Scilight

*Environmental Contamination: Causes and Solutions*
https://www.sciltp.com/journals/eccs

*Article*

# PFAS Prophet: Prioritising Per- and Polyfluoroalkyl Substances Using Mass and MS/MS Spectra

Mathieu François Feraud [1,*], Ian A. Wood [2], Saer Samanipour [1,3,4], Pradeep Dewapriya [1], Jake O'Brien [1,3,*] and Kevin Thomas [1]

[1] Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, Woolloongabba, QLD 4102, Australia
[2] School of Mathematics and Physics, The University of Queensland, Woolloongabba, QLD 4072, Australia
[3] Van't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, 1090 GD Amsterdam, The Netherlands
[4] UvA Data Science Center, University of Amsterdam, 1090 GD Amsterdam, The Netherlands
* Correspondence: mathieu.feraud@outlook.com (M.F.F.); j.obrien2@uq.edu.au (J.O.)

**How To Cite:** Feraud, M.F.; Wood, I.A.; Samanipour, S.; et al. PFAS Prophet: Prioritising Per- and Polyfluoroalkyl Substances Using Mass and MS/MS Spectra. *Environmental Contamination: Causes and Solutions* **2026**, *2*(2), 7. https://doi.org/10.53941/eccs.2026.100007

Received: 23 February 2026
Revised: 2 June 2026
Accepted: 3 July 2026
Published: 10 August 2026

**Abstract:** Per- and polyfluoroalkyl substances (PFAS) are a diverse group of over 14,700 synthetic chemicals known for their environmental persistence and association with adverse health effects. Non-target analysis (NTA) using high-resolution mass spectrometry (HRMS) is increasingly used for detecting and discovering new PFAS. However, manual analysis of HRMS data is time-consuming, and current automated systems depend on known reference standards and are restricted to Data Dependent Acquisition (DDA) methods. Here we propose a machine learning algorithm to prioritize potential PFAS-related compounds based on spectral data. The model, trained on distinct compound structures, can prioritize both known and unknown PFAS-related compounds without relying on prior statistical assumptions or reference standards, achieving an F1-macro score of 0.928, however only achieving a F1 score of 0.698 on PFAS related spectra. It leverages complex fragmentation patterns, KMD, and neutral losses to enhance detection accuracy.

**Keywords:** per- and polyfluoroalkyl substances (PFAS); non-target analysis (NTA); high-resolution mass spectrometry (HRMS); machine learning (ML); environmental contaminants; MS/MS spectra; mass spectrometry classification

## 1. Introduction

Per- and polyfluoroalkyl substances (PFAS) are a group of synthetic chemicals that have been manufactured and used in a variety of consumer products since the 1950s. Although the direct health effects of human exposure to PFAS are not well understood, there is evidence that exposure to PFAS is associated with higher blood cholesterol levels [1–12]; limited evidence that it is associated with kidney [13–16] and testicular cancers [15,16], lower levels of antibodies in humans [17–20], and damage to the liver and immune system at high levels in animals [21].

PFAS are increasingly being detected in natural waters, wastewater, sludge, and aquatic and land species [22]. Furthermore, PFAS pollution is poorly reversible as both parent compounds and their transformation products are very persistent [23,24]. To date 4730 CAS registered PFAS have been identified on the global market; and as of early 2024, over 14,700 PFAS compounds are included in the CompTox Chemicals Dashboard [25]. Under the revised definition of PFAS by the Organization for Economic Cooperation and Development (OECD), it is estimated that there are over 7 million unique PFAS and more than 21 million fluorinated compounds [26].

Current methods for the detection of PFAS mostly rely on targeted analytical methods and hence require reference standards, however, these methods have access to a limited number of reference standards (25–50 approximately) [27]. Most large-scale studies focus only on specific regulated PFAS such as perfluorooctanoic acids (PFOA) and perfluorooctane sulfonic acids (PFOS) [28]. However, these studies are insufficient as not only

are there millions of distinct PFAS, some less stable PFAS such as perfluoroalkyl acids (PFAA) can undergo transformation to become more stable PFAS once released into the environment [29]. Furthermore, more PFAS are expected to be discovered as new PFAS continue to be synthesized and released into the environment [30,31].

