Title
Prediction of 35 Target Per- and Polyfluoroalkyl Substances (PFASs) in California Groundwater Using Multilabel Semisupervised Machine Learning

Authors
Dong, Jialin
Tsai, Gabriel
Olivares, Christopher I

Publication Date
2023

DOI
10.1021/acsestwater.3c00134

Peer reviewed

ACS ES&T Water

pubs.acs.org/estwater

Article

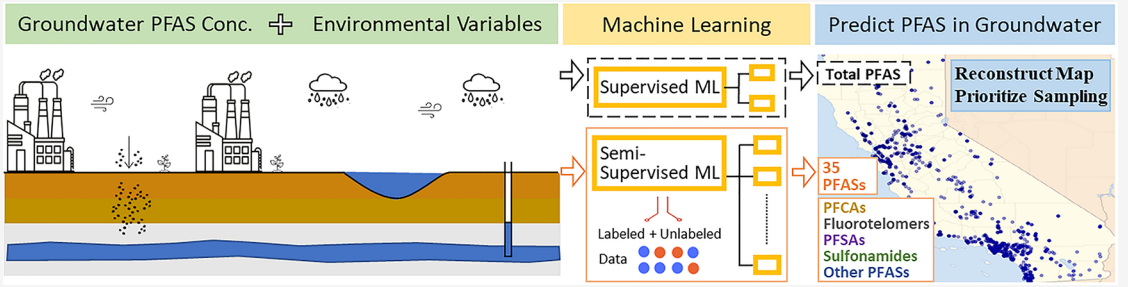
Prediction of 35 Target Per- and Polyfluoroalkyl Substances (PFASs) in California Groundwater Using Multilabel Semisupervised Machine Learning

Jialin Dong, Gabriel Tsai, and Christopher I. Olivares*

Cite This: https://doi.org/10.1021/acsestwater.3c00134

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Comprehensive monitoring of perfluoroalkyl and polyfluoroalkyl substances (PFASs) is challenging because of the high analytical cost and an increasing number of analytes. We developed a machine learning pipeline to understand environmental features influencing PFAS profiles in groundwater. By examining 23 public data sets (2016−2022) in California, we built a state-wide groundwater database (25,000 observations across 4200 wells) encompassing contamination sources, weather, air quality, soil, hydrology, and groundwater quality (PFASs and cocontaminants). We used supervised learning to prescreen total PFAS concentrations above 70 ng/L and multilabel semisupervised learning to predict 35 individual PFAS concentrations above 2 ng/L. Random forest with ADASYN oversampling performed the best for total PFASs (AUROC 99%). XGBoost with SMOTE oversampling achieved the AUROC of 73−100% for individual PFAS prediction. Contamination sources and soil variables contributed the most to accuracy. Individual PFASs were strongly correlated within each PFAS's subfamily (i.e., short- vs long-chain PFCAs, sulfonamides). These associations improved prediction performance using classifier chains, which predicts a PFAS based on previously predicted species. We applied the model to reconstruct PFAS profiles in groundwater wells with missing data in previous years. Our approach can complement monitoring programs of environmental agencies to validate previous investigation results and prioritize sites for future PFAS sampling.

KEYWORDS: *PFAS, Groundwater, Multilabel Classification, Semisupervised Learning, Pseudolabeling, Environmental Data Science*

1. INTRODUCTION

Per- and polyfluoroalkyl substances (PFASs), a family of over 5000[1] anthropogenic contaminants, are an environmental health concern[2] due to their widespread occurrence and persistence in the environment (e.g., groundwater,[3] surface water,[4] soil,[5] and precipitation[6]). State and national surveys have started to study the extent of PFAS pollution and its impact on human receptors (e.g., via drinking water). In June 2022, the U.S. Environmental Protection Agency (EPA) modified its drinking-water health advisory level of 70 ng/L[7] for combined concentrations of perfluorooctanesulfonate (PFOS) and perfluorooctanoate (PFOA) to 0.02 and 0.004 ng/L for PFOS and PFOA, respectively, in response to growing public concern. In March 2023, Maximum Contaminant Levels of 4 ng/L were proposed for PFOA and PFOS, and a Hazard Index was proposed for a combination of four other PFASs (perfluorobutanesulfonate (PFBS), perfluorohexanesulfonate (PFHxS), perfluorononoanoic acid (PFNA), and GenX).[8]

Given the rapidly evolving landscape of PFAS legislation, there are many challenges in establishing widespread monitoring programs. First, PFAS monitoring is costly because it requires advanced analytical chemistry techniques, liquid chromatography tandem mass spectrometry modified to contain low fluoropolymer PFAS cross-contamination, and robust QA/QC methods. In addition, standardized methods

Special Issue: Applications of Artificial Intelligence, Machine Learning, and Data Analytics in Water Environments

Received: March 22, 2023
Revised: August 4, 2023
Accepted: August 4, 2023

have been quickly evolving (e.g., DoD QSM Table B-15, EPA 533, Draft EPA 1633), adapting to include additional PFASs of concern based on their availability as native and mass-labeled standards. The Unregulated Contaminant Monitoring Rule 3 (UCMR3) report by the EPA[9] determined concentrations of six perfluoroalkyl acids (PFAAs; carboxylates (PFCAs) and sulfonates (PFSAs)) in public drinking water systems and groundwater. This effort found concerning PFAS concentrations (exceeding 70 ng/L, a previous health advisory level) in California water samples, where approximately 40% of drinking water is supplied by groundwater wells.[10] UCMR5 is currently underway and considers 35 individual PFASs (PFCAs, PFSAs, fluorotelomer sulfonates (FTSs), sulfonamide derivatives, chlorinated PFASs, and ether carboxylates (PFECAs)). The inclusion of additional PFASs is important to recognize shifts in PFAS manufacture and use (e.g., long- to short-chain PFAAs, PFECAs), as well as polyfluorinated transformation in the environment.[11]

The distribution of PFASs in groundwater is influenced by contamination sources and environmental fate drivers, such as hydrology,[12] geochemistry, microbial transformations,[11,13] weather events,[6,14] and air deposition.[15,16] Several studies have reported the mechanisms of transport and fate of PFASs from pollution sources to soil and groundwater.[17,18] Recently, PFAS air emissions have been studied using a Community Multiscale Air Quality model to link the emission concentrations, deposition distances, and deposition volumes of different PFASs.[15] Moreover, some individual PFASs are known to be used by applications and industries with common co-occurring contaminants (i.e., chlorinated solvents, hydrocarbons, and heavy metals), all of which can inform PFAS monitoring strategies. As more PFAS monitoring data become available, missing concentrations of individual PFAS that were not included in previous monitoring efforts could be predicted based on correlations with measured PFAS target analytes and with environmental features. Historical PFAS profiles can also inform us about the comprehensiveness of PFAS monitoring policies, remediation, and PFAS source treatment.

To complement PFAS analytical efforts that cannot sample every single well, we can leverage data science tools to predict unmeasured PFAS data from previous years and identify potential unmonitored PFAS hot spots. Machine learning (ML) is an application of artificial intelligence that learns from big data without programming explicitly. Several ML methods have been successfully applied to predict and monitor environmental pollutants in groundwater. Logistic regression, random forests, and Bayesian networks are commonly used for analytes such as heavy metals,[19] nitrate,[20] fluoride,[21] and PFAS.[22−24] ML PFAS studies have predicted the occurrence of PFAS in water resources. The PFAS source[25] and cocontaminants[22,26] were considered the most important features for these ML predictions. However, state-wide PFAS contamination in groundwater may be too complex to be explained by a single hydrological or geochemical model.[27] PFAS flux sinks in groundwater not only stem from sewage discharge but also from air deposition,[28] as well as infiltration of soil contaminants[25] and colloid-facilitated transport.[29]

Before implementing a comprehensive analysis and prediction for PFASs in groundwater, a large environmental data set with hundreds of validated data (PFASs and other environmental features) is needed. Currently, many studies use their measured data,[26] and some use partially public data,[22,25] all of which have considered limited environmental variables (industrial sources, geography, hydrology, soil, and co-contaminants). The process of self-data collection and measurement is costly and time-consuming. While other fields with more experience in data science have public data sets (e.g., ImageNet[30] for image classification, AlphaFold[31] for protein structure prediction, and AirNet[32] for air pollution monitoring), the environmental chemistry field for groundwater contamination prediction lacks such data sets. PFAS and other environmental data have been in separate data sets, limiting analysis of trends and correlations of PFASs in context with other environmental features in groundwater. Before evaluating the performance of different machine learning tools for groundwater contaminant prediction, a comprehensive public data set is needed to validate models and seek correlations of PFAS processes with environmental features.

We hypothesize that because PFASs can be grouped in subfamilies, we can leverage tools that take advantage of these subfamily groupings based on their different industrial uses, physicochemical properties, and (bio)transformation pathways. Multilabel classification is an ML approach that can leverage the nature of multiple subfamilies within PFASs. These subfamilies (carboxylic acids, sulfonates, fluorotelomers, etc.) have a range of physicochemical characteristics (i.e., mobility of short-chain vs long-chain) and are also related through transformation pathways (i.e., fluorotelomers transform to carboxylic acids). Therefore, we hypothesize that we can leverage these relationships and use PFASs with a higher detection frequency and inclusion in analytical lists to inform analytes that were not measured in previous years.

The overall objective of this study is to identify how environmental variables impact total PFAS and individual PFAS distribution in California groundwater to complement groundwater PFAS monitoring on a large scale. In this work, we 1) compile an integrated groundwater PFAS data set for California CA-PFAS-ASGWS (2016−2022) with air, water, and soil parameters, contamination sources, and other environmental variables, with 26 000 data points and hundreds of features; 2) develop an ML pipeline to screen groundwater wells for total PFAS concentration over 70 ng/L and then predict the occurrence of 35 individual PFASs above 2 ng/L, thus resolving the challenge of missing or unverifiable PFAS data; and 3) apply ML to identify potential missed individual PFAS hot spots in recent investigations and prioritize future sampling locations in California groundwater wells.

