints. For example, clusters vi and vii contain twelve and two PFASs, respectively, that are present at high relative concentrations in the AFFF-GW samples. These clusters include the C3−C4 and C6−C8 PFSAs, the C4−C8 PFCAs, the C4, C6, and C8 perfluoroalkyl sulfonamides, and perfluoroethylcyclohexane sulfonate (PFEtCHxS). Cluster ii contains three PFASs that are present at high relative concentrations in the LL samples and includes 6:2 FTCA, 8:2 FTCA, and 8:2 UFTCA. Finally, cluster iii contains 14 PFASs that are present at high relative concentrations in the LL and BL samples and includes the C9 PFSA, the C10−C12 PFCAs, 4:2, 8:2, and 10:2 FTSs, along with the other FTCAs, UFTCAs, and perfluoroalkyl sulfonamide acetic acids. These are also the PFASs with high loadings along PC3 that help separate BL from LL samples.

The HCA provides additional support that characteristic PFAS fingerprints exist for at least three distinct groups of the six source types. The HCA also provides additional insights on the PFASs within those PFAS fingerprints, which agree with the loadings from the PCA. Further, the clustering derived from these unsupervised techniques aligns with the source types, supporting the hypothesis that different PFAS sources have probabilistically distinguishable fingerprints.

Machine Learning Classifiers: Target PFASs. Machine learning classifiers were used to identify the specific target PFASs whose presence are most diagnostic of each source type. We reasoned that the weighted coefficients or feature importance from well-performing SVC, LR, and RF classifiers would allow us to identify the target PFASs that are most diagnostic of each source type. Because the PCA and HCA analyses revealed that the six source types separate into three major clusters, we evaluated the performance of each classifier for two cases. In case 1, we considered all six source classes (GW, LL, BL, WWTP, PP, and PG), and in case 2, we used the results of the PCA and HCA to define three source classes as AFFF-GW, leachates (LL and BL samples), and WWTPs (WWTP, PP, and PG samples). There was no evidence of overfitting of any classifier as the balanced accuracy ratio of the training set was not significantly higher than that of the testing set during cross-validation (Tables S8 and S9). The performance of the tuned and cross-validated classifiers for case 1 is described in the confusion matrices provided as Figures S2−S4. The results of the performance analyses for case 1 and case 2 are presented in Figure 3a,b, respectively, with the balanced accuracy thresholds to identify well-performing classifiers provided as black bars. The data in Figure 3a show that the balanced accuracies for all of the classifiers in case 1 (six source classes) ranged between 70.0 and 98.5%. The well-performing classifiers for case 1 include SVC for AFFF-GW and LL samples and RF for AFFF-GW and WWTP samples. The data in Figure 3b show that the balanced accuracies for all of the classifiers in case 2 (three source classes) ranged between 90.0 and 99.8%. The well-performing classifiers for case 2 include SVC for AFFF-GW samples and RF for leachates and WWTPs samples.

The weighted coefficients or feature importance for each of the target PFASs in each of the well-performing classifiers were used to identify target PFASs that are most diagnostic of each source class. There are five PFASs in common among the top seven PFASs with the highest weighted coefficients or feature importance in the well-performing SVC and RF classifiers for AFFF-GW samples. These include PFHxA, FHxSA, PFPeS, PFHxS, and FBSA. All five of these PFASs were also contained in clusters vi and vii of the HCA dendrogram (Figure 2), which included PFASs that exhibited high relative concentrations in AFFF-GW samples. The four PFASs with the top weighted coefficients from the SVC classifier for LL samples include 6:2 FTCA, 5:3 FTCA, PFHpA, and PFBA. The FTCAs were contained in clusters ii and iii of the HCA dendrogram (Figure 2), which included PFASs that exhibited high relative concentrations in LL samples, but PFHpA and PFBA were not, which makes their inclusion here somewhat unexpected. It is worth noting that coefficient weights of PFHpA and PFBA were also high in LR and RF (not well-performing classifiers in this case) so their importance as diagnostic PFASs for classification of LL samples seems to be robust. Finally, PFBA, PFOA, PFNA, and PFOS had the largest RF feature importance for WWTP samples. These ubiquitous PFCAs and PFSA are present in samples from nearly all source classes, and their selection as diagnostic PFASs for WWTP samples in particular is likewise unexpected. In examining the HCA in Figure 2, PFNA seems to have the most potential to be uniquely present in WWTP samples, but the presence of all four of these PFASs simultaneously may indeed be diagnostic of the WWTP source class.

There were no well-performing classifiers for BL, PP, and PG samples in case 1, which is not surprising considering the results of the PCA and HCA (Figures 1a and 2). Therefore, we considered case 2 in an effort to identify diagnostic PFASs for the combined source classes. The unique well-performing classifiers for case 2 include RF for leachates and WWTPs samples. Six of the nine PFASs with the top feature importance from the RF classifier for leachate samples were 3:3 FTCA, 5:3 FTCA, 6:2 FTCA, 7:3 FTCA, EtFOSAA, and MeFOSAA. This result suggests that FTCAs and perfluoroalkyl sulfonamidoacetic acids are diagnostic of LL and BL samples together. Finally, PFOA and PFOS have the highest feature importance in the well-performing RF classifier for the WWTPs source class. Whereas PFBA and PFNA were identified as diagnostic of the WWTP source class in case 1, these two PFASs are not among the 10 PFASs with the highest weighted coefficients for the WWTPs source class for case 2. The data from case 1 and case 2 suggest that the presence of these specific PFCAs and PFSAs (along with the absence of PFASs diagnostic of the

https://doi.org/10.1021/acs.est.3c03770
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

AFFF-GW and LL source classes) is diagnostic of the WWTPs source class.