Because of the low abundance of some PFAS compounds and high chemical diversity, high-resolution mass spectrometry (HRMS)-based non-target analysis (NTA) is gaining increasing popularity to identify new PFAS in complex environmental matrices [32,33]. However, there is currently a lack of standards and reference material to assist with their determination [34]. HRMS data can potentially result in the analysis of many thousands of compounds. Manually analyzing these compounds can be extremely time consuming and challenging. Automated methods to identify PFAS from HRMS data have been developed such as Fluoromatch and PFAScreen [28,35], however these are expert systems and depend on the use of known reference standards and knowledge of PFAS characteristics to make inferences. Although the fragmentation patterns for some PFAS are well understood, others, such as PFAS derivatives are more complex. These features, combined with multiple parameters extracted from the First Stage of Mass Spectrometry ($MS_1$), are hard to leverage effectively to detect PFAS. Even expert systems will likely not be able to leverage all potential factors to the fullest. Furthermore, the scope of the data that Fluoromatch can use is limited as it depends on Data Dependent Acquisition (DDA) and is unable to work with Data Independent Acquisition (DIA).

In this study, we present a Machine Learning (ML) algorithm to rapidly identify potential PFAS-related compounds from componentized HRMS data. The advantage of using a ML model is that it can make complex statistical inferences from the data it receives. The training and test sets for modelling and evaluation were chosen randomly under the provision that each compound would only be present in one or the other. This step ensures that the resulting trained model relies on finding common identifiers between PFAS instead of attempting to identify the exact PFAS. As such it is a dependable classifier to prioritize known and unknown PFAS-related compounds as it does not rely on previous reference standards. Furthermore, by using componentized features as input from the Self-Adjusting Feature Detection (SAFD) [36] and componentization (CompCreate) [37] packages, data can be interpreted from either DDA or DIA methods.

For the purposes of this work, the following terminology is adopted. PFAS (per- and polyfluoroalkyl substances) refers to compounds containing at least one fully fluorinated methyl ($–CF_3$) or methylene ($–CF_2–$) carbon atom, consistent with the OECD (2021) structural definition. PFAS precursors are compounds that can degrade or transform into terminal PFAS through biotic or abiotic processes. The term “PFAS-related”, as used throughout this manuscript, denotes the broader class predicted by the model: it encompasses PFAS, PFAS precursors, and structurally associated compounds (e.g., PFAS derivatives) that may not themselves contain fluorine but are linked to PFAS through their chemistry.

## 2. Method

### *2.1. Model Workflow*

The PFAS prioritization algorithm was developed using a ML algorithm. To develop the algorithm, several steps were undertaken as illustrated in the workflow in Figure 1. The MassBank repository was used as the Data source [38]. This data was first transformed and structured for the model and split for use in a grouped cross-validation (CV) (described in Section “Creating the Cross-Validation datasets”). The parameters of the model were then optimized, and the final model was tested on the test set. A final model was built using all the data available and was tested on an external test set which was obtained from Dewapriya et al. [30].

### *2.2. Data Source*

Model Data

Spectral data was retrieved from the MassBank repository [38]. The MassBank repository is an open-source experimental mass spectral library for the identification of small chemical molecules of metabolomics, exposomics, and environmental relevance. The repository can contain multiple spectra for the same compound collected under the same or different conditions, resolution, and/or on different instruments. Subsequently, an in-house algorithm was employed to extract the data from text files and populate a database. The following attributes were extracted from the MassBank data: Exact Mass, Fragment m/z, Simplified Molecular-Input Line-Entry System (SMILES), and ID.

### *2.3. Data Preparation*

#### 2.3.1. Data Filtering

Only spectra acquired through ElectroSpray Ionisation (ESI) and with a resolution of 30,000 or above were initially considered, however around 60% of the spectral data do not report resolution. For those, only spectra with at least three non-zero digits in the mantissa of all their fragment values were considered likely to be from HRMS and retained. This mantissa-based heuristic is an imperfect proxy for HRMS, low-resolution instruments can also report m/z values to several decimal places and its implications for the model are discussed in the Limitations section.

