Environmental Science Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Environ. Sci.: Adv.*, 2024, **3**, 366

Multi-class machine learning classification of PFAS in environmental water samples: a blinded test of performance on unknowns†

Tohren C. G. Kibbey, *[a] Denis M. O'Carroll, [b] Andrew Safulko [c] and Greg Coyle[d]

The ability to identify the origin of detected PFAS in environmental samples is of great interest. This work used a blinded test to explore the ability of a recently-developed multiclass classification approach to classify unknown PFAS water samples based on composition. The approach was adapted from previous work to identify similarities between the patterns of unknown samples and classes defined by the compositions of samples from more than one hundred different PFAS data sources, in addition to making an overall assessment of whether PFAS is likely of AFFF or non-AFFF origin. Methods permitting the use of data with different subsets of analyzed PFAS components allowed for the use of a training dataset of more than 13 000 samples from a highly diverse range of sites. For this work, researchers at Brown and Caldwell (BC) provided a set of 252 unknown samples to researchers at The University of Oklahoma (OU) and The University of New South Wales (UNSW) for classification. Unknown samples were provided by clients of BC, and also included a number of artificial sample compositions created to test the ability of a rejection method to identify samples too unlike the training dataset for accurate classification. Unknown samples were de-identified and placed in random order prior to being sent to OU and UNSW researchers. Only after classification results had been sent by OU and UNSW researchers to BC researchers did BC provide the actual sample descriptions to OU and UNSW. Results showed extremely strong performance of the method, both in terms of its ability to identify similarities between unknown samples and samples of known origin, and its ability to make more subtle distinctions between sample origin, such as, for example, recognizing unknown samples from an airport wastewater collection system as being compositionally similar to known samples in another airport wastewater collection system. A rejection algorithm was tested and found to be able to identify artificial sample compositions as different from those in the training dataset, a critical feature of a practical supervised machine learning application, necessary to avoid misclassification of unknown samples that are unlike those in the training dataset.

Received 8th September 2023
Accepted 10th January 2024

DOI: 10.1039/d3va00266g

rsc.li/esadvances

Environmental significance

Per-and polyfluoroalkyl substances (PFAS) are ubiquitous environmental contaminants, frequently detected in environmental samples worldwide. The ability to determine the original source of PFAS in any given sample is of great interest, because the information could be used to focus remediation efforts to create the greatest potential benefit, as well as contribute to source identification and control efforts. This work explores the use of multiclass supervised machine learning for classification of water samples based on composition. The work was designed as a blinded test, where classifications were conducted on a test dataset whose origins were unknown to researchers conducting the classifications. Results show extreme promise for the ability of machine learning to recognize patterns in PFAS from a variety of sources.

Introduction

Per- and polyfluoroalkyl substances (PFAS) are ubiquitous environmental contaminants, frequently detected in environmental samples at sites around the world. Because of their favorable physicochemical properties, particularly in applications requiring interfacial activity, PFAS have been widely used since the mid-twentieth century in a wide range of industrial and consumer applications. Because many of the PFAS

[a]*School of Civil Engineering and Environmental Science, University of Oklahoma, Norman, OK 73019, USA. E-mail: kibbey@ou.edu*

[b]*School of Civil and Environmental Engineering, Water Research Centre, University of New South Wales, Sydney, NSW 2052, Australia*

[c]*Brown and Caldwell, Lakewood, Colorado 80401, USA*

[d]*Brown and Caldwell, Andover, Massachusetts 01810, USA*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3va00266g

compounds of regulatory concern are highly recalcitrant to degradation, they persist in the environment, and PFAS from five-plus decade old sites are regularly detected. The lack of degradation means that even strongly-adsorbing PFAS can exhibit substantial environmental mobility with time, and can be transported far from their original source. The use of PFAS over multiple decades and across so many different applications means that PFAS detected in a given environmental sample could potentially come from many different candidate sites.

The ability to determine the original source of PFAS in any given sample is of great interest, because the information could be used to focus remediation efforts to create the greatest potential benefit, as well as contribute to source identification and control efforts. Information about the most likely source could reduce site investigation costs, allowing for more rapid and targeted remediation efforts.

Previous work by the authors provided a preliminary investigation of the use of supervised machine learning for classification of PFAS, both in water[1,2] and non-water (*e.g.*, biota, soil, sediment)[3] samples, based on PFAS composition. That early work focused on the use of binary classification to distinguish between PFAS from AFFF (aqueous film-forming foam, used in fire suppression applications) and non-AFFF sources. The idea of identifying PFAS source by composition is made possible by the fact that hundreds of different PFAS components have been detected in the environment, and formulations used in different applications have made use of different combinations of PFAS components. The challenge of identifying source from composition comes from the fact that due to differential mobility and transformation of some PFAS precursors, PFAS composition can vary significantly in space, even at a site where a relatively narrow range of formulations is known to have been used.[1,2] The hypothesis driving the work was that although compositions resulting from any initial formulation can differ substantially from the original composition, the environmental behaviors that produce the different compositions (differential adsorption and transport of components, transformation of precursors) are the same everywhere, so a machine learning classifier trained to recognize the family of compositions resulting from a particular formulation will recognize that pattern wherever it exists. This hypothesis was strongly supported by the results of the work, which found that supervised machine learning exhibited great promise for distinguishing between AFFF and non-AFFF sources, even for difficult subsets of sample types.[2] Recent work by Stults *et al.*[4] testing supervised machine learning for PFAS source identification in fish found similarly promising results for multiclass classification (*i.e.*, distinguishing between PFAS from multiple source types).

The work described here uses an approach modified from methods used in earlier work to conduct simultaneous multiclass and binary classification of PFAS from unknown sources. The work involved training multiclass classifiers based on PFAS concentration data from 13 572 individual water samples, and then testing the ability of the classifiers to classify 252 unknown water samples. The machine learning components of the work were conducted by the authors at The University of Oklahoma (OU) and The University of New South Wales (UNSW) in a blinded test using unknowns provided by authors at Brown and Caldwell (BC) from their own completely separate client data sources. This paper describes the methods used both for classification, and for rejection of unknowns likely not represented by the training set, a critical aspect of any PFAS classification method to avoid misclassification of unknowns that are too different from those used to train the classifier.

Methods

Unknowns

Most previously-reported work exploring supervised machine learning classification of PFAS based on composition has relied on splitting datasets into training and test sets, using training sets to train machine learning classifiers, and then testing classifier performance on test sets. For example, in recent work, Kibbey *et al.*[2] split an 8040 sample dataset into two parts, training it on one 4020 sample subset, and using the trained classifier to classify the remaining 4020 samples. (Earlier preliminary work by the same authors tested the approach with a smaller, 1197 sample dataset.[1]) While this approach does provide insight into whether classifiers can identify patterns needed to classify PFAS sources, the fact that training and test samples come from the same sites means that it is possible that some of the observed classification performance results from similarities at a given site. While Kibbey *et al.*[2] found that removing specific sites being tested from the training set had little effect on classification performance in most cases, the use of test and training sets split from the same original dataset leaves open the question of how well classification will work on unknown samples from completely different sites.

In contrast to earlier work, this work was designed from the start as a blinded test of classification performance, and did not involve splitting a single dataset into test and training sets. Rather, researchers from BC assembled an unknown dataset containing a total of 252 sample compositions, and provided the unknown dataset to researchers from OU and UNSW for classification. Researchers from OU and UNSW had no knowledge of how many sites the unknowns were taken from, or what types of samples were included, beyond the vague understanding that the samples were largely provided by clients of BC. Sample data were provided to OU and UNSW in an Excel file, anonymized and placed in random order by BC. Only after the samples were classified by OU and UNSW and the classification results sent to BC, did BC provide details about the unknown sample data sources to OU and UNSW for analysis of classification performance.

One of the challenges with supervised machine learning classification is that without inclusion of separate rejection algorithms, unknowns will be assigned to a class, even if they are completely unlike anything in the dataset used to train classifiers. For this reason, the unknown sample data provided by BC also included a small number of artificial PFAS compositions, created by BC researchers, to allow testing of a rejection algorithm to identify data too different from the training dataset to allow accurate classification. OU and UNSW researchers

had no advance knowledge of the number of artificial sample compositions included in the unknown dataset.

The input file of unknowns used in this work is provided in the accompanying online ESI Section† in the form originally provided by BC, along with classification results from OU and UNSW researchers, and finally a file containing the corresponding details on each sample, as sent by BC after classification had been completed by OU and UNSW researchers.

Classification approach

All coding for this work was conducted in Python 3.10.9, using machine learning classifiers from Scikit-Learn version 1.2.1.[5] Classification of unknowns was conducted using the Random Forest classifier. The Random Forest classifier and its variants are methods that involve creation of an ensemble of decision trees.[6] Previous work found the Random Forest method to be among the best approaches for binary classification of PFAS samples into AFFF/non-AFFF classes,[2] so it was selected for this work. For this work, the method was used to make multiclass classifications (classification of unknowns into one of multiple classes), and then additional calculations were subsequently used to estimate the probability that each unknown sample was of AFFF or non-AFFF origin, as described below.

The multiclass approach used for this work is novel, in that the classes are the individual data sources in the training dataset, split into AFFF and non-AFFF fractions. For the training set used here, that results in 125 separate classes. The primary question to be answered by the multiclass classification is: What known site has samples that exhibit patterns most similar to those observed in each unknown sample? The advantage of this approach is that it provides insights into the possible origins of a particular environmental PFAS unknown, without being susceptible to errors in labels in the training set as in the case of binary classification algorithms previously studied. While multiclass classification is well-suited to recognizing patterns in PFAS from any origin, the fact is that, with the exception of AFFF-impacted sites, sites known with high certainty to have been impacted by a single PFAS source composition are relatively rare. As such, comprehensive, accurately-labeled environmental training data for many types of specific PFAS applications could be difficult to acquire. The multiclass approach used here sidesteps this problem, essentially reporting the known site or dataset that is most reminiscent of the patterns observed in each unknown sample.

For this work, the Random Forest method made use of hyperparameters determined through initial validation in earlier work.[1] Most critically, the method was used with 1000 estimators (separate trees in the ensemble fit to different subsets of the training set created by bootstrapping), and balanced class weighting. Internal testing against a small set of samples of known origin not in the training set prior to analysis of unknowns found that balanced weighting was essential for this method of multiclass classification based on more than a hundred classes of widely varying sizes, simply because without balanced weighting, larger classes had a disproportional impact on classification. Note that the need for balanced weighting also precluded the use of many other classifiers in this work. All classifier parameters beyond those mentioned above were default values for the Scikit-Learn version used; as with earlier work, classifications with the method were found to be highly insensitive to Random Forest parameters within reasonable ranges. Note that final classifications reported in this work were the result of averaged probabilities from ten separate classifications with different random number seeds, which are used to both scramble the training set prior to training, and as an input to the Random Forest method to randomize the creation of decision trees. This is important because, like many machine learning classifiers, the training of a Random Forest classifier can result in different models depending on the order of the training data.

In addition to using multiclass classification to identify training data subsets that most closely match unknown sample fingerprints, the work also used the cumulative multiclass probabilities to estimate the overall probability that each unknown sample was of AFFF origin, an approach that is quantitatively similar to the binary classification used previously by the authors.