2. METHODS

2.1. Data. *2.1.1. Data Acquisition.* To construct a comprehensive PFAS database that can better train ML algorithms, we collected six different classes of data, including contamination source, geospatial, and sampling date, weather and air quality, soil, hydrology, and groundwater quality. We collected a total of 23 data sets (12 groundwater-related data sets from Groundwater Ambient Monitoring and Assessment (GAMA), one PFAS data set from Geotracker, three source data sets from the EPA, one PFAS source data set from the Environmental Working Group (EWG), one soil data set from the National Cooperative Soil Survey (NCSS), one weather data set from the National Oceanic and Atmospheric Administration (NOAA), two air quality data sets from the EPA and Purple Air, two hydrology data sets from the Sustainable Groundwater Management Act (SGMA) and Global Land Data Assimilation System (GLDAS-2.1) from the National Aeronautics and Space Administration (NASA)).

https://doi.org/10.1021/acsestwater.3c00134
ACS EST Water XXXX, XXX, XXX−XXX

Table 1. Summary Statistics for 38 PFAS Concentrations (ng/L) in the CA-PFAS-ASGWS Data Set[a]

Compound	n	%NA	Mean	Min	Q1	Median	Q3	Max
PFBA	2198	91.83	227.27	0	3.60	7.40	19.00	112000
PFPeA	2300	91.45	648.59	0	1.90	3.80	16.00	446000
PFHxA	25538	5.07	69.40	0	2.00	2.00	5.00	616000
PFHpA	26086	3.03	16.71	0	1.80	2.00	3.10	48700
PFOA	26885	0.06	29.28	0	2.00	2.90	8.70	79000
PFNA	26112	2.93	4.28	0	1.80	2.00	2.00	19100
PFDA	25608	4.81	2.21	0	1.70	2.00	2.00	737
PFUnA	25040	6.92	2.16	0	1.70	2.00	2.00	550
PFDoA	25040	6.92	2.07	0	1.70	2.00	2.00	280
PFTrDA	25609	4.80	2.16	0	1.70	2.00	2.00	650
PFTeDA	25608	4.81	2.11	0	1.70	2.00	2.00	370
PFHxDA	220	99.18	35.66	0	0.00	0.00	2.08	5500
PFODA	145	99.46	0.69	0	0.00	0.00	0.95	4.89
3:3FTCA	36	99.87	1.83	0	0.29	2.07	2.12	9.55
5:3FTCA	36	99.87	5.74	0	0.22	2.61	3.08	50
7:3FTCA	36	99.87	2.98	0	0.37	2.04	2.12	60
4:2FTS	2261	91.60	11.69	0.00	0.00	2.00	7.70	7210
6:2FTS	2277	91.54	703.94	0.00	0.00	7.10	8.00	5180000
8:2FTS	2277	91.54	23.90	0.00	0.00	2.21	7.70	16300
10:2FTS	146	99.46	0.88	0.00	0.00	0.00	0.95	31.6
PFBS	25380	5.65	29.40	0.00	2.00	2.00	5.00	242000
PFPeS	2252	91.63	297.62	0.00	0.49	2.10	4.00	318000
PFHxS	25381	5.65	102.67	0.00	2.00	3.00	7.30	1330000
PFHpS	2314	91.40	27.71	0.00	0.00	1.90	3.70	28300
PFOS	26879	0.08	74.63	0.00	2.00	3.60	14.00	383000
PFNS	1606	94.03	1.97	0.00	0.00	1.90	3.70	140
PFDS	2294	91.47	2.02	0.00	0.00	1.80	3.20	190
FOSA	2308	91.42	8.81	0.00	0.00	1.98	3.80	4700
ETFOSE	292	98.91	6.08	0.00	0.00	0.00	3.00	430
ETFOSA	404	98.50	4.39	0.00	0.00	0.00	0.94	440
NETFOSAA	25186	6.38	2.68	0.00	1.70	2.00	3.00	650
MEFOSE	292	98.91	10.60	0.00	0.00	0.00	3.80	700
MEFOSA	425	98.42	3.00	0.00	0.00	0.00	0.92	220
NMEFOSAA	25167	6.45	2.64	0.00	1.70	2.00	3.00	710
ADONA	21637	19.57	2.14	0.00	1.70	2.00	2.00	200
HFPO−DA	21468	20.20	2.50	0.00	1.70	2.00	2.00	750
11ClPF3OUDS	21776	19.05	2.07	0.00	1.70	2.00	2.00	190
9ClPF3ONS	21774	19.06	2.05	0.00	1.70	2.00	2.00	190

[a]n = number of observations; %NA = percentage of missing observations.

The combined data set includes 425 columns and 26,901 rows. Additional data source descriptions and the data dictionary can be found in Supporting Information (SI) Table S1 and Table S2.

2.1.2. CA-PFAS-ASGWS Data Set Preparation. After collecting the data, we cleaned data sets by dropping non-PFAS features with over 90% of the not available values (NAs), outliers (top 2 percentile only for the PFASs), and errors (invalidated observations). On the basis of PFAS measurement wells, we merged all data sets using longitude, latitude, and the sampling date. We first merged by the exact date, latitude, and longitude. For wells that did not have an exact match, we used the fuzzy matching method and tried to supplement the data with the most adjacent data points on the same date. For the rest of the missing data, we filled the NAs with the monthly average value of data points within the closest radius with available data (10, 20, 30, or 50 km). Based on the EPA report, we classified PFAS point sources into two categories, suspected source and confirmed source (facilities reporting PFAS discharges). The confirmed sources included 1247 facilities (35 different types). In the PFAS data set, we recorded the nearest facility's type and counted the total number of facilities within a 1 km, 3 km, 10 km, and 50 km radius of each well based on groundwater and aerial deposition transport distances.[15,28,33] For the soil properties, we chose to use the mean value of all soil horizons because we lacked data on the groundwater well depths. The final cleanup CA-PFAS-ASGWS data set included 38 individual PFASs (Table 1), a total of 157 columns, and 26,901 rows. The distribution map of PFAS sampling locations and the number of PFAS observations from 2016 to 2022 are shown in the SI (Figure S1 and Table S3).

2.1.3. Summary Statistics and Correlations. Summary statistics for PFASs (Table 1) and other environmental variables (SI Table S4) were performed to get the number of observations, Mean, Minimum, Median, Maximum, 25th percentile, and 75th percentile. We evaluated correlations using the Spearman's Rank coefficient (α = 0.05) among individual PFASs, as well as between other environmental variables and PFASs.

https://doi.org/10.1021/acsestwater.3c00134
ACS EST Water XXXX, XXX, XXX−XXX

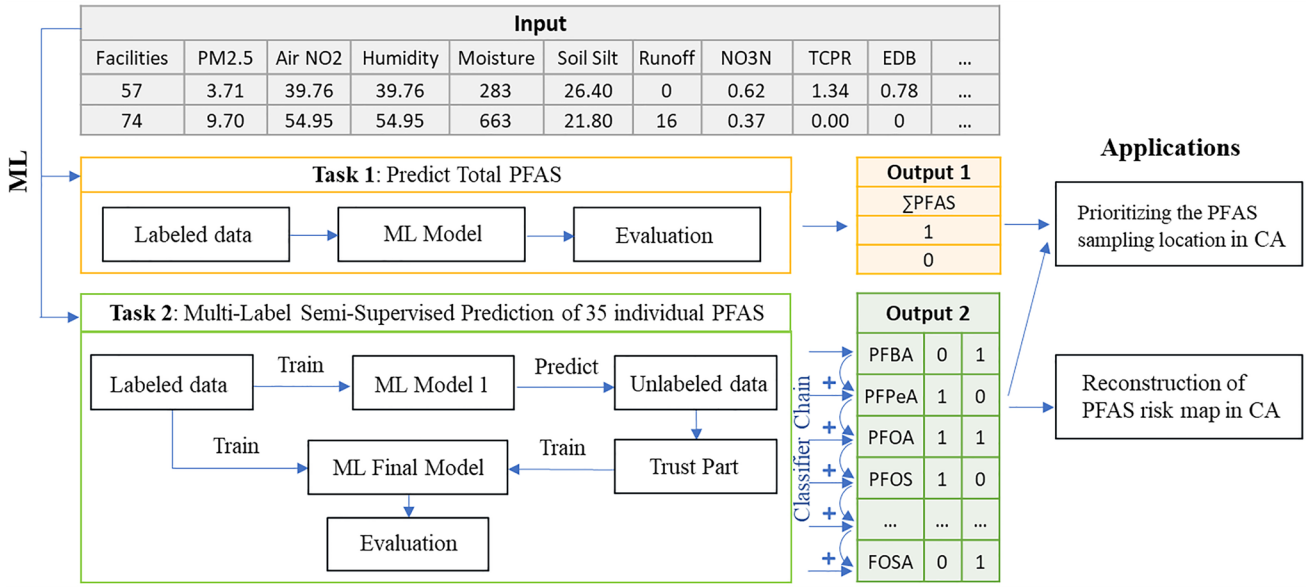


Figure 1. Workflow of supervised and semisupervised Machine Learning (ML) for binary classification and multilabel classification and its application for PFAS prediction and past PFAS profiles.

2.1.4. Data Preprocessing to Train a Machine Learning (ML) Model. To develop a model with the best performance, we preprocessed the data before training and testing the model. We dropped the variables with more than 40% of NAs. The categorical variables were encoded by a label encoder with a value between 0 and the number of classes to avoid high memory consumption. The rest of the missing data were filled in with the variable average for air quality, soil, and cocontaminant variables or zeroes for precipitation and snowpack. To diminish dependence on explicit geographical parameters and enhance the model's applicability across various locations, we decided to exclude longitude, latitude, county, and other direct spatial indicators from our input data. The final number of input variables was 112.