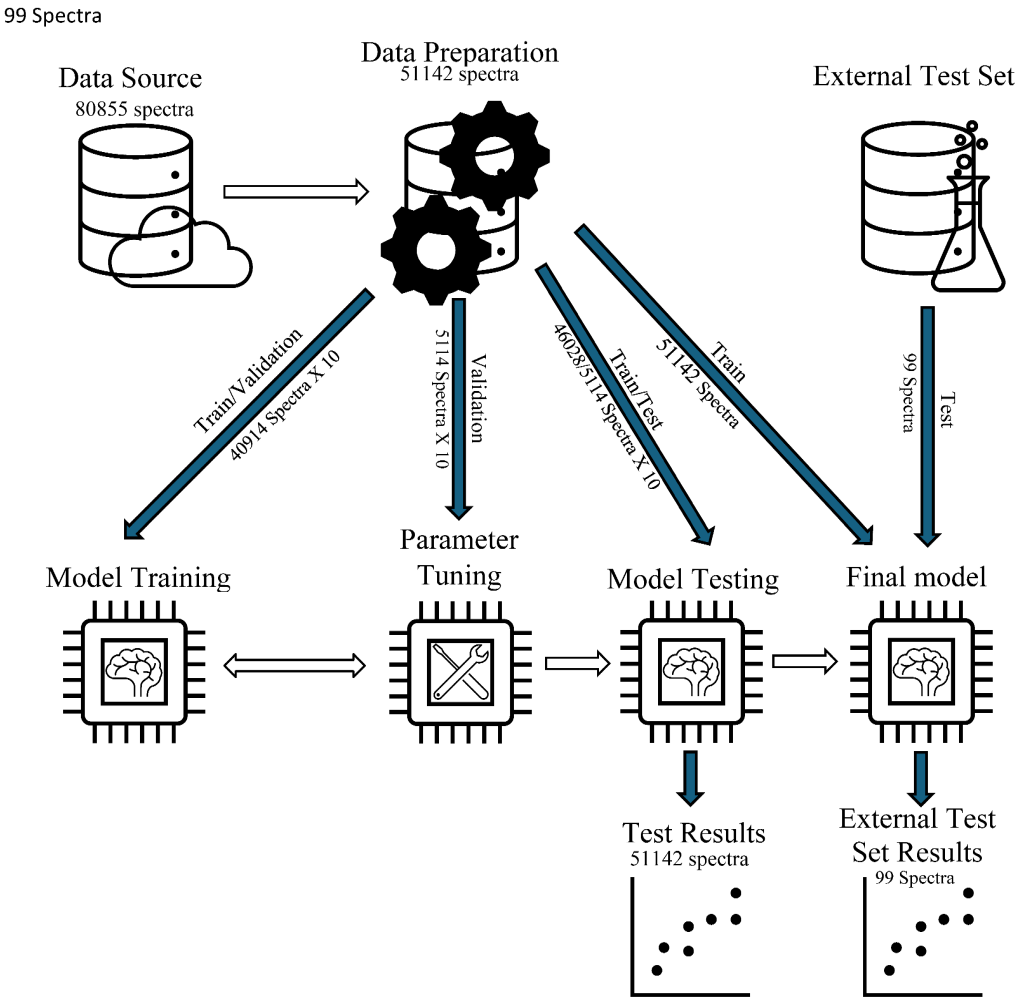


**Figure 1.** Workflow illustrating the steps needed to obtain the final model. The initial data is from the MassBank repository. The data is then selected and processed for the model to train on. A grouped fold CV method was applied (see Section "Creating the Cross-Validation datasets"). The optimal set of parameters are then selected, and the final model is trained on the whole dataset. The External test set is used to evaluate the final model.

#### 2.3.2. Synthetic Measured m/z of Parent Ion

As the reference spectra are derived from analytical standards, the measured m/z of the parent ion is not available; a synthetic measured m/z was therefore generated from the exact mass.

Most PFAS found in the environment and biota samples ionise in negative mode and they mainly deprotonate under common ionization techniques [39]. To focus mainly on environmentally related PFAS and to simplify the ionization mode, the model assumes that all PFAS compounds deprotonate. To achieve this, the exact mass was transformed into a synthetic measured m/z by subtracting the mass of a proton and then adding an error term to it. This error term is computed by assuming that all compounds and fragments were measured at a resolution of 30,000.