Evaluation of Target + Suspect PFASs Together. The preceding PCA and machine learning classifier analyses were repeated with the combined target + suspect PFASs to determine whether suspect PFASs facilitate the discovery of PFAS fingerprints for one or more source types. This is particularly relevant for the source types that contain relatively few target PFASs at low relative concentrations (e.g., WWTP, PP, PG). For this analysis, we used the concentrations of the 34 target PFASs and the semi-quantified concentrations of 188 suspect PFASs. The PCA score plot for each source type against PC1 and PC2 is provided as Figure 1b, and the 3D-PCA score plot is provided as Figure 1d. The six source types separate more clearly because most of the suspect PFASs were identified in only one source type, making that source type more compositionally distinct. However, the WWTP, PP, and PG samples still cluster closely with each other and the BL samples cluster closer to the WWTP, PP, and PG samples than the LL samples (as previously observed in Figure 1a).

The results of the PCA score plot indicate that the combined dataset is suitable for undergoing classification. Because the PCA analysis can be interpreted to conclude that each of the six source types separates into distinct clusters or that the WWTP, PP, and PG samples are close enough to be considered a single cluster, we again evaluated the performance of each classifier for two cases. In case 3, we considered all six source classes (AFFF-GW, LL, BL, WWTP, PP, and PG), and in case 4, we defined four source classes as AFFF-GW, LL, BL, and WWTPs (WWTP, PP, and PG samples combined). There was no evidence of overfitting of any classifier through the stratified cross-validation results (Tables S10 and S11). The performance of the tuned and cross-validated classifiers for case 3 is described in the confusion matrices provided as Figures S5−S7. The data in Figure 3c show that the balanced accuracies for all of the classifiers in case 3 (six source classes) ranged between 66.0 and 99.5%. The well-performing classifiers for case 3 include SVC for AFFF-GW and LL samples, LR for LL samples, and RF for LL and WWTP samples. The balanced accuracies for all of the classifiers in case 4 (four source classes) ranged between 85.5 and 99.5% (Figure 3d). The well-performing classifiers for case 4 include SVC for AFFF-GW and LL samples and LR and RF for LL samples. There were no well-performing classifiers for the BL or WWTPs source classes.

The six PFASs with the highest weighted coefficients in the well-performing SVC classifier for AFFF-GW samples include PFHxA, FHxSA, trihydrogen-substituted fluoroethernonanoic acid (3H-PFENA, NIST ID 4072), PFPeS, PFHxS, and FBSA. Five of these PFASs were also selected as diagnostic of AFFF-GW samples from the target dataset and one (3H-PFENA) is a Level 4 suspect PFAS. This result suggests that a few target PFASs may be sufficient to classify AFFF-GW samples in a one-versus-all classification. For the LL samples, 5:3 FTCA, 6:2 FTCA, 4:2 FTThA (NIST ID 3393, Level 4), MeFPrSAA (NIST ID 3343, Level 3d), and MeFBSAA (NIST ID 3344, Level 3d) were among the PFASs with the highest weighted coefficients or importance score for all three well-performing classifiers. The FTCAs are target PFASs and were selected in the previous classification, but the others are suspect PFASs and are clearly diagnostic of LL samples given their higher weighted coefficients or importance score compared to those of the target PFASs. The WWTP source class had one suspect PFAS that was statistically more diagnostic than target PFASs for classifying samples: the organofluorine-containing and broad-use insecticide fipronil (NIST ID 4820, Level 2b). It is not surprising that fipronil is measured in WWTP samples, as it has been reported as a micropollutant in wastewater effluents around the world.[65−67] Like the target analysis, PFBA, PFOA, and PFNA were also identified as diagnostic of WWTP samples, but their importance scores were lower than that of fipronil. There were no well-performing classifiers for the WWTPs source class that included WWTP, PP, and PG samples (case 4), but the RF classifier had a balanced accuracy of 97.5% and fipronil was again selected as the suspect PFAS with the highest importance score followed by PFBA, PFOA, and PFNA. The analysis of the data from case 4 further suggests that WWTP, PP, and PG samples can be best classified as a combined source class.