To develop a prescreening binary classification ML algorithm for total PFAS risk prediction, we calculated the sum of all individual PFAS concentrations, set 70 ng/L as the threshold for high risk, and added it as a new column in the data set. Then, we focused on 35 individual PFASs to predict the risk for each of them above the most stringent minimum reporting level (2 ng/L) on individual PFASs based on UCMR 5.[34] 3:3 fluorotelomer carboxylic acid (FTCA), 5:3 FTCA, and 7:3 FTCA were excluded because they only had 36 observations for the entire data set compared to at least a hundred observations for other individual PFASs.

2.2. ML Algorithms. We developed a pipeline to evaluate the PFAS risk in California groundwater (Figure 1). First, we prescreened wells to predict total PFASs greater than 70 ng/L. Second, we predicted 35 individual PFASs greater than 2 ng/L in groundwater. Because each PFAS has a different number of observations (Table 1), we split the data set into four Classes (SI Table S5) to avoid including NAs for each PFAS prediction. Class 0 included a total of 14 PFASs, each of which had more than 25,000 observations (n∼3821 wells). Class 1 had 4 PFASs with >20,0000 observations (n∼2966 wells) each, Class 2 had 10 PFASs with >1,500 observations (n∼925 wells) each, and Class 3 had 7 PFASs with >100 observations (n∼99 wells).

To make up for the lack of labeled data on PFASs, we applied Semisupervised Learning,[35] which improves the performance of the model in the supervised process using unlabeled data. Pseudolabeling[36] was used in this paper to address a common issue of data sparseness in environmental data sets.[37] Here, we first trained a classification model based on labeled data sets to make predictions on the unlabeled PFAS data. Then we selected trusted samples, which were predicted accurately with a confidence exceeding 95%, and reintroduced these samples back into the model for further training.[36]

To balance the data set, the Synthetic Minority Oversampling Technique (SMOTE)[38] and Adaptive Synthetic (ADASYN)[39] oversampling algorithms for the minority samples were performed before training the classification algorithm. The ML algorithms included Naive Bayes, logistic regression, random forest, support vector classifier (SVC), Extreme Gradient Boosting (XGBoost), categorical boosting (CatBoost), Light Gradient Boosting Machine (lightGBM), and TabNet. For the total PFAS risk prediction, we compared eight binary classification ML models. For multilabel classification, we developed 10 ML models, including seven traditional ML algorithms with problem transformation methods (Binary Relevance, Classifier Chain, and Lowest Powerset) and 3 ML models using algorithm adaptation methods. Leveraging the correlations between individual PFASs, we predicted individual PFASs based on previously predicted PFASs within a Class. Additional details on ML algorithms and approaches can be found in SI Text S1.

2.3. Training, Hyperparameter Optimization, and Evaluation. In our study, we randomly split the CA-PFAS-ASGWS data set into an 80% training set and a 20% final test set. In the training set, we further split the data using 80% for model training and the remaining 20% for hyperparameter tuning, applicable for both total and individual PFAS prediction models. For the total PFAS prediction model, we implemented stratified cross-validation, which involved dividing the training data (inclusive of the validation set) into ten subsets. Each iteration of the process trained the model on

https://doi.org/10.1021/acsestwater.3c00134
ACS EST Water XXXX, XXX, XXX−XXX

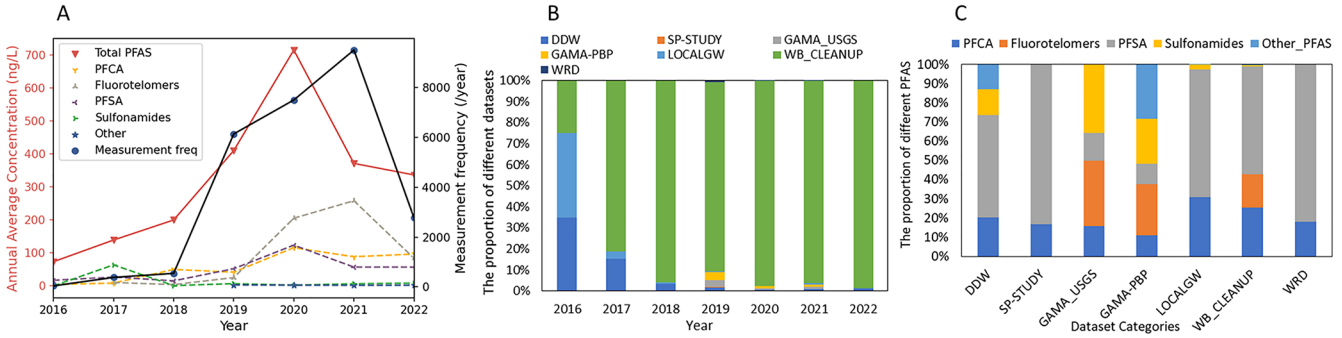


Figure 2. PFAS mean concentration and frequency of detection between 2016 and 2022 in California groundwater wells. Panel A: Total PFAS measurement frequency, total PFASs, perfluorinated carboxylic acids (PFCAs), fluorotelomers, perfluoroalkyl sulfonates (PFSAs), sulfonamides, and other PFAS annual average concentrations. Panel B: Average concentration of total PFASs in different well categories. Panel C: Proportion of the average concentration of PFAS subfamilies for each groundwater well category.

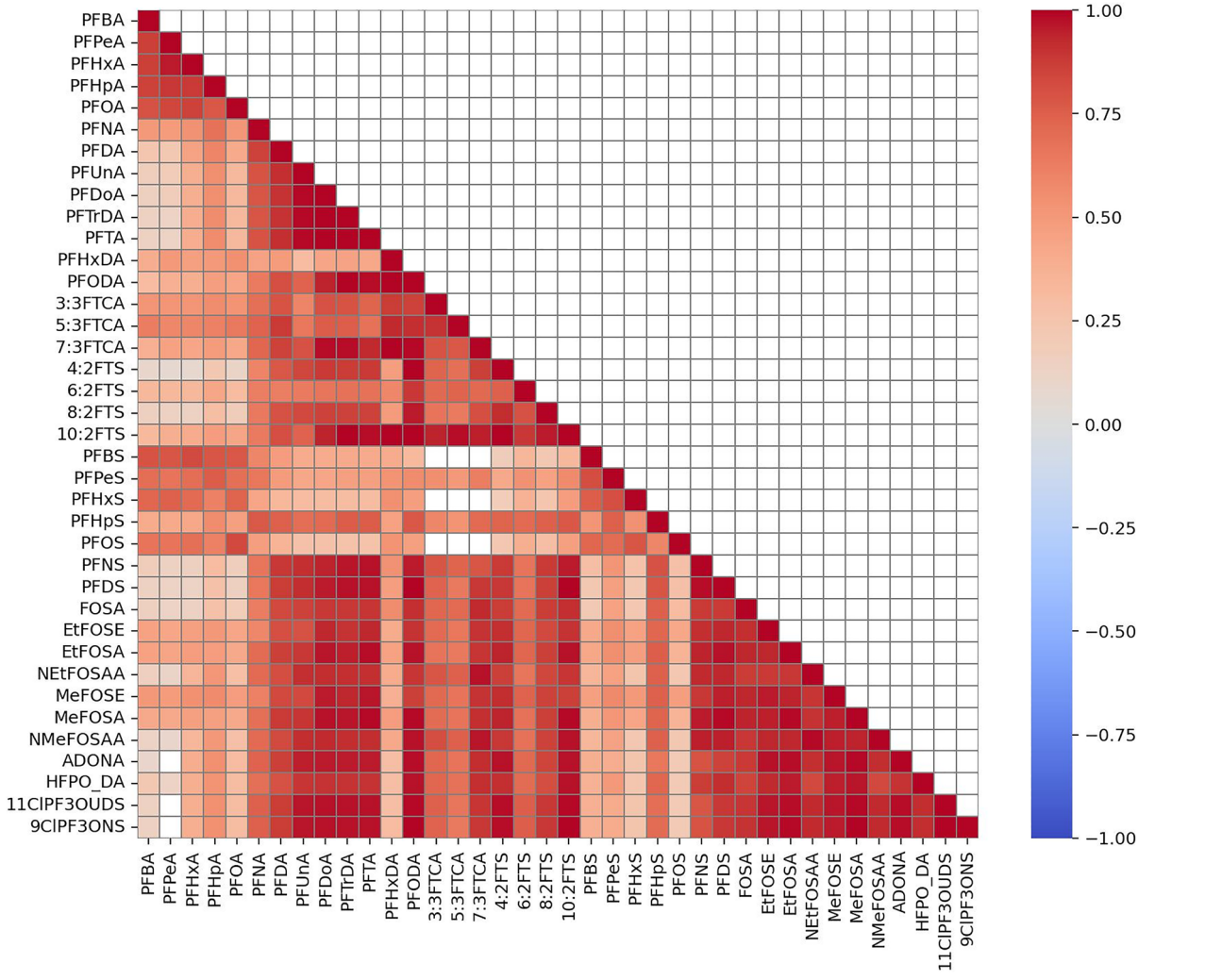


Figure 3. Heatmap of the PFAS Spearman's Rank (p 0.05) correlation coefficient analysis in California groundwater.

nine subsets and validated it on the remaining one. Grid search optimization was employed to tune the hyperparameters, and the mean F1 score of the validation set was used to find the best model. To improve the model performance, especially for the PFAS lacking labels, we used pseudolabeled data and selected the trusted labels to cotrain the best ML model. To evaluate the performance of different models for the total PFAS binary classification, we use frequently reported evaluation metrics: Accuracy, Precision, Recall, F-Score, and Area Under the Receiver Operating Characteristic curve (AUROC), each of which can range from 0 (the lowest performance) to 1 (the highest performance). The evaluation for multilabel classification is more complicated than binary classification, and thus we used evaluation metrics that have been proposed previously.[40,41] We chose to use hamming loss, which measures the proportion of the number of incorrectly

https://doi.org/10.1021/acsestwater.3c00134
ACS EST Water XXXX, XXX, XXX−XXX

Table 2. Evaluation Scores for the Baseline of Different Machine Learning Models for Total PFAS Prescreening[a]