Training dataset and preprocessing

Previous work by the authors used the concentrations of 8 (ref. 1) or 10 (ref. 2 and 3) PFAS components (*i.e.*, individual PFAS compounds in a mixed composition) as machine learning features. Supervised machine learning requires all data to have the same specific features (in this case, PFAS components), so the dataset sizes were limited by the subset of sample data that could be acquired with data for the same components. Because of changes in PFAS analytical methods over time, as well as the growing need to quantify more compounds, older datasets often contain data for fewer PFAS components, a fact that complicates their use for classification.

Methods of replacing missing data are known as imputation. While a number of different imputation methods are sometimes used to allow data sets with missing features to be used for supervised learning, it is important to emphasize that the validity of these methods for classification is more based on their effect on the overall classification behavior of a specific model, rather than any physical basis; that is, regardless of algorithm, it is impossible to determine the concentration of a PFAS component that was not quantified. Rather, imputation can allow data with missing features to be used in classification without overly skewing the resulting classification results. For this work, missing values (*i.e.*, PFAS components that were not analyzed) were replaced with zero concentration. Several other approaches were tested in preliminary internal testing (*i.e.*, prior to receipt of unknowns from BC), including the MIA ("missingness incorporated in attributes") method,[7,8] as well as two other noniterative approaches and one iterative approach. The MIA method, which is most suited to decision trees, involves creation of two new features for each feature containing missing data, one with missing data replaced with +inf, the other with missing data replaced with −inf. Noniterative methods tested based on the Scikit-Learn SimpleImputer

involved replacement of missing values with the mean or median value for that feature, while the experimental Scikit-Learn IterativeImputer, an iterative training method that tries to determine likely values based on other component values, was also tested. Ultimately, the replacement of missing values with zero concentrations appeared to produce the most predictable, consistent behavior in testing, as indicated by the ability to correctly identify classes for test samples of known origin when values are removed. The likely reason for this is that assigning a zero concentration to unmeasured components introduces bias more consistently than the other imputation methods available for comparison. The use of imputation allowed a much larger training dataset of 13 572 samples to be used compared with previous work, potentially increasing the types of data represented in the training set. Furthermore, the use of imputation allowed far more PFAS components to be considered as features than in previous work. In this work, a total of 30 PFAS components were considered as features – far more than the 8 (ref. 1) or 10 (ref. 2 and 3) in previous work. (A list of the components considered as features is included in the accompanying ESI Section.†) The benefits of the expanded training set and expanded number of components considered appear to outweigh the approximations introduced by imputation, although imputation always carries the risk that it will influence classification in some specific cases.

As was done in previous work by the authors,[1–3] all component concentrations below detection limits were replaced with zeroes in both the training dataset and the dataset containing the unknowns, an approach that is essentially equivalent to placing all non-detects into a single bin for each component. For a full discussion of the justification for and implications of this approach, see Kibbey *et al.*[1] Note that Stults *et al.*[4] used substitution with a value related to the detection limit with success; it is likely that supervised machine learning classification is relatively insensitive to the handling of non-detects due to the fact that PFAS component concentrations often vary over orders of magnitude.

For this work, a new normalization method was used, different from those used in previous work. Previous work[2] explored the use of component concentrations and mass fractions as features, both untransformed, and after logarithmic transformation. All transformations worked similarly well for Random Forest and related classifiers, but for some classifiers logarithmic scale transformation produced better results. In this work, the features are PFAS component concentrations, normalized to the maximum component concentration in each sample, *i.e.*, β in eqn (1):

$$\beta_{i,k} = \frac{C_{i,k}}{\max_{0 \le j < n}(C_{j,k})} \quad (1)$$

where i and j are indices of the individual components in sample k (n is the number of components). An advantage of this transformation is that the resulting values are not skewed by components that were not analyzed. For example, the highest concentration component in a sample will have a value of 1.0, and lower concentration components in the same sample will be scaled to the high concentration component. If two different analyses of the same sample analyze for different subsets of components, the values of β for measured components should be the same for both, provided the highest concentration component is analyzed in both cases (something that is frequently true). In contrast, if samples are transformed to mass fraction, all component values will be different, because the total measured mass will differ.

Specific data sources included in the training dataset are shown in Table 1. Full details for the data sources, including web links to original data, are included in the accompanying online ESI Section.† Note that with only a few exceptions, the data used to train the classifiers used in this work are publicly available on the Internet; in some cases, although data are public, they must be requested from the originating organization.

The data sources in Table 1 are broken down into high-certainty AFFF sources (military, non-military), high-certainty non-AFFF sources (coatings, metal plating, tannery, other), and mixed data sources (landfill, wastewater treatment plant, other). High-certainty data sources are those where an original source is known and highly likely to be the primary contributor to the detected PFAS in water samples. In contrast, mixed sources are those where there may be multiple original contributors, or where there is less certainty about the origin, for example when samples are low concentration surface water samples far from confirmed sources. In the cases of landfills and wastewater treatment plants, in particular, note that they may receive PFAS from a range of primary sources, and the mix of primary sources may differ entirely from one site to the next. As such, identifying a sample as similar to something found in landfill or wastewater treatment plant data is not the same as identifying a sample as belonging to a specific original source (*e.g.*, AFFF, metal plating). However, the ability to identify a specific data source where a similar PFAS fingerprint is observed may nevertheless provide useful clues to the origin of the PFAS in the unknown sample.

Autoclassification

Because the origins of the individual samples in mixed data sources (Table 1) are not known (and may, in fact, vary significantly from sample to sample), for purposes of estimating the probability that unknown samples are of AFFF origin, an autoclassification method was used to classify the individual samples from mixed data sources as likely AFFF or non-AFFF. The method iteratively removed each mixed source from the training set and classified its samples against all remaining samples. For this work, K-nearest neighbors classification[9] was used with $n = 15$ neighbors, and weighting calculated from the inverse of distance, a weighting method that favors more-similar samples. Using this approach, classifications converged in 6 iterations for the data set used here. It is important to note that the resulting AFFF/non-AFFF assignments made using this iterative procedure are approximate, and are primarily intended to allow samples from mixed sources to be used in determining the probability that unknown samples are of AFFF origin.

Table 1 Data sources used to train classifiers. Details for all sources are provided in the accompanying online ESI Section. GW = groundwater; SW = surface water; WWTP = wastewater treatment plant; LF = landfill; ON = onsite data; OFF = offsite data. Note that many data sources, such as landfills and wastewater treatment plants, correspond to sites that are not the original PFAS source, but rather accumulate PFAS from multiple original PFAS sources. As is described in the text, for this work, iterative autoclassification is used to estimate AFFF contributions

Data Source		Country	# AFFF	# non-AFFF	% AFFF
High certainty AFFF data sources					
Military	ALBATROSS_GW	AU	79	0	100%
	ALTUS_GW	US	36	0	100%
	AMBERLEY_OFF_GWSW	AU	88	0	100%
	AMBERLEY_ON_GWSW	AU	126	0	100%
	BANDIANA_OFF_GW	AU	11	0	100%
	BANDIANA_ON_GW	AU	41	0	100%
	BLAMEY_GW	AU	8	0	100%
	CAIRNS_GW	AU	110	0	100%
	CAIRNS_SW	AU	21	0	100%
	CALIFGAMA_GW (Military)	US	12	0	100%
	DARWIN_GW	AU	259	0	100%
	DND_Site-B_GW	CA	101	0	100%
	DND_Site-C_GW	CA	57	0	100%
	DND_Site-C_SW	CA	16	0	100%
	DND_Site-E_GW	CA	161	0	100%
	DND_Site-E_SW	CA	8	0	100%
	DND_Site-G_GW	CA	319	0	100%
	DND_Site-G_SW	CA	374	0	100%
	DND_Site-H_GW	CA	205	0	100%
	DND_Site-H_STORMWATER	CA	45	0	100%
	DND_Site-H_SW	CA	408	0	100%
	DND_Site-I_GW	CA	112	0	100%
	DND_Site-I_SW	CA	17	0	100%
	EASTSALE_ON_GW	AU	75	0	100%
	HOLSWORTHY_OFF_GW	AU	10	0	100%
	HOLSWORTHY_ON_GW	AU	32	0	100%
	JERVISBAY_GW	AU	60	0	100%
	JERVISBAY_SW	AU	114	0	100%
	JERVISBAY_TANK_SW	AU	7	0	100%
	LAVARACK_OFF_GW	AU	28	0	100%
	LAVARACK_OFF_SW	AU	61	0	100%
	LAVARACK_ON_GW	AU	58	0	100%
	LAVARACK_ON_SW	AU	38	0	100%
	OAKEY_OFF_GW	AU	57	0	100%
	OAKEY_ON_GW	AU	75	0	100%
	OAKEY_SW	AU	17	0	100%
	PEARCE_GW	AU	50	0	100%
	RICHMOND_GW	AU	69	0	100%
	ROBERTSON_DRY	AU	7	0	100%
	ROBERTSON_WET	AU	11	0	100%
	SINGLETON_OFF_GW	AU	14	0	100%
	SINGLETON_ON_GW	AU	41	0	100%
	STIRLING_GW	AU	471	0	100%
	STIRLING_SW	AU	28	0	100%
	TOWNSVILLE_OFF1_GW	AU	141	0	100%
	TOWNSVILLE_OFF2_GW	AU	27	0	100%
	TOWNSVILLE_ON_GW	AU	190	0	100%
	WAGGA_GW	AU	40	0	100%
	WILLIAMS_GW	AU	10	0	100%
	WILLIAMTOWN_GW	AU	473	0	100%
	WILLIAMTOWN_SW	AU	369	0	100%
Non-military	ALY_2020_SW	US	52	0	100%
	CLARENDON_GW_OFF	US	39	0	100%
	CLARENDON_GW_ON	US	6	0	100%
	CALIFGAMA_GW (Airport)	US	332	0	100%
	HAMILTON_AIRPORT_GWSW	CA	9	0	100%
	MARINETTE_OFF_GW	US	634	0	100%
	MARINETTE_ON_GWSW	US	72	0	100%
	PDX_GW	US	118	0	100%

Table 1 (*Contd.*)