Recursive Feature Elimination (RFE). Finally, RFE analysis was used to determine the minimum number of PFASs that must be included for successful classification (i.e., the highest balanced accuracy) of a sample of a particular source type. This analysis was done to inform the development of a PFAS analytical method that could be used for source allocation of certain source types (Table 1). The RFE analysis demonstrates that classification of AFFF-GW can be achieved with 99.1% balanced accuracy when only including PFHxA, FHxSA, 3H-PFENA, PFPeS, and PFHxS (SVC classifier). If we only consider target PFASs (because 3H-PFENA is a Level 4 suspect PFAS), classification of AFFF-GW can be achieved with 98.9% balanced accuracy when only including PFPeS, PFHxA, PFHpS, and FHxSA (RF classifier). Classification of LL can be achieved with 99.7% balanced accuracy when only including 4:2 FTThA, MeFPrSAA, 6:2 FTCA, and 5:3 FTCA (LR classifier). No well-performing classifiers were identified for BL samples, but classification of the combined leachates class can be achieved with 95.1% balanced accuracy when including 4:2 FTThA, MeFPrSAA, MeFBSAA, 6:2 FTCA, 3:3 FTCA, and 5:3 FTCA along with MeFOSAA and EtFOSAA (RF classifier). Classification of the WWTP and the combined WWTPs classes can be achieved with 98.5 and 97.8% balanced accuracy when only including fipronil along with six and eight relatively ubiquitous perfluoroalkyl acids, respectively (RF classifier). The RFE results demonstrate that each source type can be accurately classified by each well-performing classifier using some combinations of relatively few target and suspect PFASs. This shows that a future monitoring study would only need to measure the PFASs identified in Table 1 to gain insight on the potential PFAS sources. We also observed that for the WWTP source type, the predictors with the most negative coefficient weights or feature importance were those that were picked as important predictors for the AFFF-GW and LL source types. This further shows that while indeed source allocation of a given sample must contain detections of the PFASs defined in Table 1 for a given source type, in some cases it must also not include detections of important predictors for other source types.

Limitations. This study aimed to discover PFAS fingerprints for specific PFAS source types. Although successful in this endeavor, this study is limited in some ways with respect to the analytical methods and application of the machine learning classifiers. First, the dataset included in this study was acquired exclusively from LC-TOF-MS operated with ESI in negative mode. On the one hand, this approach generated a unified dataset in which all samples were prepared and

https://doi.org/10.1021/acs.est.3c03770
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Table 1. Summary of PFASs Identified as Diagnostic of Different Source Classes in the RFE Analysis

Source	Diagnostic PFASs	
	Target PFASs	Suspect PFASs
AFFF-GW (target + suspect PFASs)	PFHxA	3H-PFENA
	FHxSA	
	PFPeS	
	PFHxS	
AFFF-GW (target PFASs only)	PFPeS	
	PFHxA	
	PFHpS	
	FHxSA	
LL (target + suspect PFASs)	6:2 FTCA	4:2 FTThA
	5:3 FTCA	MeFPrSAA
Leachates (LL + BL, target + suspect PFASs)	6:2 FTCA	4:2 FTThA
	3:3 FTCA	MeFPrSAA
	5:3 FTCA	MeFBSAA
	MeFOSAA	
	EtFOSAA	
WWTP (target + suspect PFASs)	PFNA	fipronil
	PFHxS	
	PFPeA	
	PFOA	
	PFBA	
	PFOS	
WWTPs (WWTP + PP + PG, target + suspect PFASs)	PFBA	fipronil
	PFNA	
	PFOA	
	PFHpA	
	PFOS	
	FOSA	
	PFHxS	
	PFPeA	

measured in the same way. However, this approach excludes cationic and zwitterionic PFASs[68−70] and volatile PFASs and the sample preparation and analysis may be biased for more hydrophobic PFASs[71] and biased against ether-based PFASs (due to the acquisition source temperature).[65−67] Second, although we are confident in the semi-quantification approach applied to generate the suspect PFASs, we acknowledge that the semi-quantified suspect data has more uncertainty than the target data. Third, successful classification relies on sufficiently large datasets with a minimum amount of censored data. Our sample set was limited to 92 samples, and the PG source type contained only 8 samples. It is possible that more samples would have facilitated a more robust classification. Our dataset also contained a large number of censored data, and best practices (e.g., u-score normalization) were implemented to address censored data prior to data analysis and classification. Nevertheless, censored data and outliers can bias the results of classification.[72] Additionally, the specific classifiers that we selected assume that the data are linearly separable, which might not always be the case. For instance, SVC with a nonlinear kernel might work better to separate some source types but coefficient weights necessary to discover diagnostic PFASs are only available for the linear kernel. Finally, the classifiers were developed and applied on real environmental water samples derived from six different source types to define characteristic PFAS fingerprints. We acknowledge that there are other potential sources of PFASs to environmental waters that may contribute to the fingerprints defined in this study. The classifiers have not been tested on external environmental water samples from sources other than the six types described in this study.

Environmental Implications. Developing forensics tools will better enable source allocation of PFASs measured in the environment. In this study, we integrate a unified analytical method to characterize target and suspect PFASs in samples from a variety of source types with multivariate statistics and machine learning classifiers to more accurately define PFAS fingerprints that are diagnostic of specific source types (Table 1). Despite the fact that the AFFF-GW samples were collected from sites that varied in space and time, they contain the same diagnostic PFASs and our classifiers could define accurate PFAS fingerprints of both target PFASs and combined target and suspect PFASs. This indicates that these diagnostic PFASs are persistent in AFFF-GW samples over spatially and temporally distributed sites. It is important to note that FHxSA and 6:2 FTCA are target PFASs that were identified as diagnostic of certain sources but are rarely measured and are not included in the EPA Draft Method 1633. Further, authentic standards for the diagnostic suspect PFASs 3H-PFENA, 4:2 FTThA, and MeFPrSAA should be synthesized to confirm their occurrence in environmental samples and included in analytical methods. The discovery of fipronil as diagnostic of WWTP samples demonstrates how substances used more commonly in other sectors (and potentially non-PFASs) can be useful in source allocation of PFASs. Additional work is underway with this dataset to determine if other trace organic chemicals would similarly assist in classification between the different WWTP source types (WWTP, PP, and PG). Finally, the PFAS fingerprints identified in this study represent PFAS fingerprints at these specific sources. Our present study does not address how PFAS fingerprints attenuate as a function of time or distance from a source. Our ongoing work aims to address this question.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.3c03770.