	GaussianNB	LogReg	SVM	RF	XGBoost	CatBoost	lightGBM	TabNet
Accuracy	0.7110	0.7110	0.7110	**0.9492**	0.9359	0.9071	0.8909	0.8525
Recall	0.0022	0.0013	0.0013	**0.9349**	0.8344	0.7606	0.7065	0.8240
Precision	0.3359	0.1714	0.1714	**0.9737**	0.9321	0.8982	0.8852	0.6905
F1 score	0.0043	0.0026	0.0026	**0.9644**	0.8781	0.8224	0.7818	0.7514
AUROC	0.5008	0.5005	0.5005	**0.9479**	0.9058	0.8636	0.8362	0.8436

[a]GaussianNB: Gaussian Naive Bayes; LogReg: logistics regression; SVM: support vector machines; RF: random forest; XGBoost: Extreme Gradient Boosting; CatBoost: Categorical Boosting; lightGBM: Light gradient-boosting machine.

predicted labels in the total number of labels in all samples. The smaller the hamming loss, the better the model performance.[42] We also used average precision and average recall, where the partially correct concept is considered to calculate the average for all the samples.[43,44] We used the Exact Match Accuracy, where the result would be considered correct when the predicted set of labels exactly matches the true label for each sample.[45,42] We also calculated the AUROC score for each PFAS and calculated the average AUROC for each multilabel model (equation in SI Table S6). The development and evaluation of ML models were coded with the sklearn, xgboost, CatBoost, lightgbm, and PyTorch (TabNet) packages in Python. Oversampling was performed with the imblearn package in Python.[46]

3. RESULTS AND DISCUSSION

3.1. California PFAS Analysis. *3.1.1. PFAS Occurrence in California Groundwater.* From the 38 PFASs included in the CA-PFAS-ASGWS database, PFOA and PFOS are the most frequently reported species (95% detection in all samples), indicating their widespread presence in California's groundwater. For these two species, roughly 60% of the samples were higher than 2 ng/L, 40% and 48% of samples had concentrations above 4 ng/L, and 2.0% and 3.1% of samples had concentrations above 70 ng/L, for PFOA and PFOS, respectively. Compared to the PFASs above 70 ng/L in 14% of the Eastern United States' groundwater wells,[26] only ∼5% exceeded this threshold in California.

The number of PFAS analytes monitored has been increasing each year (annual PFAS concentrations, Figure S2 and Table S3). From 2016 to 2022, 17, 26, 24, 38, 35, 38, and 38 PFASs were measured in California groundwater each year, respectively. PFCAs were the most frequently detected PFASs. The average concentration of total PFAS in California also increased during this time span (Figure 2A). Since 2019, the detection frequency and average concentration of PFASs in California groundwater have increased plausibly because of the PFAS Action Plan[47] released by the US EPA as well as the PFAS source investigations and adjacent public drinking water supply sampling ordered by the California Water Boards.[48] The FTSs' average concentration has increased compared to overall PFAS frequency and the proportion of the samples from the contamination cleanup sites (SI Figure S2). The concentrations of FTCAs and FTSs were significantly higher than other PFASs in 2019 and 2020 (Figure 2 and SI Figure S3). 6:2 FTS is commonly found in sites impacted with AFFF and therefore the cleanup data set, because it is a transformation intermediate of fluorotelomer PFASs in AFFF formulations.[11,49,50]

At least 18 individual PFASs were detected in each data source between 2016 and 2022 (SI Table S3), but the PFAS profiles varied depending on the type of groundwater well (Figure 2). The average concentration of total PFASs from the unknown well type was the highest, followed by the monitoring well type. PFSAs contributed the most to the total PFAS concentration in monitoring, unknown, and other water supply well types. Groundwater wells in cleanup sites comprised about 77% of the wells monitored in California since 2018 and had detections for all 38 PFASs (in decreasing order of frequency of detection: PFSAs > PFCAs > FTSs> sulfonamides > other). The highest total amount of PFASs in a given sample (2,590,829 ng/L) was found in a well measured in 2020 within the cleanup sites. Detailed statistical analysis of well categories and PFASs is included in SI Text S2.

3.1.2. Groundwater PFAS Correlations. We performed correlation tests (Spearman's Rank coefficient, $\alpha = 0.05$) to determine association between individual PFASs within PFAS subclasses (i.e., fluorotelomers, short- vs long-chain, sulfonates, carboxylates) and known precursor-transformation product biotransformation pathways (i.e., fluorotelomer (bio)-transformation into short-chain PFCAs). Statistically, significant correlations were overall positive and generally among PFAS subfamily species (Figure 3), indicating similar physicochemical properties or common industrial applications. Short-chain PFCAs had the strongest correlation coefficients (>0.75) among each other, as well as with PFOA and PFSAs. Similarly, the less mobile longer-chain C11−C18 PFCAs had strong relationships with each other and with fluorotelomers and sulfonamides. PFOS had weaker correlation coefficients, possibly due to its high detection frequency and a lot of noise in the data. Biotransformation pathway associations were weakly detected in the data set. Regarding fluorotelomers, FTCAs had higher correlation coefficients than FTS with their expected oxidation products, short-chain PFCAs,[11] but coefficients remained weak to moderate (0.25−0.5). Interestingly, *N*-alkyl sulfonamides (i.e., ETFOSE, MEFOSAA, etc.) did not show a strong correlation with their expected aerobic terminal transformation products, PFSAs, but had strong correlations with FOSA. This is consistent with a lower extent of biotransformation of sulfonamides in the environment to PFSAs[51] and the detection of perfluorosulfonamides in AFFF-impacted groundwater.[52,53] It is worth noting that although FOSA is included in the analyte list, perfluorohexane sulfonamide is not included in standard PFAS methods. This suggests that there is a gap in PFAS profiles in groundwater. Surprisingly, ADONA and HFPO-DA (GenX) are much more mobile than sulfonamides, yet both were strongly correlated. This suggests locations of simultaneous use of both subfamilies of PFASs. Overall, the correlations encourage the development of ML models that predict and reconstruct PFAS distribution maps for an increasing number of analytes. No significant negative correlations were found between PFASs.

3.2. Total PFAS ML Performance. We selected a Random Forest algorithm using ADASYN oversampling as our best-

https://doi.org/10.1021/acsestwater.3c00134
ACS EST Water XXXX, XXX, XXX−XXX

performing total PFAS prescreening model to determine wells with total PFAS concentrations below or above 70 ng/L. Based on ML model baseline performance in Table 2, and SI Figure S4, for total PFAS prediction, random forest performed the best, followed by the boosting algorithm (XGBoost > CatBoost > LightGBM), and then the other traditional algorithms (GaussianNB > LogReg > SVM) in the CA-PFAS-ASGWS data set. The prescreening PFAS model was robust for the total PFAS risk classification task in the CA-PFAS-ASGWS data set, accurately predicting total PFAS with a concentration higher than 70 ng/L in the groundwater (AUROC = 99%, Accuracy = 96%) in the final test set. The recall rate is up to 95%, and the precision is up to 97%. Because there are NAs in the data set, the total PFASs calculated could be underestimated, but to offset false negative our criteria is more strict than using PFOA + PFOS greater than 70 ng/L. The average accuracy achieved through 10-fold cross-validation on the training set closely mirrored that of the validation set (~0.96 for both), indicating that the model did not overfit. In the final best random forest model development, the n_estimators was 200, the max_depth was 52, and the max_features was 35. Tuning the hyperparameters helped to increase the accuracy and AUROC for all the models mentioned, and the random forest model accuracy increased by 4.58%. The slightly increased model performance was similar to the result from Hu et al.[25]

Moreover, compared with a previous model to predict PFAS risk in California groundwater,[25] our study increased the AUROC by 10% for total PFAS prediction. Our finding is consistent with previous PFAS studies and other types of tabular data. Random forest has performed well for different PFAS predictive tasks in several environmental data sets.[22,25,54,55] Borisov et al.[56] proposed the traditional decision tree ensemble ML algorithms like random forest, XGBoost, LightGBM, and CatBoost still perform better than other traditional ML models and deep learning models. Therefore, it is worth comparing different decision tree ensemble algorithms.

To balance the data, we evaluated different oversampling and undersampling methods. The ADASYN oversampling method improved the prescreening model's performance the most (increased the accuracy by 1.77% compared with the model without oversampling on the same validation data set), followed by SMOTE. Both methods are better than random oversampling, which duplicates the samples. The undersampling methods, deleting or merging examples in the majority of samples, did not work. SMOTE and ADASYN oversample the data by creating synthetic data instead of duplicating it, so these methods can overcome the overfitting problem raised by replicating the samples.

3.3. Multilabel Semisupervised ML Performance for Individual PFASs. In addition to the total PFAS ML model mentioned above, we also developed multilabel classification models with semisupervised learning algorithms to predict the occurrence of each PFAS above 2 ng/L. Our model's AUROC (94%−99%) is higher than other studies (AUROC 72%−96%)[22,25,26] in the final test set. Table 3 shows the Hamming loss, exact match accuracy, average precision, average recall, and average AUROC for the final best models using the best oversampling methods and best parameters for each Class. The AUROC score for each of the PFASs is shown in SI Table S7.

Comparing all the results from different multilabel models (SI Table S8), the classifier chain XGBoost algorithm performed best for all the PFAS prediction Classes. The label powerset random forest model achieved the highest exact match accuracy in Class 0 (individual PFASs with >25,000 observations) prediction. The classifier chain random forest model achieved the highest exact match accuracy in Class 1 (>20,000 observations) and 2 (>1,000 observations) predictions. For Class 3 (>100 observations), the AUROC of the XGBoost, CatBoost, and LightBoost using the classifier chain and label powerset was up to 95%. To balance the data set, the SMOTE oversampling method performed the best for the multilabel classification, followed by ADASYN and random oversampling. Finally, we used the XGBoost model of the SMOTE oversampling method as our final multilabel PFAS model.