Data Source		Country	# AFFF	# non-AFFF	% AFFF
	PDX_SW	US	24	0	100%
	QH3_CONCENTRATE	AU	28	0	100%
	QH3_GW	AU	33	0	100%
	QH3_SEWER	AU	168	0	100%
	QH3_SW	AU	179	0	100%
	QH3_WWTP	AU	348	0	100%
	STOCKHOLM-ARLANDA_GW	SE	26	0	100%
High-certainty non-AFFF data sources					
Coatings	BENNINGTON_GW	US	0	1042	0%
	CENTRE_SW	US	0	97	0%
	GADSDEN_SW	US	0	175	0%
Metal plating	CALIFGAMA_GW (metal plating)	US	0	182	0%
	DU-WEL_DBS_VAS_GW	US	0	14	0%
	DU-WEL_MW_GW	US	0	18	0%
	DU-WEL_RES_OFF_GW	US	0	53	0%
	DU-WEL_VAS_OFF_GW	US	0	40	0%
	DU-WEL_VAS_ON_GW	US	0	102	0%
Tannery	WOLVERINE_HS_GW	US	0	99	0%
	WOLVERINE_TA_GW	US	0	108	0%
	WOLVERINE_TA_SW	US	0	14	0%
Other	CAPEFEAR_SW	US	0	456	0%
	GOBELIUS_SKIING	SE	0	8	0%
Mixed data (AFFF/non-AFFF estimated by iterative autoclassification – see text)					
Landfill	BENSKIN_2012_LF_GW	CA	2	9	18%
	BUSCH_2010_LF_WWTP	DE	5	15	25%
	CALIFGAMA_GW (LF_MSW)	US	323	343	48%
	CALIFGAMA_GW (LF_Other)	US	27	7	79%
	FUERTES_2017_LF_GW	ES	2	4	33%
	GALLEN_2017_LF_GW	AU	77	20	79%
	GOBELIUS_LF_GWSW	SE	16	7	70%
	HARRAD_2019_LF_GW	IE	12	36	25%
	HEPBURN_2019_LF_GW	AU	10	3	77%
	HUSET_2011_LF_GW	US	1	5	17%
	LANG_2017_LF_GW	US	4	81	5%
	YAN_2015_LF_GW	CN	0	5	0%
WWTP	CALIFGAMA_GW (WWTP)	US	588	1036	36%
	VTWWTF_EFF_WWTP	US	13	114	10%
	VTWWTF_INF_WWTP	US	28	92	23%
	WANG_2020_WWTP	CN	20	13	61%
	YAN_2015_LF_WWTP	CN	1	19	5%
Other	CALIFGAMA_GW (CPS)	US	420	90	82%
	CALIFGAMA_GW (fuel/refinery)	US	137	12	92%
	CALIFGAMA_GW (industrial)	US	6	1	86%
	CALIFGAMA_GW (NPDES)	US	24	11	69%
	GOBELIUS_2018_FIRE_GWSW	SE	173	9	95%
	GOBELIUS_IND_GWSW	SE	73	15	83%
		Total	9217	4355	13572

Rejection

An important part of the work was the addition of a rejection algorithm to identify cases where unknown samples were too different from anything in the training dataset to be accurately identified; a risk with supervised learning is that without additional checks, classifiers will always assign unknowns to a known class, even in cases where the unknown is unlike anything in the training dataset. While having a large training dataset spanning as many sources as possible can increase the likelihood that an unknown will be represented in the training dataset – and so will be accurately classified – the risk always exists that a novel unknown sample will be misclassified. For this work, the sum of the square distance between each unknown (m) and each sample (k) in the identified class (n is the number of samples in the identified class) was calculated, and the minimum for each unknown determined:

$$\mathrm{ssd}_{\min,m} = \min_{0 \le k < n}\left[\sum_i \left(\beta_{i,k} - \beta_{i,m}\right)^2\right] \quad (2)$$

eqn (2) is essentially a measure of how similar the unknown fingerprint is to a known fingerprint, with lower values indicating a closer match. (A value of zero would indicate an exact match.). Values larger than approximately 3.0 correspond to substantial differences, and are a strong indicator that the unknown sample is not actually a match for the assigned class; samples for which the value for $\mathrm{ssd}_{\min}$ corresponding to the highest probability class exceeds this value were marked as UNLIKE TRAINING SET. Thresholds of 0.5 and 2.0 were used to indicate LOW and VERY LOW CERTAINTY classifications, respectively – indicators that the most similar training example within the highest probability class is a relatively weak match for the unknown sample, but the classification may still be correct. Note that these thresholds were selected based on internal testing prior to classification of unknown samples from BC. Internal testing involved classification of multiple artificial samples generated by OU and UNSW researchers (generated using a range of methods different from those ultimately used by BC), and observation of the ranges of magnitudes of $\mathrm{ssd}_{\min}$ for those samples in comparison with actual internal test samples from known sources, but not in the training dataset. For the purposes of this work, the actual selected threshold values were identified based on subjective observation of typical $\mathrm{ssd}_{\min}$ values in internal test results; it was anticipated that ultimate application of the rejection algorithm to future unknown samples might require tuning of these thresholds.

As a part of this work, BC included a number of artificial sample compositions in the unknown data, to provide a test of the ability of this simple rejection approach to identify samples not in the training dataset. The number of artificial samples included was unknown to OU and UNSW authors.

Results and discussion

The test dataset of 252 unknowns provided by BC ultimately included 230 individual sample compositions from a total of 10 specific sites, as well as 22 artificial sample compositions created by BC using two different methods. The sites included two airports, one industrial site, municipal wastewater influent at seven municipal wastewater treatment plants, six from one utility, and one from another. As mentioned previously, the data from all sites were scrambled in the file sent to OU and UNSW researchers; for purposes of discussion, unknown samples are sorted into the individual known categories in the subsequent sections.

Tables 2–4 show the classification results for the three sites where AFFF is expected to be the dominant contributor to PFAS contamination. Tables 2 and 3 correspond to the two airports, while Table 4 corresponds to an industrial site where AFFF was used to extinguish a fire. Each row in each table corresponds to a sample from the unknown dataset. The Test ID is the identification code provided to OU and UNSW researchers by BC researchers, while the Plot ID is a number corresponding to the table order of samples; all plots generated during classification have been included with the accompanying online ESI Section,† and were renamed following classification to include both the Plot ID and the Test ID. In addition to the site type, description, and sample date, the table also indicates the number of components in the unknown sample for which concentrations are above detection limits (NNZ; number nonzero). Classification results shown include the class identified from the training dataset as the most like the unknown (C_1), the $\mathrm{SSD}_{\min}$ value for that class (eqn (2)), a certainty flag indicating the likelihood that the unknown may not be represented in the training dataset, and a calculated overall probability that the sample is of AFFF origin, determined from the sum of the resulting Random Forest probabilities for the classes in the training set that are categorized as being of AFFF origin, as described in the Classification approach section. Note that full details of the classification results for all unknown samples are included in the ESI Section,† including the assigned random forest probabilities for all 125 classes, as well as the calculated $\mathrm{SSD}_{\min}$ values (eqn (2)) for the top three classes identified for each unknown sample, and full β distributions (eqn (1)) for each of the unknowns, as well as the closest three samples within each of the top three classes (C_1, C_2, C_3) for each unknown. Finally, plots comparing β distributions for each of the 252 unknowns with the closest samples within the top three identified classes for each unknown are included.

From Tables 2–4, it is apparent that the vast majority of unknown samples from the two airports and the industrial site are identified as being similar to AFFF-associated classes (indicated by "(A)" in the class name). In the case of Airport 1, the top class matches for 38 of the 44 unknown samples are AFFF-associated classes, and in 41 of 44 cases (93%) at least two of the top three matches are AFFF-associated classes (ESI†). Furthermore, 43 of 44 (97.7%) samples have P_{AFFF} (the estimated probability that the sample is of AFFF origin) greater than 0.5, and the one site below 0.5 is only slightly below it, at 0.47. In the case of Airport 2, the top class matches for 95 of the 109 samples are AFFF-associated classes, and in 106 of 109 cases (97%) at least two of the top three matches are AFFF-associated classes (ESI†). Furthermore, all 109 (100%) samples have P_{AFFF} greater than 0.5. In the case of Industrial Site 1, the top site matches for all 45 of the 45 samples are AFFF-associated classes, and all 45 (100%) samples have P_{AFFF} greater than 0.5.

It is interesting to note how the types of samples at the two airports and one industrial site are captured in the classifications. In the case of Airport 1, samples cover a range of wastewater samples collected from a central lift station, as well as stormwater and wastewater samples from holding ponds. Note that 13 of the Airport 1 samples are identified as being similar to either sewer samples from the Brisbane, Australia airport (the QH3_Airport_SEWER classification), or wastewater treatment plant samples from the Brisbane, Australia airport (the QH3_Airport_WWTP classification), meaning that these samples at Airport 1 are reminiscent not just of AFFF samples, but of AFFF samples specifically associated with an airport wastewater collection system. (For full descriptions of all of the sources in the classifications in Tables 2–6, see the list of Training Dataset Sources in the ESI.†)

Table 2 Classification of data from Airport 1[a]