Additional information on sample sources and preparation, LC-QTOF operation, suspect screening, surrogate selection for quantification, method performance, and discussion of data for the various source waters; tables are provided containing information on target and surrogate PFAS; accuracy, precision, LOD, and LOQs; background in field blanks; surrogate recovery; PFASs with the highest weighted coefficients from each case and each classifier; details regarding the balanced accuracies of the testing, training, and validation data for each classifier with the best combination of hyperparameters; figures are provided containing details on confusion matrices from each case and each classifier (PDF)

https://doi.org/10.1021/acs.est.3c03770
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

AUTHOR INFORMATION

Corresponding Authors

Jennifer A. Field − *Department of Environmental and Molecular Toxicology, Oregon State University, Corvallis, Oregon 97331, United States;* orcid.org/0000-0002-9346-4693; Phone: +1 541 737 2265; Email: jennifer.field@oregonstate.edu

Damian E. Helbling − *School of Civil and Environmental Engineering, Cornell University, Ithaca, New York 14853, United States;* orcid.org/0000-0003-2588-145X; Phone: +1 607 255 5146; Email: damian.helbling@cornell.edu

Authors

Nayantara T. Joseph − *School of Civil and Environmental Engineering, Cornell University, Ithaca, New York 14853, United States;* orcid.org/0000-0003-4022-7675

Trever Schwichtenberg − *Chemistry Department, Oregon State University, Corvallis, Oregon 97331, United States*

Dunping Cao − *Chemistry Department, Oregon State University, Corvallis, Oregon 97331, United States*

Gerrad D. Jones − *Department of Biological & Ecological Engineering, Oregon State University, Corvallis, Oregon 97331, United States;* orcid.org/0000-0002-1529-9506

Alix E. Rodowa − *National Institutes of Standards and Technology, Gaithersburg, Maryland 20899, United States*

Morton A. Barlaz − *Department of Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh, North Carolina 27695, United States;* orcid.org/0000-0001-8028-3917

Joseph A. Charbonnet − *Department of Civil, Construction, and Environmental Engineering, Iowa State University, Ames, Iowa 50011, United States; Department of Civil and Environmental Engineering, Colorado School of Mines, Golden, Colorado 80401, United States;* orcid.org/0000-0001-8766-6072

Christopher P. Higgins − *Department of Civil and Environmental Engineering, Colorado School of Mines, Golden, Colorado 80401, United States;* orcid.org/0000-0001-6220-8673

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.3c03770

Author Contributions

#N.T.J. and T.S. contributed equally.

Notes

The authors declare no competing financial interest.

These opinions, recommendations, findings, and conclusions do not necessarily reflect the views or policies of NIST or the United States Government.

ACKNOWLEDGMENTS

Funding for this project was provided by SERDP ER20-1375. We thank Pat Gorski at the Wisconsin Department of Natural Resources for the pulp and paper mill and power generation effluents. We thank Derek Muensterman for helpful discussions. Oregon State University in Corvallis, Oregon, is located within the traditional homelands of the Mary's River or Ampinefu Band of Kalapuya.

REFERENCES

(1) Wang, Z.; DeWitt, J. C.; Higgins, C. P.; Cousins, I. T. A Never-Ending Story of Per- and Polyfluoroalkyl Substances (PFASs)? *Environ. Sci. Technol.* **2017**, *51*, 2508−2518.

(2) Backe, W. J.; Day, T. C.; Field, J. A. Zwitterionic, Cationic, and Anionic Fluorinated Chemicals in Aqueous Film Forming Foam Formulations and Groundwater from U.S. Military Bases by Nonaqueous Large-Volume Injection HPLC-MS/MS. *Environ. Sci. Technol.* **2013**, *47*, 5226−5234.

(3) Liu, M.; Munoz, G.; Duy, S. V.; Sauve, S.; Liu, J. X. Per- and Polyfluoroalkyl Substances in Contaminated Soil and Groundwater at Airports: A Canadian Case Study. *Environ. Sci. Technol.* **2022**, *56*, 885−895.

(4) Sepulvado, J. G.; Blaine, A. C.; Hundal, L. S.; Higgins, C. P. Occurrence and Fate of Perfluorochemicals in Soil Following the Land Application of Municipal Biosolids. *Environ. Sci. Technol.* **2011**, *45*, 8106−8112.

(5) Johnson, G. R. PFAS in Soil and Groundwater Following Historical Land Application of Biosolids. *Water Res.* **2022**, 211.

(6) Pepper, I. L.; Brusseau, M. L.; Prevatt, F. J.; Escobar, B. A. Incidence of Pfas in Soil Following Long-Term Application of Class B Biosolids. *Sci. Total Environ.* **2021**, 793.

(7) Benskin, J. P.; Li, B.; Ikonomou, M. G.; Grace, J. R.; Li, L. Y. Per- and Polyfluoroalkyl Substances in Landfill Leachate: Patterns, Time Trends, and Sources. *Environ. Sci. Technol.* **2012**, *46*, 11532−11540.