Table 3. Model Evaluation for the Semisupervised Learning Models

	Class 0	Class 1	Class 2	Class 3
Hamming loss	0.045	0.034	0.063	0.007
Exact match accuracy	0.883	0.961	0.789	0.950
Average precision	0.833	0.602	0.831	0.217
Average recall	0.839	0.604	0.815	0.219
Average AUROC	0.938	0.965	0.942	0.985

To further improve the model's performance, we applied semisupervised learning, adding more trusted unlabeled data (missing PFAS measurements) with the labeled data to train the model for the individual PFASs with several missing labels (Classes 1, 2, and 3). For Classes 1 and 3, the average AUROC improvement is less than 1%. For Class 2 PFAS, the average AUROC increased by 2.9% (SI Table S7). The ratio of the labeled data to the unlabeled data was around 3, 0.1, and 0.01 for Classes 1, 2, and 3, respectively.

Our results indicate both the number and quality of initial labeled points have an impact on pseudolabeling. This result is consistent with previous research,[57,58] that pseudolabeling works best with ~1000 initial labeled samples (Class 2). When the labeled data set is reduced (e.g., less than 100 points in Class 3), the pseudolabeled performance starts to decrease. For our Class 2 prediction, pseudolabeling can improve the model performance when the labeled data is large enough to get a robust prediction model, the initial unlabeled data is large enough to offer more information for the model, and the algorithm has enough complexity to benefit from the additional data.[57] For Class 1, the initial unlabeled data may not be large enough to improve the model. For Class 3, the initially labeled data set is too small, so the pseudolabeling may not be robust (low recall). Moreover, if the labeled data set contains outliers, the prediction for the unlabeled data may be incorrect and further reduce the model's performance.[59]

The multilabel PFAS semisupervised learning model is robust in the CA-PFAS-ASGWS data set and can accurately predict individual PFASs with concentrations higher than 2 ng/L efficiently. For the Class 0 to 3 predictions, the average AUROC values were 0.938, 0.965, 0.914, and 0.952. For each of the individual PFASs, the lowest AUROC (0.729) is for the prediction of PFBA. For the rest of the PFASs, the AUROC ranged from 0.908 to 1.00. By leveraging the strong correlations of PFAS subfamilies (i.e., PFCAs, PFSAs, and FTSs), instead of building 35 binary classification models (below or above 2 ng/L) for each PFAS, our study used the classifier chain set ensemble and applied XGBoost as a classifier. Four multilabel algorithms for each data Class were

https://doi.org/10.1021/acsestwater.3c00134
ACS EST Water XXXX, XXX, XXX−XXX

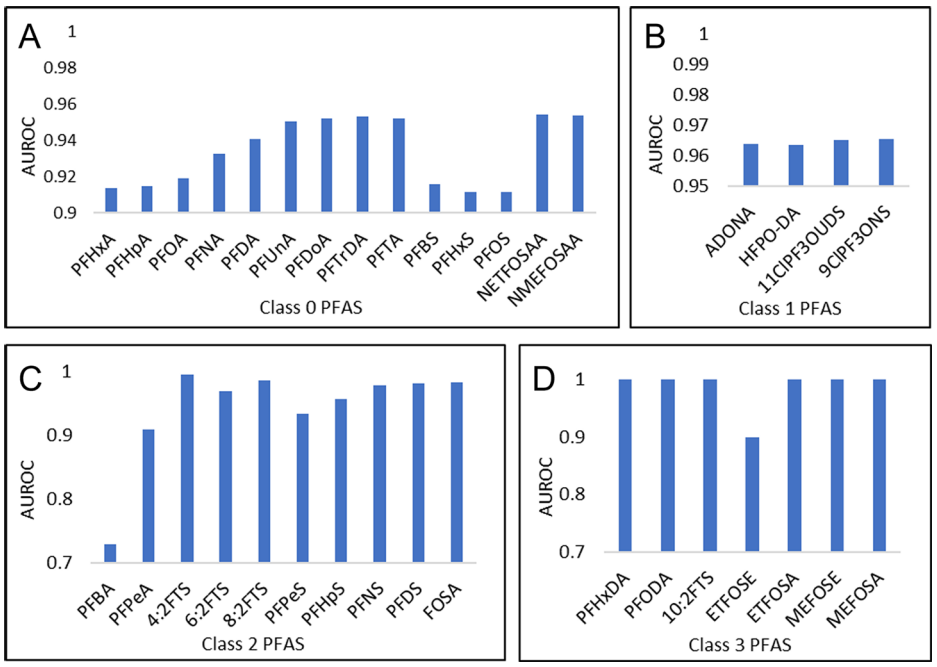


Figure 4. Model evaluation AUROC results for each PFAS in semisupervised learning models. Panel A: Class 0 includes 14 PFASs (~3800 wells, over 24,000 observations). Panel B: Class 1 includes PFASs (~3000 wells, over 21,000 observations). Panel C: Class 2 includes PFASs (~1000 wells, over 1,500 observations). Panel D: Class 3 includes PFASs (~100 wells, over 100 observations).

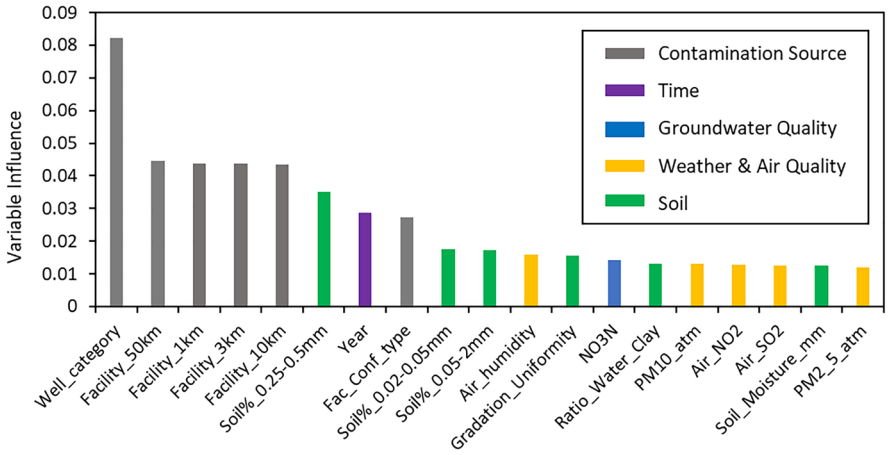


Figure 5. Relative contributions of the top 20 variables sorted from the largest to smallest importance for the PFAS prescreening model. Facility_1km, 3km, 10km, and 50km represent the number of all facilities (PFAS confirmed source and potential source) within the radius. Soil%_0.25−0.5mm (medium sand), Soil%_0.02−0.05 mm (silt), and Soil%_0.05−2mm (all sand)[61] are the gravimetric percent of particles reported on a clay free < 2 mm base, respectively. Gradation_Uniformity is the coefficient of uniformity, calculated from the cumulative grain size distribution curve (gradation curve) to evaluate the grading characteristics of coarse materials in the Unified Soil Classification System. Ratio_Water_Clay is the ratio of 15 bar water percent/clay percent, calculated by (W15AD or W15FM)/CLAY_TOT, reported as grams per gram, on a <2 mm base. Additional details on each variable are listed in the SI Table S9.

developed using semisupervised learning to learn from the labeled and unlabeled data. The model was used to predict the risk of several PFAS agents in a single step.

The performance using the classifier chain was better than the binary relevance, indicating that the PFAS concentrations and the occurrence of PFASs were indeed related to each other. The accuracy of the PFAS prediction can be improved by adding the previous PFASs label as an input feature. In the experiment, we ordered individual PFASs based on PFCA, FTS, PFSA, sulfonamides, other PFASs subfamilies, and within each group, based on increasing perfluoroalkyl chain length. The AUROC of each PFAS also increased within each prediction Class for the correlated PFASs (Figure 4 and SI Table S7). Specifically, for the PFCAs, the prediction accuracy for the later predicted PFCAs and FTS was higher than for the short-chain PFCAs that were predicted first (Figure 4, Class 1 and Class 2). PFSA and its precursors, sulfonamides, also show the same pattern (Figure 4, Class 1 and Class 2).

Although the prescreening and the multilabel semisupervised learning models achieved high accuracy, there are limitations. The model has a higher probability of predicting data from recent years, because PFAS detection frequency and average concentration have been increasing each year (Figure 2) and most of the data come from the cleanup data set, which