Plot ID	Test ID	Site	Description	Sample date	NNZ	C_1	SSD_1	Certainty flag	P_{AFFF} overall
A1-001	T-121	Airport 1	Combined wastewater discharge	18/04/2019	13	CALIFGAMA_Fuel/Refinery_GW_(A)	0.821	LOW CERTAINTY	78%
A1-002	T-25	Airport 1	Combined wastewater discharge	26/08/2019	14	CALIFGAMA_Fuel/Refinery_GW_(A)	1.343	LOW CERTAINTY	81%
A1-003	T-190	Airport 1	Combined wastewater discharge	02/10/2019	11	QH3_Airport_SEWER_(A)	0.017		94%
A1-004	T-36	Airport 1	Combined wastewater discharge	10/01/2020	13	QH3_Airport_SEWER_(A)	0.297		80%
A1-005	T-245	Airport 1	Combined wastewater discharge	16/06/2020	13	DND_Site-G_Military_SW_(A)	1.005	LOW CERTAINTY	74%
A1-006	T-75	Airport 1	Combined wastewater discharge	07/07/2020	14	DND_Site-G_Military_SW_(A)	0.298		80%
A1-007	T-24	Airport 1	Combined wastewater discharge	15/10/2020	15	CALIFGAMA_Fuel/Refinery_GW_(A)	0.768	LOW CERTAINTY	80%
A1-008	T-249	Airport 1	Combined wastewater discharge	19/01/2021	8	GOBELIUS_Skiing_(nA)	0.349		67%
A1-009	T-92	Airport 1	Combined wastewater discharge	04/05/2021	14	DND_Site-G_Military_SW_(A)	0.288		82%
A1-010	T-165	Airport 1	Combined wastewater discharge	16/09/2021	9	QH3_Airport_SEWER_(A)	0.151		93%
A1-011	T-154	Airport 1	Combined wastewater discharge	16/11/2021	10	QH3_Airport_SEWER_(A)	0.262		86%
A1-012	T-133	Airport 1	Combined wastewater discharge	04/01/2022	10	QH3_Airport_SEWER_(A)	0.040		96%
A1-013	T-232	Airport 1	Combined wastewater discharge	06/05/2022	13	DND_Site-G_Military_SW_(A)	0.066		88%
A1-014	T-100	Airport 1	Combined wastewater discharge	29/08/2022	15	DND_Site-G_Military_SW_(A)	0.384		75%
A1-015	T-99	Airport 1	Combined wastewater discharge	29/08/2022	11	QH3_Airport_WWTP_(A)	0.203		80%
A1-016	T-68	Airport 1	Combined wastewater discharge	29/08/2022	12	CALIFGAMA_WWTP_GW_(nA)	1.003	LOW CERTAINTY	76%
A1-017	T-206	Airport 1	Combined wastewater discharge	29/08/2022	12	QH3_Airport_WWTP_(A)	0.201		81%
A1-018	T-93	Airport 1	Combined wastewater discharge	29/08/2022	12	QH3_Airport_SEWER_(A)	0.207		86%
A1-019	T-19	Airport 1	Combined wastewater discharge	30/08/2022	15	CALIFGAMA_Fuel/Refinery_GW_(A)	0.546	LOW CERTAINTY	77%
A1-020	T-169	Airport 1	Combined wastewater discharge	20/10/2022	14	DND_Site-G_Military_SW_(A)	0.240		87%
A1-021	T-88	Airport 1	Combined wastewater discharge	10/02/2023	13	DND_Site-G_Military_SW_(A)	0.315		81%
A1-022	T-41	Airport 1	Industrial stormwater pond	15/11/2021	11	CALIFGAMA_WWTP_GW_(nA)	0.389		62%
A1-023	T-139	Airport 1	Industrial stormwater pond	21/12/2021	15	CALIFGAMA_Fuel/Refinery_GW_(A)	0.432		69%
A1-024	T-179	Airport 1	Industrial stormwater pond	05/05/2022	11	MARINETTE_ON_AFFF-Mfg_GWSW_(A)	0.778	LOW CERTAINTY	65%
A1-025	T-70	Airport 1	Industrial stormwater pond	15/11/2021	13	DND_Site-G_Military_SW_(A)	0.692	LOW CERTAINTY	73%
A1-026	T-197	Airport 1	Industrial stormwater pond	21/12/2021	15	CALIFGAMA_Fuel/Refinery_GW_(A)	0.693	LOW CERTAINTY	90%
A1-027	T-9	Airport 1	Industrial stormwater pond	05/05/2022	10	DND_Site-G_Military_SW_(A)	0.979	LOW CERTAINTY	73%
A1-028	T-3	Airport 1	Industrial stormwater pond	15/11/2021	6	CALIFGAMA_WWTP_GW_(A)	0.177		79%
A1-029	T-80	Airport 1	Industrial stormwater pond	21/12/2021	17	QH3_Airport_WWTP_(A)	0.066		89%
A1-030	T-226	Airport 1	Industrial stormwater pond	05/05/2022	19	QH3_Airport_SEWER_(A)	0.543	LOW CERTAINTY	89%
A1-031	T-223	Airport 1	Industrial stormwater pond	21/12/2021	18	QH3_Airport_SEWER_(A)	0.287		87%
A1-032	T-209	Airport 1	Industrial stormwater pond	05/05/2022	15	DND_Site-G_Military_SW_(A)	0.731	LOW CERTAINTY	77%
A1-033	T-125	Airport 1	Industrial stormwater pond	15/11/2021	14	GOBELIUS_2018_Fire_GWSW_(A)	0.305		74%
A1-034	T-55	Airport 1	Industrial stormwater pond	21/12/2021	16	QH3_Airport_SEWER_(A)	2.028	*VERY LOW CERTAINTY*	85%
A1-035	T-251	Airport 1	Industrial stormwater pond	05/05/2022	16	GOBELIUS_2018_Fire_GWSW_(A)	0.053		79%
A1-036	T-221	Airport 1	Industrial stormwater pond	15/11/2021	12	DND_Site-G_Military_SW_(A)	0.488		72%
A1-037	T-152	Airport 1	Industrial stormwater pond	21/12/2021	15	DND_Site-G_Military_SW_(A)	0.554	LOW CERTAINTY	73%
A1-038	T-235	Airport 1	Industrial stormwater pond	05/05/2022	13	CALIFGAMA_WWTP_GW_(nA)	0.494		55%
A1-039	T-137	Airport 1	Industrial wastewater pond	15/11/2021	10	QH3_Airport_CONCENTRATE_(A)	0.001		98%
A1-040	T-95	Airport 1	Industrial wastewater pond	21/12/2021	11	QH3_Airport_CONCENTRATE_(A)	0.002		98%
A1-041	T-101	Airport 1	Industrial wastewater pond	05/05/2022	8	QH3_Airport_SEWER_(A)	0.008		97%
A1-042	T-91	Airport 1	Industrial stormwater pond	15/11/2021	13	CALIFGAMA_WWTP_GW_(nA)	0.373		64%
A1-043	T-22	Airport 1	Industrial stormwater pond	21/12/2021	15	DND_Site-G_Military_SW_(A)	0.783	LOW CERTAINTY	68%
A1-044	T-111	Airport 1	Industrial stormwater pond	05/05/2022	10	CALIFGAMA_WWTP_GW_(nA)	0.246		47%

[a] NNZ = number of nonzero PFAS components in the unknown. C_1 = class from training dataset most like the unknown sample; (A) = AFFF-associated subset; (nA) = non-AFFF-associated subset. $SSD_1 = SSD_{min}$ for this unknown corresponding to class C_1. Certainty flag = indicator of the likelihood that the unknown may not be represented in the training dataset. P_{AFFF} overall = estimated probability that sample is of AFFF origin.

Table 3 Classification of data from Airport 2. Only the first 45 unknown samples are shown; see ESI for the full table[a]

Plot ID	Test ID	Site	Description	Sample date	NNZ	C_1	SSD_1	Certainty flag	P_{AFFF} overall
A2–001	T-66	Airport 2	Groundwater	03/12/2021	11	QH3_Airport_WWTP_(A)	0.002		96%
A2–002	T-109	Airport 2	Groundwater	30/11/2021	14	CALIFGAMA_WWTP_GW_(nA)	0.021		73%
A2–003	T-108	Airport 2	Groundwater	01/12/2021	16	QH3_Airport_WWTP_(A)	0.227		87%
A2–004	T-246	Airport 2	Groundwater	03/12/2021	17	QH3_Airport_WWTP_(A)	0.173		90%
A2–005	T-185	Airport 2	Groundwater	18/01/2022	11	QH3_Airport_WWTP_(A)	0.153		82%
A2–006	T-11	Airport 2	Groundwater	24/05/2022	23	CALIFGAMA_MSW_Landfill_GW_(nA)	0.324		52%
A2–007	T-195	Airport 2	Groundwater	23/05/2022	14	CALIFGAMA_Airport_GW_(A)	0.100		81%
A2–008	T-211	Airport 2	Groundwater	24/05/2022	23	QH3_Airport_WWTP_(A)	0.092		80%
A2–009	T-217	Airport 2	Groundwater	24/05/2022	23	CALIFGAMA_Airport_GW_(A)	0.105		84%
A2–010	T-214	Airport 2	Groundwater	25/05/2022	9	CALIFGAMA_WWTP_GW_(nA)	0.003		73%
A2–011	T-194	Airport 2	Groundwater	25/05/2022	10	CALIFGAMA_Airport_GW_(A)	0.111		90%
A2–012	T-172	Airport 2	Groundwater	01/06/2022	18	CALIFGAMA_WWTP_GW_(A)	0.001		91%
A2–013	T-138	Airport 2	Groundwater	31/05/2022	10	CALIFGAMA_Airport_GW_(A)	0.072		92%
A2–014	T-28	Airport 2	Groundwater	01/06/2022	18	CALIFGAMA_CPS_GW_(A)	0.054		99%
A2–015	T-236	Airport 2	Groundwater	01/06/2022	21	DND_Site-G_Military_SW_(A)	0.227		80%
A2–016	T-153	Airport 2	Groundwater	01/06/2022	14	CALIFGAMA_Airport_GW_(A)	0.331		90%
A2–017	T-10	Airport 2	Groundwater	02/06/2022	16	CALIFGAMA_WWTP_GW_(nA)	0.354		58%
A2–018	T-79	Airport 2	Groundwater	01/06/2022	4	CALIFGAMA_WWTP_GW_(A)	0.027		79%
A2–019	T-188	Airport 2	Groundwater	02/06/2022	10	CALIFGAMA_Airport_GW_(A)	0.102		89%
A2–020	T-53	Airport 2	Groundwater	02/06/2022	12	CALIFGAMA_Airport_GW_(A)	0.531	LOW CERTAINTY	88%
A2–021	T-146	Airport 2	Groundwater	01/06/2022	23	CALIFGAMA_Airport_GW_(A)	0.138		77%
A2–022	T-123	Airport 2	Groundwater	01/06/2022	7	CALIFGAMA_Airport_GW_(A)	0.105		88%
A2–023	T-48	Airport 2	Groundwater	01/06/2022	23	CALIFGAMA_Airport_GW_(A)	0.050		94%
A2–024	T-2	Airport 2	Groundwater	01/06/2022	23	CALIFGAMA_Airport_GW_(A)	0.047		94%
A2–025	T-205	Airport 2	Groundwater	01/06/2022	23	CALIFGAMA_Airport_GW_(A)	0.123		86%
A2–026	T-170	Airport 2	Groundwater	31/05/2022	14	CALIFGAMA_Airport_GW_(A)	0.044		95%
A2–027	T-113	Airport 2	Groundwater	31/05/2022	12	QH3_Airport_WWTP_(A)	0.019		98%
A2–028	T-201	Airport 2	Groundwater	02/06/2022	10	CALIFGAMA_Airport_GW_(A)	0.126		88%
A2–029	T-131	Airport 2	Groundwater	31/05/2022	14	QH3_Airport_SW_(A)	0.089		89%
A2–030	T-49	Airport 2	Groundwater	31/05/2022	12	CALIFGAMA_Airport_GW_(A)	0.029		94%
A2–031	T-116	Airport 2	Groundwater	31/05/2022	11	CALIFGAMA_Airport_GW_(A)	0.041		94%
A2–032	T-233	Airport 2	Groundwater	27/05/2022	12	CALIFGAMA_Airport_GW_(A)	0.003		93%
A2–033	T-78	Airport 2	Groundwater	27/05/2022	17	CALIFGAMA_Airport_GW_(A)	0.079		86%
A2–034	T-124	Airport 2	Groundwater	26/05/2022	16	QH3_Airport_GW_(A)	0.148		97%
A2–035	T-30	Airport 2	Groundwater	26/05/2022	13	CALIFGAMA_Airport_GW_(A)	0.012		95%
A2–036	T-74	Airport 2	Groundwater	26/05/2022	14	CALIFGAMA_Airport_GW_(A)	0.046		94%
A2–037	T-252	Airport 2	Groundwater	26/05/2022	15	QH3_Airport_GW_(A)	0.089		87%
A2–038	T-127	Airport 2	Groundwater	25/05/2022	12	QH3_Airport_GW_(A)	0.066		94%
A2–039	T-218	Airport 2	Groundwater	25/05/2022	14	CALIFGAMA_Airport_GW_(A)	0.050		94%
A2–040	T-135	Airport 2	Groundwater	24/05/2022	23	QH3_Airport_WWTP_(A)	0.138		91%
A2–041	T-182	Airport 2	Groundwater	24/05/2022	23	QH3_Airport_WWTP_(A)	0.015		96%
A2–042	T-72	Airport 2	Groundwater	24/05/2022	23	QH3_Airport_WWTP_(A)	0.253		80%
A2–043	T-54	Airport 2	Groundwater	23/05/2022	9	CALIFGAMA_WWTP_GW_(A)	0.373		82%
A2–044	T-160	Airport 2	Groundwater	23/05/2022	11	QH3_Airport_WWTP_(A)	0.132		84%
A2–045	T-159	Airport 2	Groundwater	27/05/2022	15	CALIFGAMA_Airport_GW_(A)	0.031		97%

[a] NNZ = number of nonzero PFAS components in the unknown. C_1 = class from training dataset most like the unknown sample; (A) = AFFF-associated subset; (nA) = non-AFFF-associated subset. $SSD_1 = SSD_{min}$ for this unknown corresponding to class C_1. Certainty flag = indicator of the likelihood that the unknown may not be represented in the training dataset. P_{AFFF} overall = estimated probability that sample is of AFFF origin.