(8) Allred, B. M.; Lang, J. R.; Barlaz, M. A.; Field, J. A. Orthogonal Zirconium Diol/C18 Liquid Chromatography-Tandem Mass Spectrometry Analysis of Poly and Perfluoroalkyl Substances in Landfill Leachate. *J Chromatogr A* **2014**, *1359*, 202−211.

(9) Gallen, C.; Drage, D.; Eaglesham, G.; Grant, S.; Bowman, M.; Mueller, J. F. Australia-Wide Assessment of Perfluoroalkyl Substances (PFASs) in Landfill Leachates. *J. Hazard. Mater.* **2017**, *331*, 132−141.

(10) Lang, J. R.; Allred, B. M. K.; Field, J. A.; Levis, J. W.; Barlaz, M. A. National Estimate of Per- and Polyfluoroalkyl Substance (PFAS) Release to U.S. Municipal Landfill Leachate. *Environ. Sci. Technol.* **2017**, *51*, 2197−2205.

(11) Maldonado, V. Y.; Schwichtenberg, T.; Schmokel, C.; Witt, S. E.; Field, J. A. Electrochemical Transformations of Perfluoroalkyl Acid (PFAA)Precursors and PFAAs in Landfill Leachates. *Acs Es&T Water* **2022**, *2*, 624−634.

(12) Huang, X. Y.; Wei, X. X.; Liu, H. Z.; Li, W.; Shi, D. Z.; Qian, S. H.; Sun, W. J.; Yue, D. B.; Wang, X. M. Occurrence of Per- and Polyfluoroalkyl Substances (PFAS) in Municipal Solid Waste Landfill Leachates from Western China. *Environ. Sci. Pollut. Res.* **2022**, *29*, 69588−69598.

(13) Sims, J. L.; Stroski, K. M.; Kim, S.; Killeen, G.; Ehalt, R.; Simcik, M. F.; Brooks, B. W. Global Occurrence and Probabilistic Environmental Health Hazard Assessment of Per- and Polyfluoroalkyl Substances (PFASs) in Groundwater and Surface Waters. *Sci. Total Environ.* **2022**, 816.

(14) Barisci, S.; Suri, R. Occurrence and Removal of Poly/Perfluoroalkyl Substances (PFAS) in Municipal and Industrial Wastewater Treatment Plants. *Water Sci. Technol.* **2021**, *84*, 3442−3468.

(15) Lenka, S. P.; Kah, M.; Padhye, L. P. A Review of the Occurrence, Transformation, and Removal of Poly- and Perfluoroalkyl Substances (PFAS) in Wastewater Treatment Plants. *Water Res.* **2021**, 199.

(16) Franco, M. E.; Burket, S. R.; Sims, J. L.; Lovin, L. M.; Scarlett, K. R.; Stroski, K.; Steenbeek, R.; Ashcroft, C.; Luers, M.; Brooks, B. W.; Lavado, R. Multi-Approach Assessment for the Evaluation of Spatio-Temporal Estrogenicity in Fish from Effluent-Dominated Surface Waters under Low Instream Flow. *Environ. Pollut.* **2020**, 265.

(17) Gallen, C.; Baduel, C.; Lai, F. Y.; Thompson, K.; Thompson, J.; Warne, M.; Mueller, J. F. Spatio-Temporal Assessment of Perfluorinated Compounds in the Brisbane River System, Australia: Impact of a Major Flood Event. *Mar. Pollut. Bull.* **2014**, *85*, 597−605.

(18) Coggan, T. L.; Moodie, D.; Kolobaric, A.; Szabo, D.; Shimeta, J.; Crosbie, N. D.; Lee, E.; Fernandes, M.; Clarke, B. O. An