https://doi.org/10.1021/acsestwater.3c00134
ACS EST Water XXXX, XXX, XXX−XXX

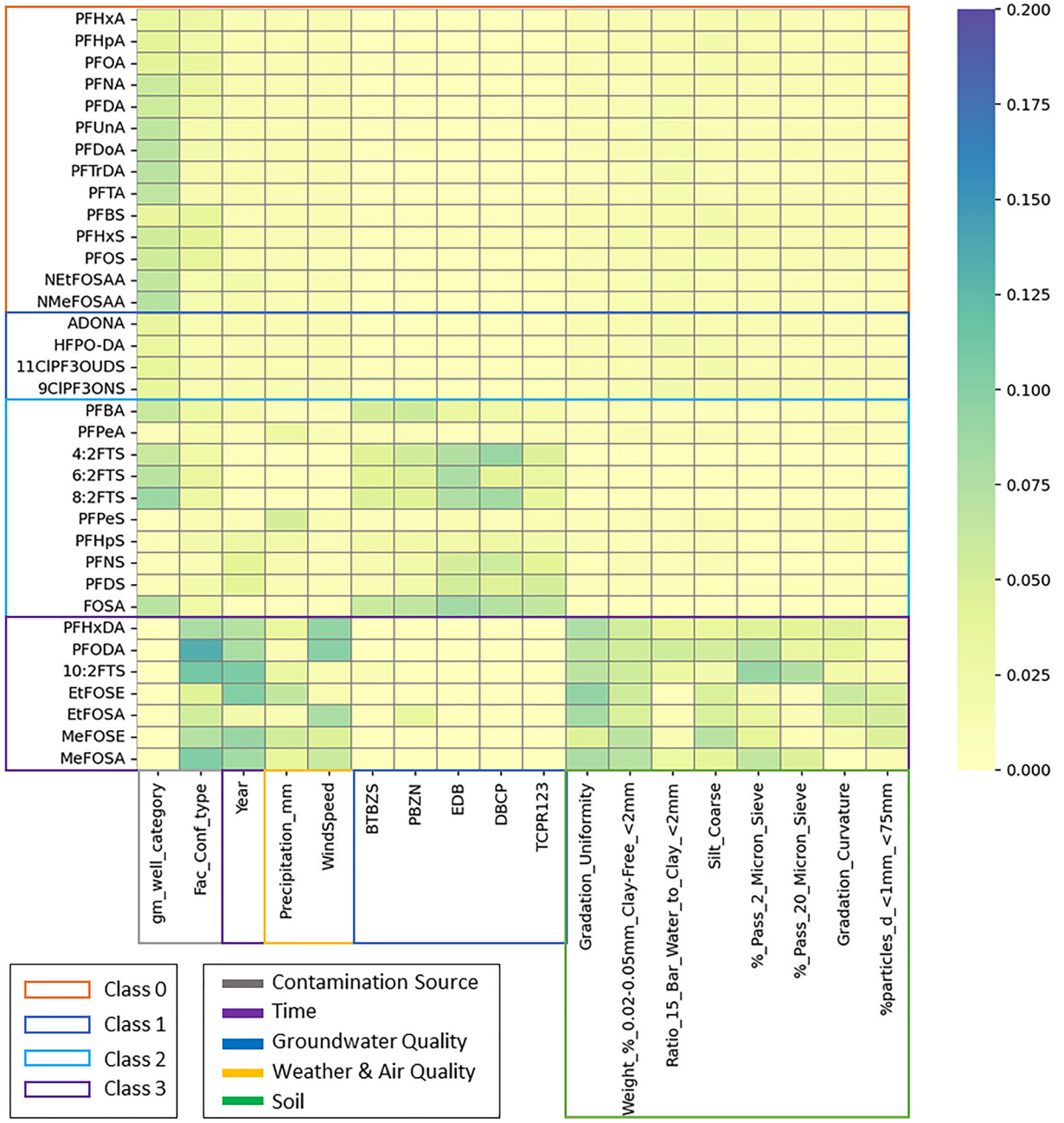


Figure 6. Relative contributions of the top 18 variables for each PFAS prediction model. gm_well_category represents the well type. Fac_Conf_type is the type of the nearest confirmed PFAS contamination facility. BTBZS is *sec*-butylbenzene. PBZN is n-propylbenzene. EDB is 1,2-dibromoethane. DBCP is 1,2-dibromo-3-chloropropane. TCPR123 is 1,2,3-trichloropropane. Gradation_Uniformity is the coefficient of uniformity. Weight_%_0.02−0.05mm_Clay-Free_<2 mm represents the gravimetric percent of particles with a 0.02−0.05 mm diameter (medium sand), reported on a clay free < 2 mm base. Ratio_15_Bar_Water_to_Clay_<2 mm is the 15 bar water percent/clay percent ratio on a < 2 mm base. Silt_Coarse is the soil separate with a 0.02 to 0.05 mm particle size. %_Pass_2_Micron_Sieve is the cumulative gravimetric percentage of particles with diameters < 2 μm (0.002 mm), reported on a < 3 in. base. %_Pass_20_Micron_Sieve is the cumulative gravimetric percentage of particles with diameters < 20 μm (0.02 mm), reported on a < 3 in. base. Gradation_Curvature is a descriptive parameter calculated from the cumulative grain size distribution curve to evaluate grading characteristics of coarse materials. %particles_d_<1mm_<75mm is the cumulative gravimetric percentage of particles with a < 1.0 mm diameter, reported on a < 3 in. base. Additional details on each variable are shown in SI Table S2.

has a high concentration of PFASs (SI Figure 2). The PFAS profile in the cleanup sites may differ from other sites. To improve the generalizability of the model, more data sourced from different types of sites and some missing important features' information (e.g., groundwater depth) are needed. Moreover, more data should be collected for PFASs with roughly one hundred observations in the current data set (Class 3).

3.4. Environmental Variables Influencing PFAS Prediction in Groundwater. Overall, contamination sources had the highest influence on total and individual PFAS prediction (Figure 5 and Figure 6), which is consistent with a previous

https://doi.org/10.1021/acsestwater.3c00134
ACS EST Water XXXX, XXX, XXX−XXX

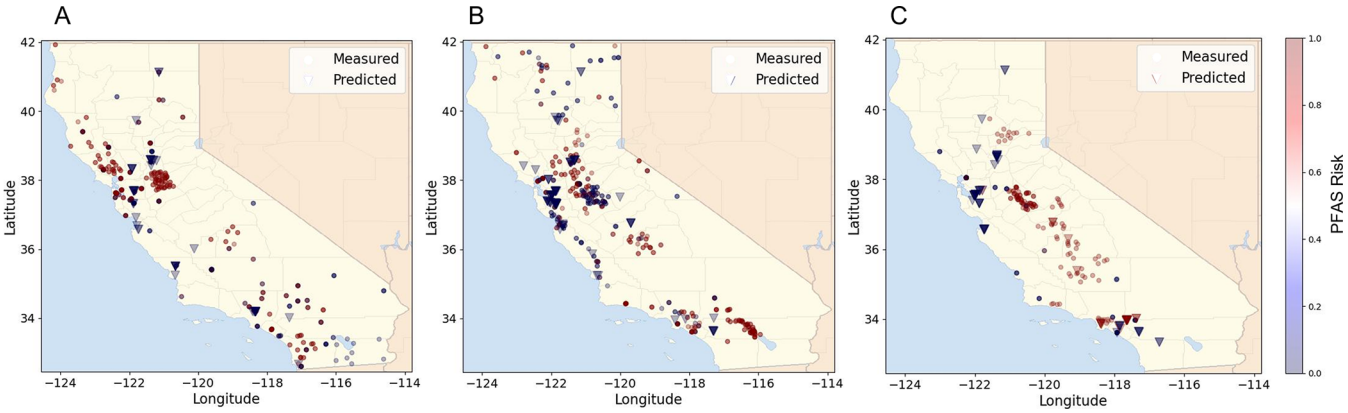


Figure 7. Reconstruction of PFAS recent maps in California groundwater. (A) The distribution map of the perfluoropentanoic acid (PFPeA) contamination risk in all groundwater wells in 2019. (B) The distribution map of the PFPeA contamination risk in all groundwater wells in 2020. (C) The distribution map of the 4:2 fluorotelomer sulfonate (FTS) contamination risk in all groundwater wells in 2021. Circles depict the wells that measured PFASs. Triangles depict the prediction results by the multilabel PFAS semisupervised learning model for the unmeasured groundwater wells. The color of the points represents the possibility of the well detecting PFASs above the notification limit (2 ng/L) during the year (0 lowest - 1 highest), calculated by the number of samples tested or predicted above 2 ng/L divided by the number of samples at this well during the year.

study.[25] The soil particle size distribution also showed as a top important variable for total and individual PFAS prediction.

For total PFAS risk prediction, we found that the number of facilities/industries within 1−50 km was more important than the type and the distance from the nearest PFAS confirmed source. This suggests that the number of facilities using or producing PFASs is currently underestimated. Therefore, future PFAS investigations or regulations should be extended to a larger group of industries to monitor PFAS releases. The influence of the number of facilities within the largest radii (10 and 50 km), expected air emission ranges of PFASs,[28] was as important as the number of facilities within short distances (1 and 3 km), expected from PFAS infiltration into groundwater. The weather and air quality variables (humidity, particulate matter (PM)10, NO2, SO2, PM 2.5 in decreasing order of influence) were as influential as soil and groundwater quality (Figure 5 and SI Table S9). Previous research suggests PFASs can be transported with other air pollutants and deposited to the soil, becoming a new source that continually introduces PFASs to the groundwater in the long term.[15,18,28] Unexpectedly, only one groundwater co-contaminant (nitrate) showed in the top variables. More research is needed to understand the extent to which air and soil variables impact PFAS groundwater contamination at different ranges from the pollution source. In addition, detailed mechanistic studies of air deposition of PFASs and its links to groundwater pollution are needed beyond statistical associations between air pollution and facility/industry PFAS emissions and discharge[60] (SI Figure S5 and Figure S6).

For individual PFASs, the top important variables vary based on the number of observations for each PFAS (Figure 6). No single type of variable was especially relevant for the prediction of the most frequently detected PFASs (>21,000 observations, Classes 0 and 1). For PFASs detected with less frequency, <1000 observations (Class 2 and 3), the nearest facility type was more important than the number of facilities within different radii. Groundwater co-contaminants, especially chlorinated solvents and aromatic hydrocarbons, influence the prediction of less frequently detected PFASs, especially FTSs and FOSA (Figure 6). Our findings suggest that PFASs that are less frequently detected may serve as indicators of specific pollution sources. To identify the PFAS risk and signature from different industries, more PFAS emissions and cocontaminant data from different industries should be collected.

3.5. Reconstruction of Recent California PFAS Groundwater Contamination Map. To fill in the gap in the PFAS investigation in previous years, we applied our model to reconstruct the PFAS contamination in California groundwater. The probability of annual measured PFASs higher than 2 ng/L and the probability of predicted PFASs above 2 ng/L for unmeasured wells are shown in Figure 7 and SI Figure S7. Most of the unmeasured groundwater wells in 2019 and 2020 were found to be at low risk of PFASs. This suggests that previous PFAS investigations covered most groundwater well locations with PFAS contamination. However, some predicted wells had a high risk of PFAS concentrations for 4:2 FTS (Figure 7 Panel C), indicating a high possibility of measuring this PFAS in some locations near Los Angeles which were not included in the PFAS investigation plan. Locations with a high risk PFAS detection overlap between different years and between different PFASs, supporting our correlation analysis that shows that individual PFASs within a subfamily tend to cooccur. Locations with high PFAS risk were concentrated in the San Francisco Bay Area and Southern California. The PFAS distribution pattern is similar to the distribution of industrial facilities (SI Figure S8), consistent with our previous finding that the number of facilities is an important factor in the prediction of PFASs.