In the case of Airport 2, samples are groundwater samples, and are identified as similar to range of largely airport-associated AFFF classes, although in some cases they are also identified as similar to other AFFF-associated classes, such as wastewater treatment plant influent and effluent of AFFF origin, or military sites.

In the case of industrial site 1, the top matches for the unknown samples are all AFFF-associated classes, although the top classes are generally different from those at Airport 1 or 2, with large numbers of samples matched to offsite residential well data near an AFFF manufacturing facility in Wisconsin (MARINETTE_OFF_AFFF-Mfg_GW_(A)), as well as landfill leachate data from Australian landfills (GALLEN_2017_Landfill_GW_(A)), a dataset[10] that appears to be dominated by PFAS of AFFF origin, as indicated by auto-classification results (Table 1).

Table 4 Classification of data from industrial site 1[a]

Plot ID	Test ID	Site	Description	Sample date	NNZ	C_1	SSD_1	Certainty flag	P_{AFFF} overall
I1-001	T-247	Industrial	Groundwater	15/11/2022	7	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.007		93%
I1-002	T-12	Industrial	Groundwater	15/11/2022	8	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.020		95%
I1-003	T-181	Industrial	Groundwater	15/11/2022	8	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.009		94%
I1-004	T-17	Industrial	Groundwater	15/11/2022	7	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.018		99%
I1-005	T-184	Industrial	Groundwater	03/11/2021	8	PDX_Airport_GW_(A)	0.041		69%
I1-006	T-161	Industrial	Groundwater	09/06/2021	8	GALLEN_2017_Landfill_GW_(A)	0.122		91%
I1-007	T-29	Industrial	Groundwater	11/08/2022	9	GALLEN_2017_Landfill_GW_(A)	0.051		94%
I1-008	T-40	Industrial	Groundwater	09/06/2021	8	GALLEN_2017_Landfill_GW_(A)	0.009		97%
I1-009	T-177	Industrial	Groundwater	11/08/2022	9	GALLEN_2017_Landfill_GW_(A)	0.009		96%
I1-010	T-142	Industrial	Groundwater	11/08/2022	9	GALLEN_2017_Landfill_GW_(A)	0.006		97%
I1-011	T-33	Industrial	Groundwater	10/06/2021	8	GALLEN_2017_Landfill_GW_(A)	0.016		96%
I1-012	T-224	Industrial	Groundwater	09/08/2022	8	GALLEN_2017_Landfill_GW_(A)	0.014		96%
I1-013	T-8	Industrial	Groundwater	10/06/2021	8	GALLEN_2017_Landfill_GW_(A)	0.008		96%
I1-014	T-60	Industrial	Groundwater	09/08/2022	8	GALLEN_2017_Landfill_GW_(A)	0.009		97%
I1-015	T-31	Industrial	Groundwater	03/09/2020	7	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.045		99%
I1-016	T-180	Industrial	Groundwater	08/06/2021	7	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.016		99%
I1-017	T-97	Industrial	Groundwater	08/06/2021	7	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.012		99%
I1-018	T-207	Industrial	Groundwater	11/08/2022	7	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.022		99%
I1-019	T-115	Industrial	Groundwater	08/06/2021	6	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.095		98%
I1-020	T-1	Industrial	Groundwater	10/08/2022	7	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.212		87%
I1-021	T-158	Industrial	Groundwater	03/09/2020	8	GALLEN_2017_Landfill_GW_(A)	0.056		90%
I1-022	T-227	Industrial	Groundwater	03/09/2020	8	GALLEN_2017_Landfill_GW_(A)	0.066		89%
I1-023	T-203	Industrial	Groundwater	10/06/2021	7	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.356		80%
I1-024	T-37	Industrial	Groundwater	10/08/2022	8	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.238		89%
I1-025	T-83	Industrial	Groundwater	08/06/2021	9	MARINETTE_ON_AFFF-Mfg_GWSW_(A)	0.219		95%
I1-026	T-47	Industrial	Groundwater	10/08/2022	7	MARINETTE_ON_AFFF-Mfg_GWSW_(A)	0.213		88%
I1-027	T-69	Industrial	Groundwater	09/06/2021	8	GALLEN_2017_Landfill_GW_(A)	0.034		94%
I1-028	T-168	Industrial	Groundwater	11/08/2022	8	PDX_Airport_GW_(A)	0.178		94%
I1-029	T-239	Industrial	Groundwater	07/06/2021	8	GALLEN_2017_Landfill_GW_(A)	0.010		95%
I1-030	T-21	Industrial	Groundwater	10/08/2022	8	GALLEN_2017_Landfill_GW_(A)	0.026		91%
I1-031	T-5	Industrial	Groundwater	07/06/2021	7	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.035		98%
I1-032	T-192	Industrial	Groundwater	12/08/2022	7	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.009		99%
I1-033	T-148	Industrial	Groundwater	09/06/2021	7	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.006		99%
I1-034	T-122	Industrial	Groundwater	10/08/2022	7	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.015		99%
I1-035	T-129	Industrial	Groundwater	12/08/2022	8	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.102		86%
I1-036	T-250	Industrial	Groundwater	09/08/2022	9	MARINETTE_ON_AFFF-Mfg_GWSW_(A)	0.001		85%
I1-037	T-16	Industrial	Groundwater	11/08/2022	9	GALLEN_2017_Landfill_GW_(A)	0.007		97%
I1-038	T-23	Industrial	Groundwater	03/09/2020	9	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.065		92%
I1-039	T-26	Industrial	Groundwater	09/06/2021	8	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.031		96%
I1-040	T-126	Industrial	Groundwater	10/08/2022	8	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.025		96%
I1-041	T-119	Industrial	Surface water	12/08/2020	10	GALLEN_2017_Landfill_GW_(A)	0.099		82%
I1-042	T-120	Industrial	Surface water	12/08/2020	12	MARINETTE_ON_AFFF-Mfg_GWSW_(A)	0.557	LOW CERTAINTY	79%
I1-043	T-117	Industrial	Surface water	12/08/2020	12	GALLEN_2017_Landfill_GW_(A)	0.128		82%
I1-044	T-178	Industrial	Surface water	12/08/2020	11	GALLEN_2017_Landfill_GW_(A)	0.164		82%
I1-045	T-173	Industrial	Surface water	24/08/2020	13	MARINETTE_ON_AFFF-Mfg_GWSW_(A)	0.097		85%

[a] NNZ = number of nonzero PFAS components in the unknown. C_1 = class from training dataset most like the unknown sample; (A) = AFFF-associated subset; (nA) = non-AFFF-associated subset. $SSD_1 = SSD_{min}$ for this unknown corresponding to class C_1. Certainty flag = indicator of the likelihood that the unknown may not be represented in the training dataset. P_{AFFF} overall = estimated probability that sample is of AFFF origin.

It is important to note that the classification method used here effectively functions as a similarity checker, looking for classes whose sample compositional patterns are consistent with those in each unknown sample. As such, it is reasonable to anticipate that some samples at the classes identified as matches for the unknown samples will be quite similar in composition to the unknown samples. Fig. 1 compares the unknown composition with that of the closest matching samples from each of the top three classes identified through classification for four selected unknown samples from each of the three AFFF-dominated sites (Airport 1, Airport 2, Industrial Site 1). Note that Fig. 1 shows only 12 samples for purposes of discussion, selected to illustrate the range of different compositions observed, and the matches to samples in identified classes; plots for all 198 samples from the three sites are included in the accompanying ESI Section.† It is interesting to

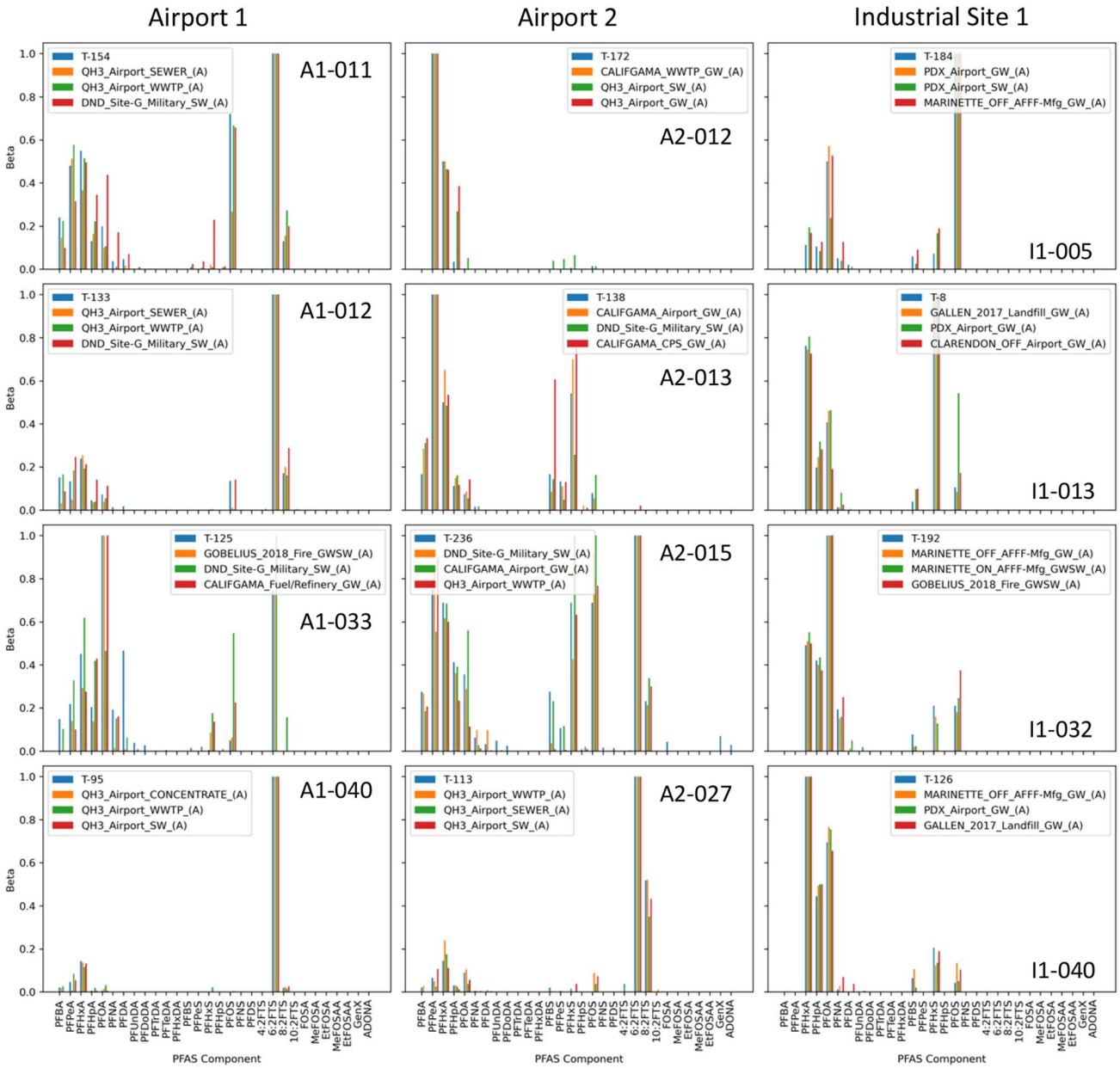

Fig. 1 Component distributions (β) in selected unknowns at three AFFF-dominated sites, shown with closest matching known distributions in the top three selected classes from the training set, as identified by Random Forest classification. The blue bars are the unknown samples (indicated with code T-__), while the orange, green and red bars correspond to the first, second and third identified classes (C_1, C_2, C_3), respectively. Note that plots corresponding to all unknowns are included in the accompanying online ESI.†