The resulting error was derived through the following calculations:

$$\mathcal{R} = \frac{\mathrm{m}/_{\mathrm{z}}}{\mathrm{FWHM}}$$

$$\mathrm{FWHM} = 2 * 2 * \ln 2 * \sigma$$

$$\sigma = \frac{\mathrm{m}_{/\mathrm{z}}}{\mathcal{R} * 2 * \sqrt{2 * \ln(2)}}$$

where:

$\mathcal{R}$ is the resolution, FWHM is the FWHM, $\sigma$ is the standard deviation, and $\mathrm{m}_{/\mathrm{z}}$ is the mass-to-charge ratio of an ion.

The equation above is derived from the probability density function of the normal distribution, for further information on the derivations, please see [40].

Assuming the error is normally distributed, the estimated error term is obtained by sampling from the normal distribution.

$$e \sim \mathcal{N}(0, \sigma^2)$$

The sampled error term is then added to the exact m/z to create a synthetic measured m/z.

The justification of the resolution choice and the Gaussian error assumption can be found in Text S1.

### 2.3.3. Homologous Series

The synthetic measured m/z is too compound-specific for the model to generalize from given the limited training data. To obtain more generalizable input variables, the Kendrick Mass Defect (KMD) was computed across twelve repeating units [41] repeating units include the seven fluorinated units ($CF_2$, $CF_2O$, $C_2F_4O$, CF, $C_2H_2F_2$, $C_3F_6O$), which aid in detecting PFAS homologue series, and five non-fluorinated units (CN, CO, $CH_2$, CCl, CS), which help classify non-PFAS homologous series [42]. The rationale for the KMD transformation, the choice of repeat units, and algorithm is provided in Text S2.

### 2.3.4. Oversampling

To present a more accurate error distribution to the model, the following was performed: where there were less than 30 reference spectra for a distinct PFAS-related compound, further spectra are created by copying the original spectrum with the exact m/z of the parent ion, and creating a new synthetic measured m/z of the parent ion.

To have an even representation of the spectra for the same compound, the original spectra were copied *x* times where *x* is the quotient of 30 divided by the number of original spectra, and *y* copies were selected without replacement from the original spectra, where *y* was the modulus of the division above.

The rationale for oversampling is provided in Text S3.

### 2.3.5. Addition of Neutral Loss to Fragments

Characteristic neutral losses have been shown to aid in the identification of PFAS [43,44], so neutral loss was also added to the model. Neutral losses were computed by subtracting the pseudo measured precursor ion m/z from the fragment m/z. They represent the remainder weight of the compound and can be considered a fragment themselves. Hence, further discussion about fragments also includes neutral losses.

### 2.3.6. Binning

The following preprocessing steps were implemented:

- Fragments were grouped into uniformly sized bins starting from zero, with bin width treated as a hyperparameter optimized on the validation sets.
- A minimum occurrence threshold per bin was applied as a hyperparameter; bins with too few fragments were excluded.
- Only fragments associated with at least one parent compound in a PFAS-related class were retained, to prevent the model from predicting class solely on fragment presence.
- Imposing a minimum m/z value for fragments. A threshold of −10 Daltons was set to exclude implausible values. This only occurs with neutral loss data as it is suspected that some of the entries require spectra cleaning as there are many MassBank entries that have reported fragments much larger than the precursor ion.

The rationale of these steps can be found in Text S1.

### 2.3.7. Classes

A class titled "PFAS-related" was derived using the PFAS-Map algorithm [45]. The algorithm uses standardized SMILES, which were extracted using the Python package RDKit (version 2025.09.6) [46] to categorize compounds as not being PFAS related, or into various PFAS types and PFAS-related compounds, including silicon PFAS, side-chain fluorinated aromatic PFAS, aliphatic PFAS, PFAS derivatives, PFAA, perfluoro PFAA precursors, non-PFAA perfluoroalkyls, fluorotelomer alkyl sulfonamides (FASA)-based PFAA precursors, fluorotelomer-based PFAA precursors.

The final input variables utilized for this model include the KMDs and the fragment bins. The model attempts to predict the "PFAS-related CLASS" class. This class includes PFAS compounds, but also PFAS derivatives that contain no fluoride and thus are not a PFAS yet related to PFAS.

### 2.3.8. Preliminary Exploratory Data Analysis

In relation to PFAS, the spectra found in the MassBank database can be found in Table 1.