4. IMPLICATIONS

Given the current lack of comprehensive large-scale environmental data sets (including pollution sources, contaminant transport and transformation-related information, and co-contaminants), we believe that the benchmark PFAS data set CA-PFAS-ASGWS (over 26,000 observations) will facilitate the development of ML-based pollutant prediction methods, not only for PFASs but also for other pollutants in the data set. We proposed a general ML pipeline for multicontaminant prediction using semisupervised learning, overcoming the missing data/label problem in environmental data sets. This approach makes it possible to predict the occurrence of

https://doi.org/10.1021/acsestwater.3c00134
ACS EST Water XXXX, XXX, XXX−XXX

contaminants lacking measured data. Our result suggests that ML can help us to understand contaminant fate and transport using domain knowledge from a particular field (i.e., contaminant physicochemical properties and transformation pathways) and vice versa. Here, the classifier chain algorithm accurately predicted correlated PFASs. Moreover, weather and air quality were unexpectedly important variables influencing groundwater PFAS prediction, which were not included in previous studies. We suggest that PFAS flux studies in the environment not only consider soil and water matrices but also include air emissions and deposition.[15,18,28] Applying our algorithm to reconstruct the groundwater PFAS map in past years, we validate that previous PFAS investigation plans have likely covered most hot spots of PFAS contamination in California, but there are still some underestimated PFASs (i.e., 4:2 FTS).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsestwater.3c00134.

Summaries of source data sets, CA-PFAS-ASGWS data set data dictionary and statistical summary, ML algorithm methods and description of multilabel classification evaluation metrics, ML model evaluations, correlations, and figures of applications of ML model for recent PFAS map reconstruction (PDF)

CA-PFAS-ASGWS data set (CSV)

■ AUTHOR INFORMATION

Corresponding Author

Christopher I. Olivares − *Department of Civil and Environmental Engineering, University of California, Irvine, California 92697, United States;* orcid.org/0000-0001-6213-7158; Email: chris.olivares@uci.edu

Authors

Jialin Dong − *Department of Civil and Environmental Engineering, University of California, Irvine, California 92697, United States*

Gabriel Tsai − *Department of Civil and Environmental Engineering, University of California, Irvine, California 92697, United States; Sage High School, Newport Beach, California 92657, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsestwater.3c00134

Author Contributions

CRediT: **Jialin Dong** conceptualization (equal), data curation (equal), formal analysis (equal), investigation (equal), methodology (equal), software (equal), writing-original draft (equal); **Gabriel Tsai** data curation (supporting), investigation (supporting); **Christopher I Olivares** conceptualization (equal), formal analysis (equal), funding acquisition (equal), investigation (equal), project administration (equal), resources (equal), supervision (equal), visualization (equal), writing-review & editing (equal).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was supported by setup funds for C.I.O.

■ REFERENCES

(1) Simon, J. A.; Abrams, S.; Bradburne, T.; Bryant, D.; Burns, M.; Cassidy, D.; Cherry, J.; Chiang, S.-Y. D.; Cox, D.; Crimi, M.; Denly, E.; DiGuiseppi, B.; Fenstermacher, J.; Fiorenza, S.; Guarnaccia, J.; Hagelin, N.; Hall, L.; Hesemann, J.; Houtz, E.; Koenigsberg, S. S.; Lauzon, F.; Longsworth, J.; Maher, T.; McGrath, A.; Naidu, R.; Newell, C. J.; Parker, B. L.; Singh, T.; Tomiczek, P.; Wice, R. PFAS Experts Symposium: Statements on Regulatory Policy, Chemistry and Analytics, Toxicology, Transport/fate, and Remediation for Per- and Polyfluoroalkyl Substances (PFAS) Contamination Issues. *Remediation* **2019**, *29* (4), 31−48.

(2) Sunderland, E. M.; Hu, X. C.; Dassuncao, C.; Tokranov, A. K.; Wagner, C. C.; Allen, J. G. A Review of the Pathways of Human Exposure to Poly- and Perfluoroalkyl Substances (PFASs) and Present Understanding of Health Effects. *Journal of Exposure Science & Environmental Epidemiology* **2019**, *29*, 131−147.

(3) Prevedouros, K.; Cousins, I. T.; Buck, R. C.; Korzeniowski, S. H. Sources, Fate and Transport of Perfluorocarboxylates. *Environ. Sci. Technol.* **2006**, *40* (1), 32−44.

(4) Podder, A.; Sadmani, A. H. M. A.; Reinhart, D.; Chang, N.-B.; Goel, R. Per and Poly-Fluoroalkyl Substances (PFAS) as a Contaminant of Emerging Concern in Surface Water: A Transboundary Review of Their Occurrences and Toxicity Effects. *J. Hazard. Mater.* **2021**, *419*, 126361.

(5) Brusseau, M. L.; Anderson, R. H.; Guo, B. PFAS Concentrations in Soils: Background Levels versus Contaminated Sites. *Sci. Total Environ.* **2020**, *740*, 140017.

(6) Pike, K. A.; Edmiston, P. L.; Morrison, J. J.; Faust, J. A. Correlation Analysis of Perfluoroalkyl Substances in Regional U.S. Precipitation Events. *Water Res.* **2021**, *190*, 116685.

(7) Epa, U. S. Lifetime Health Advisories and Health Effects Support Documents for Perfluorooctanoic Acid and Perfluorooctane Sulfonate. *Fed. Regist.* 2016.

(8) Us Epa, O. A. Biden-Harris Administration Proposes First-Ever National Standard to Protect Communities from PFAS in Drinking Water. 2023.

(9) Crone, B. C.; Speth, T. F.; Wahman, D. G.; Smith, S. J.; Abulikemu, G.; Kleiner, E. J.; Pressman, J. G. Occurrence of per- and Polyfluoroalkyl Substances (PFAS) in Source Water and Their Treatment in Drinking Water. *Critical Reviews in Environmental Science and Technology* **2019**, *49*, 2359−2396.

(10) Carle, D. *Introduction to Water in California*; Univ of California Press: 2015.

(11) Choi, Y. J.; Helbling, D. E.; Liu, J.; Olivares, C. I.; Higgins, C. P. Microbial Biotransformation of Aqueous Film-Forming Foam Derived Polyfluoroalkyl Substances. *Sci. Total Environ.* **2022**, *824*, 153711.

(12) Brusseau, M. L.; Yan, N.; Van Glubt, S.; Wang, Y.; Chen, W.; Lyu, Y.; Dungan, B.; Carroll, K. C.; Holguin, F. O. Comprehensive Retention Model for PFAS Transport in Subsurface Systems. *Water Res.* **2019**, *148*, 41−50.

(13) Liu, J.; Mejia Avendaño, S. Microbial Degradation of Polyfluoroalkyl Chemicals in the Environment: A Review. *Environ. Int.* **2013**, *61*, 98−114.

(14) Cousins, I. T.; Johansson, J. H.; Salter, M. E.; Sha, B.; Scheringer, M. Outside the Safe Operating Space of a New Planetary Boundary for Per- and Polyfluoroalkyl Substances (PFAS). *Environ. Sci. Technol.* **2022**, *56* (16), 11172−11179.

(15) D'Ambro, E. L.; Pye, H. O. T.; Bash, J. O.; Bowyer, J.; Allen, C.; Efstathiou, C.; Gilliam, R. C.; Reynolds, L.; Talgo, K.; Murphy, B. N. Characterizing the Air Emissions, Transport, and Deposition of Per- and Polyfluoroalkyl Substances from a Fluoropolymer Manufacturing Facility. *Environ. Sci. Technol.* **2021**, *55* (2), 862−870.

(16) Borthakur, A.; Leonard, J.; Koutnik, V. S.; Ravi, S.; Mohanty, S. K. Inhalation Risks of Wind-Blown Dust from Biosolid-Applied Agricultural Lands: Are They Enriched with Microplastics and PFAS? *Current Opinion in Environmental Science & Health* **2022**, *25*, 100309.

(17) Weber, A. K.; Barber, L. B.; LeBlanc, D. R.; Sunderland, E. M.; Vecitis, C. D. Geochemical and Hydrologic Factors Controlling