observe that the compositions at the three sites in Fig. 1 vary considerably between samples at each site, as well as between the sites. Not only do PFAS compositions change as a result of differential transport and the transformation of precursors,[1] but many sites have histories of use of more than one AFFF, resulting in mixed compositional signatures. Fortunately (from a classification standpoint) AFFF has been so widely used that even these mixed signatures are recognizable by comparison with existing environmental data. The AFFF formulations used at Airport 1 are unknown, although many of the samples are dominated by 6 : 2 FTS. Unknown samples are identified by classification as being similar to samples from an AFFF release at the Brisbane Airport where Angus Tridol S3 was released, so it is probable the main formulation used at Airport 1 is compositionally similar to that formulation. Like Airport 1, many samples at Airport 2 are dominated by 6 : 2 FTS, but many also show evidence of PFOS and PFHxS. There is a known history of use of newer AFFFs at Airport 2, including T-Storm C6 foams and Buckeye Platinum 3% AFFF, as well as historical use of legacy PFOS-based AFFFs. For the industrial site, it is important to note that the original sample data for the site did not include any analyses for PFAS compounds to the left of PFHxA or to the

Table 5 Classification of data from municipal wastewater plant influents[a]

Plot ID	Test ID	Site	Description	Sample date	NNZ	C_1	SSD_1	Certainty flag	P_{AFFF} overall
U1.1-001	T-18	Utility 1, Plant 1	Muni wastewater influent, Plant 1	16/12/2019	10	CALIFGAMA_WWTP_GW_(nA)	0.267		33%
U1.1-002	T-77	Utility 1, Plant 1	Muni wastewater influent, Plant 1	05/03/2020	2	CALIFGAMA_WWTP_GW_(A)	0.000		99%
U1.1-003	T-73	Utility 1, Plant 1	Muni wastewater influent, Plant 1	08/06/2020	4	CALIFGAMA_WWTP_GW_(nA)	0.034		11%
U1.1-004	T-238	Utility 1, Plant 1	Muni wastewater influent, Plant 1	13/09/2021	6	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.210		86%
U1.1-005	T-151	Utility 1, Plant 1	Muni wastewater influent, Plant 1	23/09/2021	6	CALIFGAMA_WWTP_GW_(A)	0.091		87%
U1.2-001	T-225	Utility 1, Plant 2	Muni wastewater influent, Plant 2	17/12/2019	11	CALIFGAMA_WWTP_GW_(nA)	0.126		19%
U1.2-002	T-163	Utility 1, Plant 2	Muni wastewater influent, Plant 2	04/03/2020	6	CALIFGAMA_WWTP_GW_(A)	0.366		69%
U1.2-003	T-39	Utility 1, Plant 2	Muni wastewater influent, Plant 2	11/06/2020	6	CALIFGAMA_WWTP_GW_(A)	0.307		65%
U1.2-004	T-43	Utility 1, Plant 2	Muni wastewater influent, Plant 2	15/09/2021	5	CALIFGAMA_WWTP_GW_(nA)	0.156		70%
U1.2-005	T-46	Utility 1, Plant 2	Muni wastewater influent, Plant 2	24/09/2021	0			NO DETECTS IN UNKNOWN	
U1.3-001	T-157	Utility 1, Plant 3	Muni wastewater influent, Plant 3	17/12/2019	11	CALIFGAMA_CPS_GW_(A)	0.094		93%
U1.3-002	T-166	Utility 1, Plant 3	Muni wastewater influent, Plant 3	04/03/2020	6	DND_Site-G_Military_SW_(A)	0.080		92%
U1.3-003	T-155	Utility 1, Plant 3	Muni wastewater influent, Plant 3	11/06/2020	6	DND_Site-G_Military_SW_(A)	0.131		90%
U1.3-004	T-7	Utility 1, Plant 3	Muni wastewater influent, Plant 3	15/09/2021	6	CALIFGAMA_CPS_GW_(A)	0.154		90%
U1.3-005	T-114	Utility 1, Plant 3	Muni wastewater influent, Plant 3	24/09/2021	2	JERVISBAY_Military_GW_(A)	0.000		100%
U1.4-001	T-145	Utility 1, Plant 4	Muni wastewater influent, Plant 4	18/12/2019	10	CALIFGAMA_MSW_Landfill_GW_(nA)	0.165		30%
U1.4-002	T-85	Utility 1, Plant 4	Muni wastewater influent, Plant 4	02/03/2020	10	CALIFGAMA_MSW_Landfill_GW_(nA)	0.249		31%
U1.4-003	T-104	Utility 1, Plant 4	Muni wastewater influent, Plant 4	10/06/2020	7	CENTRE_Coatings_SW_(nA)	0.021		51%
U1.4-004	T-187	Utility 1, Plant 4	Muni wastewater influent, Plant 4	14/09/2021	6	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.198		93%
U1.4-005	T-103	Utility 1, Plant 4	Muni wastewater influent, Plant 4	23/09/2021	6	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.216		86%
U1.5-001	T-86	Utility 1, Plant 5	Muni wastewater influent, Plant 5	18/12/2019	8	CALIFGAMA_WWTP_GW_(nA)	0.170		17%
U1.5-002	T-27	Utility 1, Plant 5	Muni wastewater influent, Plant 5	02/03/2020	9	CALIFGAMA_MSW_Landfill_GW_(nA)	0.048		20%
U1.5-003	T-63	Utility 1, Plant 5	Muni wastewater influent, Plant 5	10/06/2020	6	CALIFGAMA_WWTP_GW_(nA)	0.050		53%
U1.5-004	T-6	Utility 1, Plant 5	Muni wastewater influent, Plant 5	14/09/2021	6	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.292		87%
U1.5-005	T-204	Utility 1, Plant 5	Muni wastewater influent, Plant 5	23/09/2021	6	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.262		88%
U1.6-001	T-105	Utility 1, Plant 6	Muni wastewater influent, Plant 6	14/09/2021	4	LAVARACK_OFF_Military_SW_(A)	0.309		93%
U1.6-002	T-62	Utility 1, Plant 6	Muni wastewater influent, Plant 6	23/09/2021	8	PDX_Airport_GW_(A)	0.023		96%
U2-001	T-44	Utility 2	Muni wastewater influent	27/05/2020	11	CALIFGAMA_WWTP_GW_(nA)	0.088		14%
U2-002	T-234	Utility 2	Muni wastewater influent	25/08/2020	8	CALIFGAMA_WWTP_GW_(nA)	0.147		13%
U2-003	T-242	Utility 2	Muni wastewater influent	28/09/2021	8	CALIFGAMA_WWTP_GW_(nA)	0.063		16%
U2-004	T-248	Utility 2	Muni wastewater influent	20/10/2021	12	CALIFGAMA_WWTP_GW_(nA)	0.224		26%
U2-005	T-56	Utility 2	Muni wastewater influent	23/11/2021	10	CALIFGAMA_WWTP_GW_(nA)	0.108		21%

[a] NNZ = Number of nonzero PFAS components in the unknown. C_1 = class from training dataset most like the unknown sample; (A) = AFFF-associated subset; (nA) = non-AFFF-associated subset. $SSD_1 = SSD_{min}$ for this unknown corresponding to class C_1. Certainty flag = indicator of the likelihood that the unknown may not be represented in the training dataset. P_{AFFF} overall = estimated probability that sample is of AFFF origin.

right of PFOS in the plot, so if other compounds are present (*e.g.* 6 : 2 FTS), they would not appear in the distributions. This difference may at least in part explain the largely different subset of identified classes compared with the two airports, although the identified classes are still predominantly of AFFF origin. Note that many sites for which experimental data have been measured over a span of years often exhibit differences in the number of analyzed compounds over time, often with fewer compounds analyzed in older data. Taking into account the differences in compounds analyzed, the compositions at the industrial site are reminiscent of those at Airport 2, although PFOA is more prominent in some of the industrial site compositions. The AFFF used to extinguish the fire at the industrial site is thought to have been National Foam Universal Gold.

Table 5 shows the classification results for samples taken from the influents of seven different wastewater treatment plants. Because wastewater treatment plant influents come from multiple sources, there is a high likelihood that they will consist of PFAS from multiple, mixed sources. Not surprisingly, a large fraction of the unknowns in Table 5 are identified as being similar to samples from other mixed sources, including wastewater treatment plant data sources and landfill leachate data sources. The overall AFFF probability for these mixed samples is likely influenced by the highest concentration contributors to the mixtures, although more work is needed to better understand how classification is influenced by mixture composition. (It should be noted that one of the unknown samples (T-46) contained no detected PFAS, so classification is not possible; this is indicated in the Certainty Flag column.)