https://doi.org/10.1021/acs.est.3c03770
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Investigation into Per- and Polyfluoroalkyl Substances (PFAS) in Nineteen Australian Wastewater Treatment Plants (WWTPs). *Heliyon* **2019**, *5*, No. e02316.
(19) Tavasoli, E.; Luek, J. L.; Malley, J. P.; Mouser, P. J. Distribution and Fate of Per- and Polyfluoroalkyl Substances (PFAS) in Wastewater Treatment Facilities. *Environmental Science-Processes & Impacts* **2021**, *23*, 903−913.
(20) Langberg, H. A.; Arp, H. P. H.; Breedveld, G. D.; Slinde, G. A.; Høiseter, Å.; Grønning, H. M.; Jartun, M.; Rundberget, T.; Jenssen, B. M.; Hale, S. E. Paper Product Production Identified as the Main Source of Per- and Polyfluoroalkyl Substances (PFAS) in a Norwegian Lake: Source and Historic Emission Tracking. *Environ. Pollut.* **2021**, *273*, No. 116259.
(21) States Environmental Protection Agency. *United States Environmental Protection Agency Multi-Industry Per-and Polyfluoroalkyl Substances (PFAS) Study-2021 Preliminary Report*; 2021.
(22) Bao, J.; Liu, L.; Wang, X.; Jin, Y. H.; Dong, G. H. Human Exposure to Perfluoroalkyl Substances near a Fluorochemical Industrial Park in China. *Environ. Sci. Pollut. Res.* **2017**, *24*, 9194−9201.
(23) Bao, J.; Yu, W. J.; Liu, Y.; Wang, X.; Jin, Y. H.; Dong, G. H. Perfluoroalkyl Substances in Groundwater and Home-Produced Vegetables and Eggs around a Fluorochemical Industrial Park in China. *Ecotoxicol. Environ. Saf.* **2019**, *171*, 199−205.
(24) Lu, G. H.; Jiao, X. C.; Piao, H. T.; Wang, X. C.; Chen, S.; Tan, K. Y.; Gai, N.; Yin, X. C.; Yang, Y. L.; Pan, J. The Extent of the Impact of a Fluorochemical Industrial Park in Eastern China on Adjacent Rural Areas. *Arch. Environ. Contam. Toxicol.* **2018**, *74*, 484−491.
(25) Wang, Y.; Yu, N.; Zhu, X.; Guo, H.; Jiang, J.; Wang, X.; Shi, W.; Wu, J.; Yu, H.; Wei, S. Suspect and Nontarget Screening of Per- and Polyfluoroalkyl Substances in Wastewater from a Fluorochemical Manufacturing Park. *Environ. Sci. Technol.* **2018**, *52*, 11007−11016.
(26) Yong, Z. Y.; Kim, K. Y.; Oh, J. E. The Occurrence and Distributions of Per- and Polyfluoroalkyl Substances (PFAS) in Groundwater after a PFAS Leakage Incident in 2018. *Environ. Pollut.* , *2021*, 268, DOI: 10.1016/j.envpol.2020.115395.
(27) Jacob, P.; Barzen-Hanson, K. A.; Helbling, D. E. Target and Nontarget Analysis of Per- and Polyfluoroalkyl Substances in Wastewater from Electronics Fabrication Facilities. *Environ. Sci. Technol.* **2021**, *55*, 2346−2356.
(28) Joerss, H.; Schramm, T. R.; Sun, L. T.; Guo, C.; Tang, J. H.; Ebinghaus, R. Per- and Polyfluoroalkyl Substances in Chinese and German River Water - Point Source- and Country-Specific Fingerprints Including Unknown Precursors. *Environ. Pollut.* **2020**, 267.
(29) Benotti, M. J.; Fernandez, L. A.; Peaslee, G. F.; Douglas, G. S.; Uhler, A. D.; Emsbo-Mattingly, S. A Forensic Approach for Distinguishing PFAS Materials. *Environ. Forensics* **2020**, *21*, 319−333.
(30) Kibbey, T. C. G.; Jabrzemski, R.; O'Carroll, D. M. Supervised Machine Learning for Source Allocation of Per- and Polyfluoroalkyl Substances (PFAS) in Environmental Samples. *Chemosphere* **2020**, *252*, No. 126593.
(31) Charbonnet, J. A.; Rodowa, A. E.; Joseph, N. T.; Guelfo, J. L.; Field, J. A.; Jones, G. D.; Higgins, C. P.; Helbling, D. E.; Houtz, E. F. Environmental Source Tracking of Per- and Polyfluoroalkyl Substances within a Forensic Context: Current and Future Techniques. *Environ. Sci. Technol.* **2021**, *55*, 7237−7245.
(32) Peter, K. T.; Kolodziej, E. P.; Kucklick, J. R. Assessing Reliability of Non-Targeted High-Resolution Mass Spectrometry Fingerprints for Quantitative Source Apportionment in Complex Matrices. *Anal. Chem.* **2022**, *94*, 2723−2731.
(33) Houtz, E. F.; Higgins, C. P.; Field, J. A.; Sedlak, D. L. Persistence of Perfluoroalkyl Acid Precursors in AFFF-Impacted Groundwater and Soil. *Environ. Sci. Technol.* **2013**, *47*, 8187−8195.