Subsurface Transport of Poly- and Perfluoroalkyl Substances, Cape Cod, Massachusetts. *Environ. Sci. Technol.* **2017**, *51* (8), 4269−4279.
(18) Guelfo, J. L.; Korzeniowski, S.; Mills, M. A.; Anderson, J.; Anderson, R. H.; Arblaster, J. A.; Conder, J. M.; Cousins, I. T.; Dasu, K.; Henry, B. J.; Lee, L. S.; Liu, J.; McKenzie, E. R.; Willey, J. Environmental Sources, Chemistry, Fate, and Transport of per- and Polyfluoroalkyl Substances: State of the Science, Key Knowledge Gaps, and Recommendations Presented at the August 2019 SETAC Focus Topic Meeting. *Environ. Toxicol. Chem.* **2021**, *40* (12), 3234−3260.
(19) Alizamir, M.; Sobhanardakani, S.; Shahrabadi, A. H. Prediction of Heavy Metals Concentration in the Groundwater Resources in Razan Plain: Extreme Learning Machine vs. Artificial Neural Network and Multivariate Adaptive Regression Spline. *Annals of Military and Health Sciences Research* **2019**, *17*, e98554.
(20) Rahmati, O.; Choubin, B.; Fathabadi, A.; Coulon, F.; Soltani, E.; Shahabi, H.; Mollaefar, E.; Tiefenbacher, J.; Cipullo, S.; Ahmad, B. B.; Tien Bui, D. Predicting Uncertainty of Machine Learning Models for Modelling Nitrate Pollution of Groundwater Using Quantile Regression and UNEEC Methods. *Sci. Total Environ.* **2019**, *688*, 855−866.
(21) Ling, Y.; Podgorski, J.; Sadiq, M.; Rasheed, H.; Eqani, S. A. M. A. S.; Berg, M. Monitoring and Prediction of High Fluoride Concentrations in Groundwater in Pakistan. *Sci. Total Environ.* **2022**, *839*, 156058.
(22) George, S.; Dixit, A. A Machine Learning Approach for Prioritizing Groundwater Testing for per-and Polyfluoroalkyl Substances (PFAS). *J. Environ. Manage.* **2021**, *295*, 113359.
(23) 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*, 126593.
(24) Li, R.; MacDonald Gibson, J. Predicting the Occurrence of Short-Chain PFAS in Groundwater Using Machine-Learned Bayesian Networks. *Front. Environ. Sci. Eng.* **2022**, *10*, 958784.
(25) Hu, X. C.; Ge, B.; Ruyle, B. J.; Sun, J.; Sunderland, E. M. A Statistical Approach for Identifying Private Wells Susceptible to Perfluoroalkyl Substances (PFAS) Contamination. *Environ. Sci. Technol. Lett.* **2021**, *8* (7), 596−602.
(26) McMahon, P. B.; Tokranov, A. K.; Bexfield, L. M.; Lindsey, B. D.; Johnson, T. D.; Lombard, M. A.; Watson, E. Perfluoroalkyl and Polyfluoroalkyl Substances in Groundwater Used as a Source of Drinking Water in the Eastern United States. *Environ. Sci. Technol.* **2022**, *56* (4), 2279−2288.
(27) Xu, B.; Liu, S.; Zhou, J. L.; Zheng, C.; Weifeng, J.; Chen, B.; Zhang, T.; Qiu, W. PFAS and Their Substitutes in Groundwater: Occurrence, Transformation and Remediation. *J. Hazard. Mater.* **2021**, *412*, 125159.
(28) Schroeder, T.; Bond, D.; Foley, J. PFAS Soil and Groundwater Contamination via Industrial Airborne Emission and Land Deposition in SW Vermont and Eastern New York State, USA. *Environ. Sci. Process. Impacts* **2021**, *23* (2), 291−301.
(29) Borthakur, A.; Olsen, P.; Dooley, G. P.; Cranmer, B. K.; Rao, U.; Hoek, E. M. V.; Blotevogel, J.; Mahendra, S.; Mohanty, S. K. Dry-Wet and Freeze-Thaw Cycles Enhance PFOA Leaching from Subsurface Soils. *Journal of Hazardous Materials Letters* **2021**, *2*, 100029.
(30) Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*; 2009; pp 248−255.
(31) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.
(32) Zhao, S.; Yuan, X.; Xiao, D.; Zhang, J.; Li, Z. AirNet: A Machine Learning Dataset for Air Quality Forecasting, 2018. https://openreview.net/pdf?id=SkymMAxAb (accessed 2023-01-07).
(33) Newell, C. J.; Adamson, D. T.; Kulkarni, P. R.; Nzeribe, B. N.; Connor, J. A.; Popovic, J.; Stroo, H. F. Monitored Natural Attenuation to Manage PFAS Impacts to Groundwater: Scientific Basis. *Ground Water Monit. Remediat.* **2021**, *41* (4), 76−89.
(34) US Environmental Protection Agency. Fifth Unregulated Contaminant Monitoring Rule (UCMR5). https://www.epa.gov/dwucmr/fifth-unregulated-contaminant-monitoring-rule (accessed 2023-01-11).
(35) Zhu, X.; Goldberg, A. B. Introduction to Semi-Supervised Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **2009**, *3* (1), 1−130.
(36) Lee, D.-H. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. 2013.
(37) 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. *Environ. Sci. Technol.* **2022**, *56* (7), 4080−4090.
(38) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-Sampling Technique. *jair* **2002**, *16*, 321−357.
(39) He, H.; Bai, Y.; Garcia, E. A.; Li, S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*; 2008; pp 1322−1328.
(40) Tsoumakas, G.; Katakis, I. Multi-Label Classification: An Overview. https://intelligence.csd.auth.gr/wp-content/uploads/2019/03/tsoumakas-ijdwm.pdf (accessed 2022-11-30).
(41) Tsoumakas, G.; Katakis, I.; Vlahavas, I. A review of multi-label classification methods. In *Proceedings of the 2nd ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD 2006)*; 2006.
(42) ElKafrawy, P.; Mausad, A.; Esmail, H. Experimental Comparison of Methods for Multi-Label Classification in Different Application Domains. *Int. J. Comput. Appl.* **2015**, *114* (19), 1−9.
(43) Mustafa, G.; Usman, M.; Yu, L.; Afzal, M. T.; Sulaiman, M.; Shahid, A. Multi-Label Classification of Research Articles Using Word2Vec and Identification of Similarity Threshold. *Sci. Rep.* **2021**, *11* (1), 21900.
(44) Tawiah, C. A.; Sheng, V. S. A Study on Multi-Label Classification. In *Advances in Data Mining. Applications and Theoretical Aspects*; Springer Berlin Heidelberg: 2013; pp 137−150.
(45) Zhou, L.; Zheng, X.; Yang, D.; Wang, Y.; Bai, X.; Ye, X. Application of Multi-Label Classification Models for the Diagnosis of Diabetic Complications. *BMC Med. Inform. Decis. Mak.* **2021**, *21* (1), 182.
(46) Lemaitre, G.; Nogueira, F.; Aridas, C. K. Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* **2017**, *18*, 1−5.
(47) EPA's Per- and Polyfluoroalkyl Substances (PFAS) Action Plan; EPA 823R18004 | ; U.S. Environmental Protection Agency, 2019. https://www.epa.gov/sites/default/files/2019-02/documents/pfas_action_plan_021319_508compliant_1.pdf (accessed 2023-08-09).
(48) PFAS Timeline. California State Water Resources Control Board. https://www.waterboards.ca.gov/pfas/ca_pfas_timeline.html (accessed 2022-12-29).
(49) Yi, S.; Harding-Marjanovic, K. C.; Houtz, E. F.; Gao, Y.; Lawrence, J. E.; Nichiporuk, R. V.; Iavarone, A. T.; Zhuang, W.-Q.; Hansen, M.; Field, J. A.; Sedlak, D. L.; Alvarez-Cohen, L. Biotransformation of AFFF Component 6:2 Fluorotelomer Thioether Amido Sulfonate Generates 6:2 Fluorotelomer Thioether Carboxylate under Sulfate-Reducing Conditions. *Environ. Sci. Technol. Lett.* **2018**, *5* (5), 283−288.

https://doi.org/10.1021/acsestwater.3c00134
ACS EST Water XXXX, XXX, XXX−XXX

(50) Harding-Marjanovic, K. C.; Houtz, E. F.; Yi, S.; Field, J. A.; Sedlak, D. L.; Alvarez-Cohen, L. Aerobic Biotransformation of Fluorotelomer Thioether Amido Sulfonate (Lodyne) in AFFF-Amended Microcosms. *Environ. Sci. Technol.* **2015**, *49* (13), 7666−7674.
(51) Cook, E. K.; Olivares, C. I.; Antell, E. H.; Yi, S.; Nickerson, A.; Choi, Y. J.; Higgins, C. P.; Sedlak, D. L.; Alvarez-Cohen, L. Biological and Chemical Transformation of the Six-Carbon Polyfluoroalkyl Substance N-Dimethyl Ammonio Propyl Perfluorohexane Sulfonamide (AmPr-FHxSA). *Environ. Sci. Technol.* **2022**, *56* (22), 15478−15488.
(52) Rodowa, A. E.; Knappe, D. R. U.; Chiang, S.-Y. D.; Pohlmann, D.; Varley, C.; Bodour, A.; Field, J. A. Pilot Scale Removal of per- and Polyfluoroalkyl Substances and Precursors from AFFF-Impacted Groundwater by Granular Activated Carbon. *Environmental Science: Water Research & Technology* **2020**, *6* (4), 1083−1094.
(53) D'Agostino, L. A.; Mabury, S. A. Certain Perfluoroalkyl and Polyfluoroalkyl Substances Associated with Aqueous Film Forming Foam Are Widespread in Canadian Surface Waters. *Environ. Sci. Technol.* **2017**, *51* (23), 13603−13613.
(54) Cheng, W.; Ng, C. A. Using Machine Learning to Classify Bioactivity for 3486 Per- and Polyfluoroalkyl Substances (PFASs) from the OECD List. *Environ. Sci. Technol.* **2019**, *53* (23), 13970−13980.
(55) 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*, 130124.
(56) Borisov, V.; Leemann, T.; Seßler, K.; Haug, J.; Pawelczyk, M.; Kasneci, G. Deep Neural Networks and Tabular Data: A Survey. *arXiv [cs.LG]*, 2021. http://arxiv.org/abs/2110.01889 (accessed 2023-08-09).
(57) Shenoy, A. Pseudo-Labeling to deal with small datasets — What, Why & How?. *Towards Data Science*. https://towardsdatascience.com/pseudo-labeling-to-deal-with-small-datasets-what-why-how-fd6f903213af (accessed 2023-01-04).
(58) Oliver, A.; Odena, A.; Raffel, C.; Cubuk, E. D.; Goodfellow, I. J. Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. *arXiv [cs.LG]*, 2018. http://arxiv.org/abs/1804.09170 (accessed 2023-08-09).
(59) Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; McGuinness, K. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. *arXiv [cs.CV]*, 2019. http://arxiv.org/abs/1908.02983 (accessed 2023-08-09).
(60) Streets, S.; Kvale, D. PFAS Air and Deposition Monitoring Report; tdr-g1-23; Minnesota Pollution Control Agency, 2022. https://www.pca.state.mn.us/sites/default/files/tdr-g1-23.pdf (accessed 2023-08-09).
(61) *Environmental Monitoring and Characterization*; Artiola, J. F., Pepper, I. L., Brusseau, M. L., Eds.; 2004.

https://doi.org/10.1021/acsestwater.3c00134
ACS EST Water XXXX, XXX, XXX−XXX