Fig. 2 compares the unknown composition with that of the closest match from each of the top three classes identified through classification for three selected samples from each of three plant influents. Note that Fig. 2 shows only 9 samples for purposes of discussion, selected to illustrate the range of different compositions observed, and the matches to samples in identified classes; plots for all 32 samples from the three sites are included in the accompanying ESI Section.† One of the interesting features of all of the influents is the temporal variability of compositions for a given plant. Some of the plants, such as Utility 1 Plant 3 influent, appear to be dominated by AFFF sources, although the compositions at Utility 1 Plant 3 are

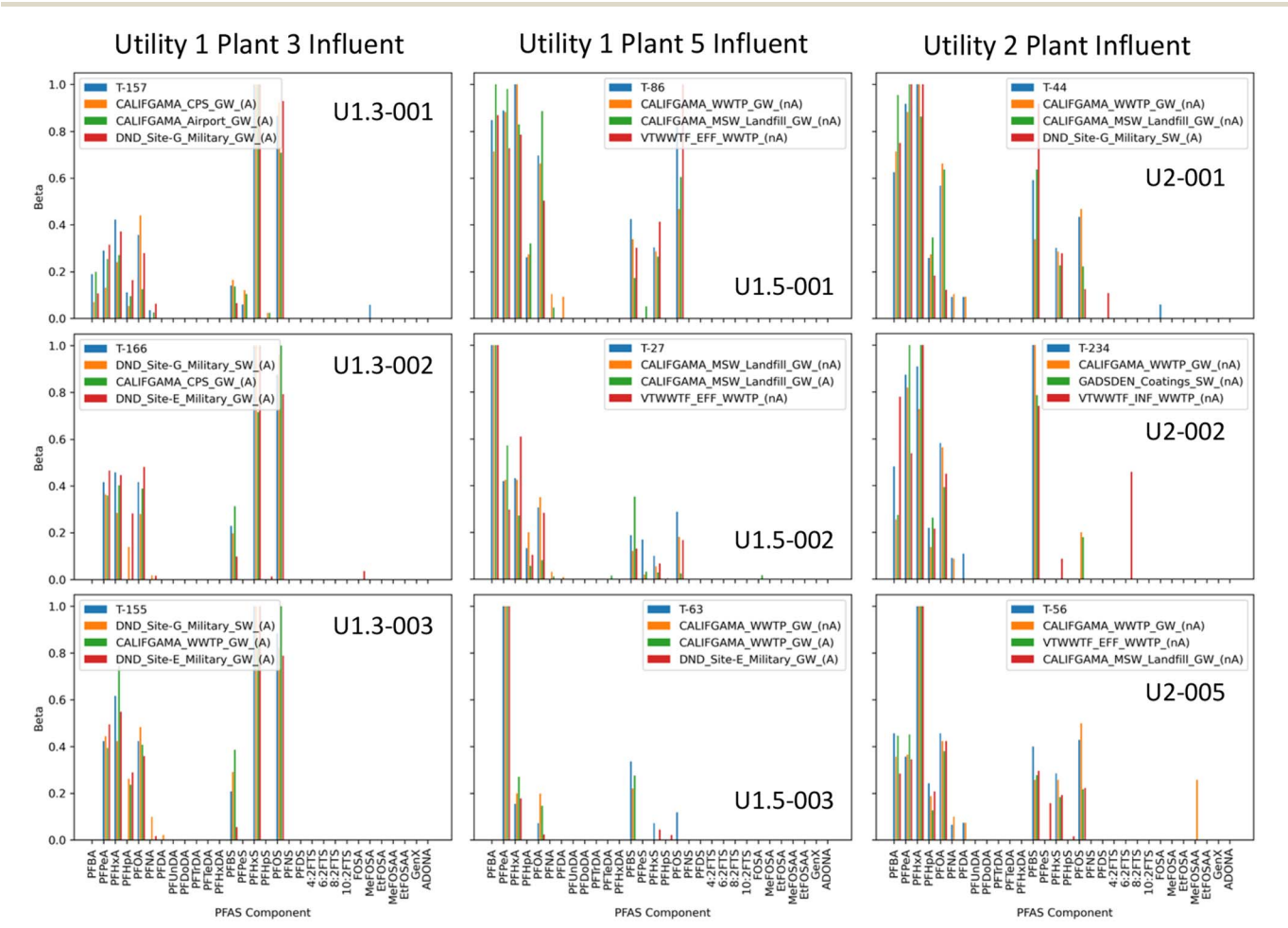

Fig. 2 Component distributions (β) in selected unknowns from three different municipal wastewater treatment plant influents, shown with closest matching known distributions in the top three selected classes from the training set, as identified by Random Forest classification. The blue bars are the unknown samples (indicated with code T-__), while the orange, green and red bars correspond to the first, second and third identified classes (C_1, C_2, C_3), respectively. Note that plots corresponding to all unknowns are included in the accompanying online ESI.†

Table 6 Classification of data from artificial samples[a]

Plot ID	Test ID	Site	Description	Sample date	NNZ	C_1	SSD_1	Certainty flag	P_{AFFF} overall
Art1-001	T-14	Artificial 1	Random select, normalized to 500, inverse	n/a	10	CALIFGAMA_WWTP_GW_(nA)	3.277	**UNLIKE TRAINING SET!**	51%
Art1-002	T-156	Artificial 1	Random select, normalized to 500, inverse	n/a	12	DND_Site-G_Military_SW_(A)	3.505	**UNLIKE TRAINING SET!**	63%
Art1-003	T-102	Artificial 1	Random select, normalized to 500, inverse	n/a	5	MARINETTE_OFF_AFFF-Mfg_GW_(A)	0.801	LOW CERTAINTY	78%
Art1-004	T-243	Artificial 1	Random select, normalized to 500, inverse	n/a	7	CALIFGAMA_WWTP_GW_(nA)	0.867	LOW CERTAINTY	50%
Art1-005	T-164	Artificial 1	Random select, normalized to 500, inverse	n/a	22	CAPEFEAR_GenX_WWTP_(nA)	15.891	**UNLIKE TRAINING SET!**	65%
Art1-006	T-132	Artificial 1	Random select, normalized to 500, inverse	n/a	8	CALIFGAMA_Metal_Plating_GW_(nA)	2.087	*VERY LOW CERTAINTY*	51%
Art1-007	T-4	Artificial 1	Random select, normalized to 500, inverse	n/a	7	MARINETTE_ON_AFFF-Mfg_GWSW_(A)	2.270	*VERY LOW CERTAINTY*	61%
Art1-008	T-38	Artificial 1	Random select, normalized to 500, inverse	n/a	10	GOBELIUS_Industrial_GWSW_(A)	2.543	*VERY LOW CERTAINTY*	73%
Art1-009	T-213	Artificial 1	Random select, normalized to 500, inverse	n/a	7	GOBELIUS_2018_Fire_GWSW_(A)	1.622	LOW CERTAINTY	83%
Art1-010	T-222	Artificial 1	Random select, normalized to 500, inverse	n/a	9	CALIFGAMA_WWTP_GW_(nA)	2.435	*VERY LOW CERTAINTY*	63%
Art1-011	T-45	Artificial 1	Random select, normalized to 500, inverse	n/a	14	CALIFGAMA_WWTP_GW_(nA)	6.907	**UNLIKE TRAINING SET!**	57%
Art1-012	T-96	Artificial 1	Random select, normalized to 500, inverse	n/a	7	LAVARACK_OFF_Military_SW_(A)	2.061	*VERY LOW CERTAINTY*	88%
Art2-001	T-65	Artificial 2	Random concentrations, zero to 100	n/a	23	OAKEY_Military_SW_(A)	6.721	**UNLIKE TRAINING SET!**	77%
Art2-002	T-143	Artificial 2	Random concentrations, zero to 101	n/a	22	CALIFGAMA_WWTP_GW_(nA)	4.665	**UNLIKE TRAINING SET!**	72%
Art2-003	T-71	Artificial 2	Random concentrations, zero to 102	n/a	22	OAKEY_Military_SW_(A)	6.059	**UNLIKE TRAINING SET!**	78%
Art2-004	T-61	Artificial 2	Random concentrations, zero to 103	n/a	23	CALIFGAMA_WWTP_GW_(nA)	6.260	**UNLIKE TRAINING SET!**	63%
Art2-005	T-84	Artificial 2	Random concentrations, zero to 104	n/a	23	CALIFGAMA_WWTP_GW_(nA)	5.030	**UNLIKE TRAINING SET!**	65%
Art2-006	T-210	Artificial 2	Random concentrations, zero to 105	n/a	22	CALIFGAMA_WWTP_GW_(nA)	4.213	**UNLIKE TRAINING SET!**	65%
Art2-007	T-107	Artificial 2	Random concentrations, zero to 106	n/a	22	DND_Site-G_Military_SW_(A)	3.973	**UNLIKE TRAINING SET!**	74%
Art2-008	T-13	Artificial 2	Random concentrations, zero to 107	n/a	22	QH3_Airport_WWTP_(A)	6.535	**UNLIKE TRAINING SET!**	73%
Art2-009	T-87	Artificial 2	Random concentrations, zero to 108	n/a	22	DND_Site-G_Military_SW_(A)	4.705	**UNLIKE TRAINING SET!**	64%
Art2-010	T-193	Artificial 2	Random concentrations, zero to 109	n/a	23	CALIFGAMA_WWTP_GW_(nA)	5.104	**UNLIKE TRAINING SET!**	69%

[a] NNZ = number of nonzero PFAS components in the unknown. C_1 = class from training dataset most like the unknown sample; (A) = AFFF-associated subset; (nA) = non-AFFF-associated subset. $SSD_1 = SSD_{min}$ for this unknown corresponding to class C_1. Certainty flag = indicator of the likelihood that the unknown may not be represented in the training dataset. P_{AFFF} overall = estimated probability that sample is of AFFF origin.

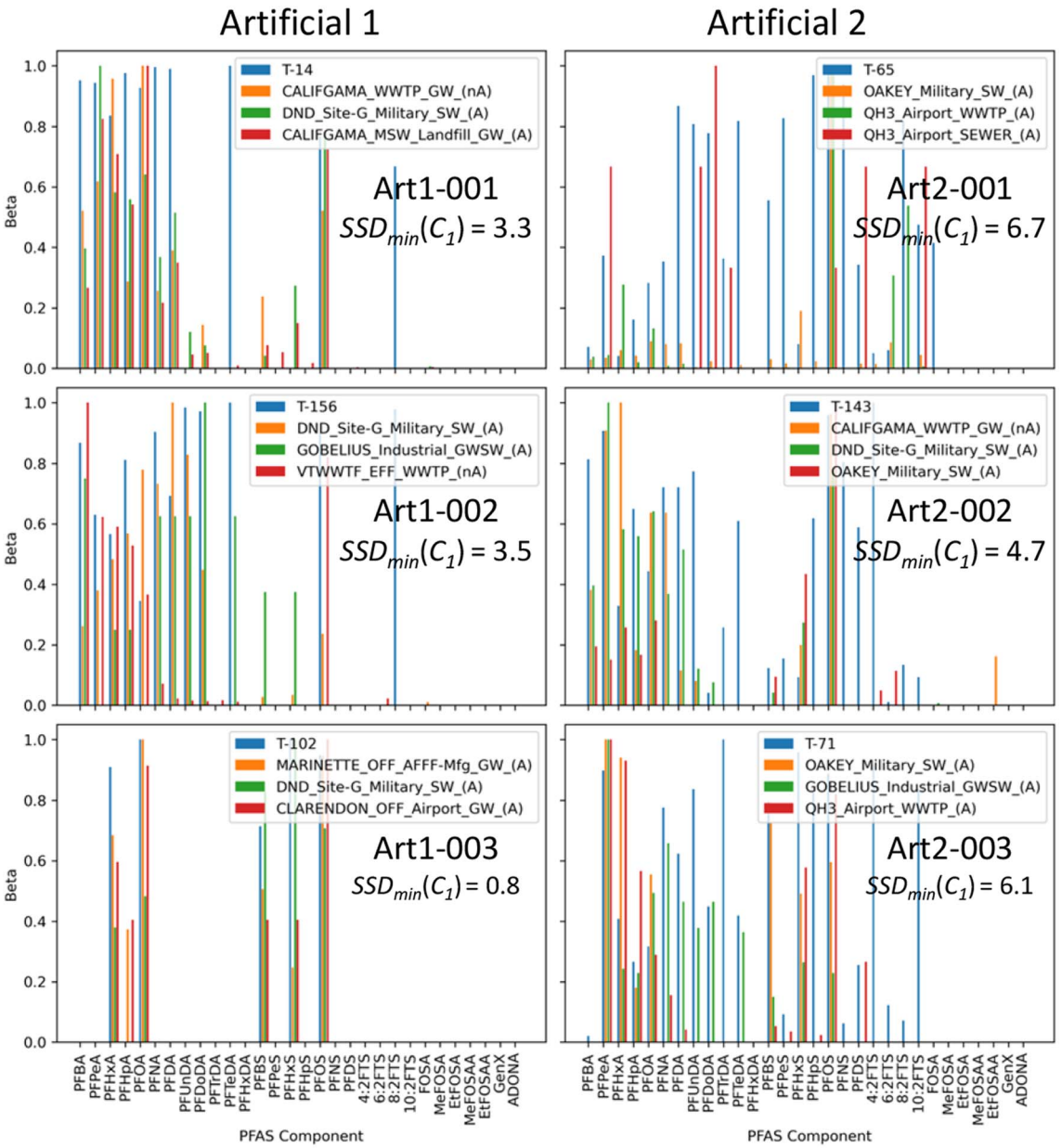

Fig. 3 Component distributions (β) in selected randomly-generated synthetic unknown samples, generated using two different methods. Note that all of the artificial samples exhibit substantial differences from any training set samples. The blue bars are the unknown samples (indicated with code T-__), while the orange, green and red bars correspond to the first, second and third identified classes (C_1, C_2, C_3), respectively. Plots corresponding to all unknowns are included in the accompanying online ESI.†

different from those in Fig. 1. Other plant influents such as the Utility 2 Plant influent tend to be dominated by non-AFFF sources, although in the case of the Utility 2 Plant influent, most of the classifications are to autoclassified mixed sources.