(34) Zhang, X.; Lohmann, R.; Dassuncao, C.; Hu, X. C.; Weber, A. K.; Vecitis, C. D.; Sunderland, E. M. Source Attribution of Poly- and Perfluoroalkyl Substances (PFASs) in Surface Waters from Rhode Island and the New York Metropolitan Area. *Environ. Sci. Technol. Lett.* **2016**, *3*, 316−321.
(35) Kibbey, T. C. G.; Jabrzemski, R.; O'Carroll, D. M. Source Allocation of Per- and Polyfluoroalkyl Substances (PFAS) with Supervised Machine Learning: Classification Performance and the Role of Feature Selection in an Expanded Dataset. *Chemosphere* **2021**, *275*, No. 130124.
(36) Yao, Y.; Zhu, H.; Li, B.; Hu, H.; Zhang, T.; Yamazaki, E.; Taniyasu, S.; Yamashita, N.; Sun, H. Distribution and Primary Source Analysis of Per- and Poly-Fluoroalkyl Substances with Different Chain Lengths in Surface and Groundwater in Two Cities, North China. *Ecotoxicol. Environ. Saf.* **2014**, *108*, 318−328.
(37) Qi, Y.; Huo, S.; Xi, B.; Hu, S.; Zhang, J.; He, Z. Spatial Distribution and Source Apportionment of PFASs in Surface Sediments from Five Lake Regions, China. *Sci. Rep.* **2016**, *6*, 1−11.
(38) Dasu, K.; Xia, X.; Siriwardena, D.; Klupinski, T. P.; Seay, B. Concentration Profiles of Per- and Polyfluoroalkyl Substances in Major Sources to the Environment. *J. Environ. Manage.* **2022**, 301.
(39) 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.
(40) Hepburn, E.; Madden, C.; Szabo, D.; Coggan, T. L.; Clarke, B.; Currell, M. Contamination of Groundwater with Per- and Polyfluoroalkyl Substances (PFAS) from Legacy Landfills in an Urban Re-Development Precinct. *Environ. Pollut.* **2019**, *248*, 101−113.
(41) US National Institutes of Standards and Technology. *Suspect List of Possible Per- and Polyfluoroalkyl Substances (PFAS)*, https://data.nist.gov/od/id/mds2-2387.
(42) Charbonnet, J. A.; McDonough, C. A.; Xiao, F.; Schwichtenberg, T.; Cao, D.; Kaserzon, S.; Thomas, K. V.; Dewapriya, P.; Place, B. J.; Schymanski, E. L.; Field, J. A.; Helbling, D. E.; Higgins, C. P. Communicating Confidence of Per- and Polyfluoroalkyl Substance (PFAS) Identification via High Resolution Mass Spectrometry. *Environ. Sci. Technol. Lett.* **2022**, 473−481.
(43) Cao, D.; Schwichtenberg, T.; Duan, C.; Xue, L.; Muensterman, D.; Field, J. Practical Semiquantification Strategy for Estimating Suspect Per- and Polyfluoroalkyl Substance (PFAS) Concentrations. *J. Am. Soc. Mass Spectrom.* **2023**, *34*, 939−947.
(44) Helsel, D. R. *Statistics for Censored Environmental Data Using Minitab and R*; John Wiley & Sons Inc.: Hoboken, NJ, 2012, pp. 268-296.
(45) Huston, C.; Juarez-Colunga, E. *Guidelines for Computing Summary Statistics for Data-Sets Containing Non-Detects*; British Columbia, 2009. http://bvcentre.ca/files/research_reports/08-03GuidanceDocument.pdf.
(46) Smith, L. J. *Literature Review Comparing the Use of One-Half the Reporting Limit to ProUCL Methods of Estimating Non-Detects*; 2018.
(47) Baccarelli, A.; Pfeiffer, R.; Consonni, D.; Pesatori, A. C.; Bonzini, M.; Patterson, D. G.; Bertazzi, P. A.; Landi, M. T. Handling of Dioxin Measurement Data in the Presence of Non-Detectable Values: Overview of Available Methods and Their Application in the Seveso Chloracne Study. *Chemosphere* **2005**, *60*, 898−906.
(48) Farnham, I. M.; Singh, A. K.; Stetzenbach, K. J.; Johannesson, K. H. Treatment of Nondetects in Multivariate Analysis of Groundwater Geochemistry Data. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 265−281.
(49) Dávila-Santiago, E.; Shi, C.; Mahadwar, G.; Medeghini, B.; Insinga, L.; Hutchinson, R.; Good, S.; Jones, G. D. Machine Learning Applications for Chemical Fingerprinting and Environmental Source Tracking Using Non-Target Chemical Data. *Cite This: Environ. Sci. Technol* **2022**, *2022*, 4090.
(50) Chicco, D. Ten Quick Tips for Machine Learning in Computational Biology. *BioData Mining*. BioMed Central Ltd. December 1, 2017.
(51) Nguyen, Q. H.; Ly, H. B.; Ho, L. S.; Al-Ansari, N.; Van Le, H.; Tran, V. Q.; Prakash, I.; Pham, B. T. Influence of Data Splitting on