An important part of this work was exploring a rejection algorithm to identify unknown samples not sufficiently represented in the training dataset for accurate classification. The challenge with classification algorithms such as the Random Forest method is that the calculated probabilities for all training set classes add to 100%, even if, in reality, the unknown sample is entirely unlike anything in the training set. Table 6 shows classification results for artificial compositions generated by BC using two different methods. Artificial 1 samples had compositions calculated by randomly selecting another sample from the unknown dataset, and then processing the concentrations of the components in that sample to replace any non-zero detected concentration with a value of 500 ng L^{-1} minus the original concentration normalized to 500 by scaling between the minimum and maximum concentration in the sample. This method yielded something with a composition different from actual samples, but with the same set of detected components. Artificial 2 samples were simply generated randomly with values between zero and 100 ng L^{-1}.

From Table 6, it is apparent that SSD_{min} values for the artificial samples are generally greater than most of the values in Tables 1–5 for the actual unknown samples. All of the Artificial 2 samples are correctly identified as UNLIKE TRAINING SET, while the Artificial 1 samples are mostly identified as either VERY LOW CERTAINTY or UNLIKE TRAINING SET, although three are flagged as LOW CERTAINTY.

Fig. 3 shows some example compositions for selected artificial samples from Table 6, along with compositions of the closest match from each of the top three classes identified through classification. It's easy to see why the Artificial 2 samples are flagged as UNLIKE TRAINING SET, because they genuinely look nothing like any of the closest matches in the training set. Most of the Artificial 1 samples do look quite different from the closest matches (for example, T-14 and T-156 in Fig. 3), although a few – often those with a small number of nonzero compounds, such as T-102 – look somewhat similar to existing samples, so are flagged as LOW CERTAINTY. This is not necessarily a problem with the rejection algorithm, but rather simply a reflection of the fact that if a sample composition looks similar to something in the training set – even if it was artificially generated – there is no mathematical way to identify it as an artificial sample.

It is important to discuss the results of this work within the broader context of PFAS forensics, where the objective is identification of the original source associated with PFAS detected in environmental samples. Methods explored by others have included a number of different approaches, many focused on searching for specific compounds or combinations of compounds unique to a specific source of PFAS, or using multivariate statistical methods to look for patterns in PFAS from different sources (*e.g.* (ref. 4 and 11–16)). Some proposed methods have potential pitfalls, such as susceptibility to changing PFAS composition with transport or transformation of precursors, or potential challenges associated with detection limits, where specific compounds are too low in concentration to be detected in some samples. For this reason and others, it has been suggested (*e.g.* (ref. 13)) that source identification should ideally be based on multiple lines of evidence. The method described in this work can be thought of as providing a very direct additional line of evidence for source identification, by looking for similarities between unknowns and existing environmental samples of known origin. This work builds on earlier supervised learning work studying binary classification of PFAS between AFFF and non-AFFF sources,[1–3] illustrating that the same underlying idea that works for binary classification also works for multiclass classification to distinguish between multiple sources. Because the method is trained on thousands of actual environmental samples, the resulting classification automatically takes into account compositional changes that result from differential transport and precursor transformation. Both the method itself and the accompanying rejection method could be thought of a reality check on any PFAS forensics method, in that if an identified source is, in fact, correct, then it is highly likely that there are other environmental samples with similar compositions to the unknown sample associated with the same type of source elsewhere. The absence of evidence that this is the case may be taken as an indicator that a proposed source assignment is suspect.

Conclusions

The results of this work show that supervised machine learning provides a highly-capable tool for identifying unknown PFAS samples based on composition. The approach tested made use of the Random Forest method for multiclass classification, with the individual classes defined based on individual existing data sources. The method effectively functions as a similarity checker, looking for known sites whose compositional patterns are the closest match to those in each unknown sample. The method was found to be able to recognize samples of AFFF origin at sites with a known history of AFFF use, in some cases making more subtle distinctions in classification. For example, despite significant variability in sample compositions across and between airport sites, samples from airports were largely identified as being similar to samples from other airports, and some samples from an airport wastewater collection system were even identified as looking like samples from another airport wastewater collection system. In the case of municipal wastewater treatment facility influents, where the influent composition varies widely over time and between facilities, and is likely to result from a changing mixture of different original sources, the classifier identified a large fraction of unknown samples as being similar to samples from other mixed sources, such as wastewater treatment plants or landfill leachates, although some exhibited distinct AFFF signatures.

While the use of mixed data sources (*e.g.*, data from wastewater treatment plants or landfill leachates) to train classifiers appears to work well in classification, and sidesteps the substantial challenges associated with finding sufficient single-application non-AFFF environmental data for a training set, the obvious limitation of the approach is that one wastewater treatment plant influent, for example, may ultimately be classified as looking like another wastewater treatment plant influent. Unless more is known about the influent in the training set, this result may or may not be useful. As such, future work aimed at learning more about the true origins of mixed data could be extremely valuable. For example, data collected from within a wastewater collection system close to known sources could be extremely valuable for providing more insight in classifications. Similarly, it is probable that a machine learning classifier could be trained to identify specific dominant AFFF types in different samples, or even specific mixtures of dominant types, if enough information could be obtained about AFFF types used in training set data sources.

The ability to reject samples as not in the training dataset is a critical component of the use of machine learning for PFAS classification, because most supervised classifiers will assign unknowns to a known class, even in cases where they are quite different from all known sets. The rejection method tested here appears quite promising, and was able to accurately flag artificially-generated samples as being unlike those in the training dataset.

In the broader context of PFAS forensics for source identification, the results of this work could be thought of as a reality check, providing a direct line of evidence as to the likely origin of a particular unknown sample. If the proposed sample source type identified by any forensic method is correct, it is highly likely that other examples of the same composition will be present in other environmental samples. Both the method used here and the accompanying rejection method are designed to look for this evidence.

Conflicts of interest

There are no conflicts of interest to report.

References

1 T. C. G. Kibbey, R. Jabrzemski and D. M. O'Carroll, Supervised machine learning for source allocation of per- and polyfluoroalkyl substances (PFAS) in environmental samples, *Chemosphere*, 2020, **252**, 126593.
2 T. C. G. Kibbey, R. Jabrzemski and D. M. O'Carroll, 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.
3 T. C. G. Kibbey, R. Jabrzemski and D. M. O'Carroll, Predicting the relationship between PFAS component signatures in water and non-water phases through mathematical transformation: Application to machine learning classification, *Chemosphere*, 2021, **282**, 131097.
4 J. F. Stults, C. P. Higgins and D. E. Helbling, Integration of Per- and Polyfluoroalkyl Substance (PFAS) Fingerprints in Fish with Machine Learning for PFAS Source Tracking in Surface Water, *Environ. Sci. Technol. Lett.*, 2023, **10**, 1052–1058.
5 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, *et al.*, Scikit-learn: Machine Learning in Python, *J Mach Learn Res*, 2011, **12**, 2825.
6 L. Breiman, Random Forests, *Mach. Learn.*, 2001, **45**, 5–32.
7 B. E. T. H. Twala, M. C. Jones and D. J. Hand, Good methods for coping with missing data in decision trees, *Pattern Recognit. Lett.*, 2008, **29**, 950.
8 J. Josse, N. Prost, E. Scornet and G. Varoquaux, On the consistency of supervised learning with missing values, *arXiv*, 2020, preprint, arXiv:1902.06931, DOI: **10.48550/arXiv.1902.06931**.
9 C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006, p. 738.
10 C. Gallen, D. Drage, G. Eaglesham, S. Grant, M. Bowman and J. F. Mueller, Australia-wide assessment of perfluoroalkyl substances (PFASs) in landfill leachates, *J. Hazard. Mater.*, 2017, **331**, 132–141.
11 X. C. Hu, D. Q. Andrews, A. B. Lindstrom, T. A. Bruton, L. A. Schaider, P. Grandjean, *et al.*, Detection of Poly- and Perfluoroalkyl Substances (PFASs) in U.S. Drinking Water Linked to Industrial Sites, Military Fire Training Areas, and Wastewater Treatment Plants, *Environ. Sci. Technol. Lett.*, 2016, **3**, 344–350.
12 X. Zhang, R. Lohmann, C. Dassuncao, X. C. Hu, A. K. Weber, C. D. Vecitis, *et al.*, 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.
13 J. A. Charbonnet, A. E. Rodowa, N. T. Joseph, J. L. Guelfo, J. A. Field, G. D. Jones, *et al.*, Environmental Source Tracking of Per- and Polyfluoroalkyl Substances within a Forensic Context: Current and Future Techniques, *Environ. Sci. Technol.*, 2021, **55**, 7237–7245.
14 E. Dávila-Santiago, C. Shi, G. Mahadwar, B. Medeghini, L. Insinga, R. Hutchinson, *et al.*, Machine Learning Applications for Chemical Fingerprinting and Environmental Source Tracking Using Non-target Chemical Data, *Environ. Sci. Technol.*, 2022, **56**, 4080–4090.
15 N. T. Joseph, T. Schwichtenberg, D. Cao, G. D. Jones, A. E. Rodowa, M. A. Barlaz, *et al.*, Target and Suspect Screening Integrated with Machine Learning to Discover Per- and Polyfluoroalkyl Substance Source Fingerprints, *Environ. Sci. Technol.*, 2023, **57**, 14351–143162.
16 E. H. Antell, S. Yi, C. I. Olivares, B. J. Ruyle, J. T. Kim, K. Tsou, *et al.*, The Total Oxidizable Precursor (TOP) Assay as a Forensic Tool for Per- and Polyfluoroalkyl Substances (PFAS) Source Apportionment, *ACS ES&T Water*, 2023, DOI: **10.1021/acsestwater.3c00106**.