https://doi.org/10.1021/acs.est.3c03770
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Performance of Machine Learning Models in Prediction of Shear Strength of Soil. *Math Probl Eng* **2021**, *2021*, 4832864.
(52) Tien Bui, D.; Tuan, T. A.; Klempe, H.; Pradhan, B.; Revhaug, I. Spatial Prediction Models for Shallow Landslide Hazards: A Comparative Assessment of the Efficacy of Support Vector Machines, Artificial Neural Networks, Kernel Logistic Regression, and Logistic Model Tree. *Landslides* **2016**, *13*, 361−378.
(53) Solo-Gabriele, H. M.; Jones, A. S.; Lindstrom, A. B.; Lang, J. R. Waste Type, Incineration, and Aeration Are Associated with per- and Polyfluoroalkyl Levels in Landfill Leachates. *Waste Manage.* **2020**, *107*, 191−200.
(54) Robey, N. M.; da Silva, B. F.; Annable, M. D.; Townsend, T. G.; Bowden, J. A. Concentrating Per- and Polyfluoroalkyl Substances (PFAS) in Municipal Solid Waste Landfill Leachate Using Foam Separation. *Environ. Sci. Technol.* **2020**, *54*, 12550−12559.
(55) Liu, T.; Hu, L. X.; Han, Y.; Dong, L. L.; Wang, Y. Q.; Zhao, J. H.; Liu, Y. S.; Zhao, J. L.; Ying, G. G. Non-Target and Target Screening of per- and Polyfluoroalkyl Substances in Landfill Leachate and Impact on Groundwater in Guangzhou, China. *Sci. Total Environ.* **2022**, 844.
(56) Lindstrom, A. B.; Strynar, M. J.; Delinsky, A. D.; Nakayama, S. F.; McMillan, L.; Libelo, E. L.; Neill, M.; Thomas, L. Application of WWTP Biosolids and Resulting Perfluorinated Compound Contamination of Surface and Well Water in Decatur, Alabama, USA. *Environ. Sci. Technol.* **2011**, *45*, 8015−8021.
(57) Gallen, C.; Bignert, A.; Taucare, G.; O'Brien, J.; Braeunig, J.; Reeks, T.; Thompson, J.; Mueller, J. F. Temporal Trends of Perfluoroalkyl Substances in an Australian Wastewa-Ter Treatment Plant: A Ten-Year Retrospective Investigation. *Sci. Total Environ.* **2022**, 804.
(58) Masoner, J. R.; Kolpin, D. W.; Cozzarelli, I. M.; Smalling, K. L.; Bolyard, S. C.; Field, J. A.; Furlong, E. T.; Gray, J. L.; Lozinski, D.; Reinhart, D.; Rodowa, A.; Bradley, P. M. Landfill Leachate Contributes Per-/Poly-Fluoroalkyl Substances (PFAS) and Pharmaceuticals to Municipal Wastewater. *Environ Sci (Camb)* **2020**, *6*, 1300−1311.
(59) Gremmel, C.; Froemel, T.; Knepper, T. P. HPLC-MS/MS Methods for the Determination of 52 Perfluoroalkyl and Polyfluoroalkyl Substances in Aqueous Samples. *Anal. Bioanal. Chem.* **2017**, *409*, 1643−1655.
(60) Eriksson, U.; Haglund, P.; Karrman, A. Contribution of Precursor Compounds to the Release of Per- and Polyfluoroalkyl Substances (PFASs) from Waste Water Treatment Plants (WWTPs). *Journal of Environmental Sciences* **2017**, *61*, 80−90.
(61) Szabo, D.; Coggan, T. L.; Robson, T. C.; Currell, M.; Clarke, B. O. Investigating Recycled Water Use as a Diffuse Source of Per- and Polyfluoroalkyl Substances (PFASs) to Groundwater in Melbourne. *Australia. Science of the Total Environment* **2018**, *644*, 1409−1417.
(62) Wang, N.; Liu, J. X.; Buck, R. C.; Korzeniowski, S. H.; Wolstenholme, B. W.; Folsom, P. W.; Sulecki, L. M. 6:2 Fluorotelomer Sulfonate Aerobic Biotransformation in Activated Sludge of Waste Water Treatment Plants. *Chemosphere* **2011**, *82*, 853−858.
(63) Dauchy, X.; Boiteux, V.; Bach, C.; Colin, A.; Hemard, J.; Rosin, C.; Munoz, J. F. Mass Flows and Fate of Per- and Polyfluoroalkyl Substances (PFASs) in the Wastewater Treatment Plant of a Fluorochemical Manufacturing Facility. *Sci. Total Environ.* **2017**, *576*, 549−558.
(64) US Environmental Protection Agency. *Multi-Industry Per- and Polyfluoroalkyl Substances (PFAS) Study − 2021 Preliminary Report*; Washington DC, 2021.
(65) Sutton, R.; Xie, Y.; Moran, K. D.; Teerlink, J. Occurrence and Sources of Pesticides to Urban Wastewater and the Environment. *ACS Symp. Ser.* **2019**, *1308*, 63−88.
(66) Sadaria, A. M.; Sutton, R.; Moran, K. D.; Teerlink, J.; Brown, J. V.; Halden, R. U. Passage of Fiproles and Imidacloprid from Urban Pest Control Uses through Wastewater Treatment Plants in Northern California, USA. *Environ. Toxicol. Chem.* **2017**, *36*, 1473−1482.
(67) Sadaria, A. M.; Labban, C. W.; Steele, J. C.; Maurer, M. M.; Halden, R. U. Retrospective Nationwide Occurrence of Fipronil and Its Degradates in U.S. Wastewater and Sewage Sludge from 2001 - 2016. *Water Res.* **2019**, *155*, 465−473.
(68) Nickerson, A.; Maizel, A. C.; Kulkarni, P. R.; Adamson, D. T.; Kornuc, J. J.; Higgins, C. P. Enhanced Extraction of AFFF-Associated PFASs from Source Zone Soils. *Environ. Sci. Technol.* **2020**, *54*, 4952−4962.
(69) Nickerson, A.; Rodowa, A. E.; Adamson, D. T.; Field, J. A.; Kulkarni, P. R.; Kornuc, J. J.; Higgins, C. P. Spatial Trends of Anionic, Zwitterionic, and Cationic PFASs at an AFFF-Impacted Site. *Environ. Sci. Technol.* **2021**, *55*, 313−323.
(70) Adamson, D. T.; Kulkarni, P. R.; Nickerson, A.; Higgins, C.; Field, J. A.; Schwichtenberg, T.; Newell, C.; Kornuc, J. J. Characterization of Relevant Site Specific PFAS Fate and Transport Processes at Multiple AFFF Sites. *Environ Adv* **2022**, No. 100167.
(71) Woudneh, M. B.; Chandramouli, B.; Hamilton, C.; Grace, R. Effect of Sample Storage on the Quantitative Determination of 29 PFAS: Observation of Analyte Interconversions during Storage. *Environ. Sci. Technol.* **2019**, *53*, 12576−12585.
(72) Helsel, D. R. More Than Obvious: Better Methods for Interpreting Nondetect Data. *Environ. Sci. Technol.* **2005**, *39*, 419A−423A.

https://doi.org/10.1021/acs.est.3c03770
Environ. Sci. Technol. XXXX, XXX, XXX−XXX