**Table 1.** Class count per spectra and per distinct compounds of the MassBank database.

| | Before Filter | | After Filter | |
|---|---|---|---|---|
| **CLASS** | **Distinct Reference Spectra** | **Distinct Compounds** | **Distinct Spectra** | **Distinct Compounds** |
| Not PFAS-Related | 62,999 | 7284 | 45,790 | 5406 |
| PFAS derivatives | 847 | 73 | 685 | 54 |
| Side-chain aromatics | 1113 | 87 | 941 | 69 |
| Fluorotelomer PFAA precursors | 124 | 13 | 121 | 13 |
| PFAAs | 244 | 31 | 232 | 31 |
| Other aliphatics | 28 | 5 | 28 | 5 |
| Perfluoro PFAA precursors | 8 | 1 | 8 | 1 |
| FASA based PFAA precursors | 26 | 4 | 26 | 4 |

Class Count of the MassBank database.

The PFAS distribution in the m/z spectrum is well represented (Figure 2, Figure S3). For individual PFAS classes, there are some that could potentially have adequate representation such as PFAS derivatives and side-chain aromatics (see Figure S2). However, there are many that are not sufficiently represented to be considered significant, and even some that are only represented by one compound (Figures S1 and S2). An ideal model would predict the PFAS subclass, however, given the limited amount of data and the increase of complexity for the model, the final class variable is a binary value for "PFAS-related", and "not PFAS-related". "PFAS-related" contains all classes besides "Not PFAS-related".

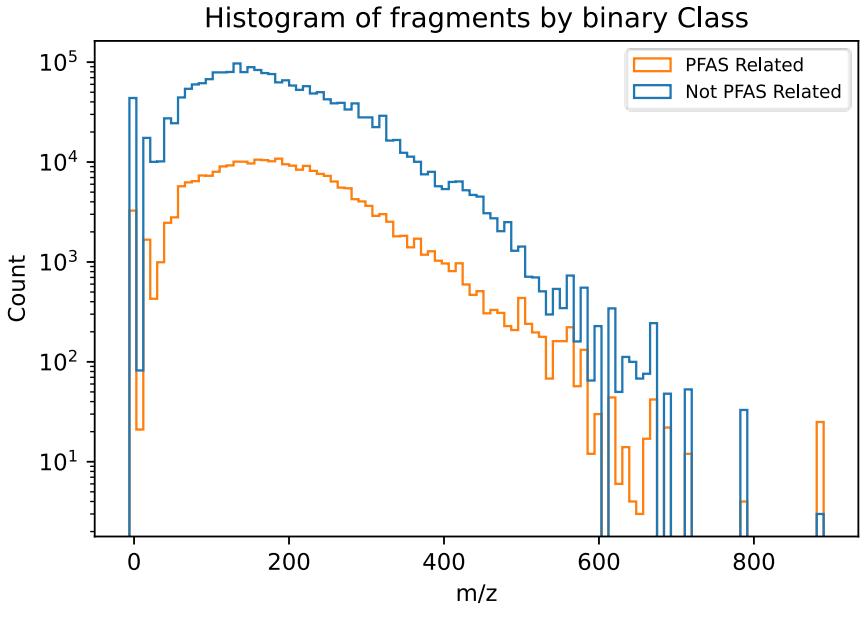


**Figure 2.** Histogram of the collection of fragments and losses of the dataset, grouped by PFAS and non PFAS class, after removal of suspected low-quality spectra and removal of fragments which do not appear in any PFAS spectra.

### *2.4. External Test Data*

The measured mass used in the model construction is a synthetic measured mass extracted from the exact mass, whereas the model is expected to use the mass from HRMS measurements. Furthermore, for the model construction, the compounds were measured by using targeted methods, whereas the model will need to process data coming from untargeted methods. These create some differences between the training/testing data and data from NTA. To test if these differences affect the model performance, an external NTA test set obtained from Dewapriya et al. [30] was used to test the model and evaluate performance.

The raw dataset was from pooled blood and serum extracted from cattle exposed to Aqueous Film Forming Foam (AFFF) impacted groundwater. The HRMS data was acquired using a SCIEX X500R Quadrupole Time-of-Flight (QTOF) mass spectrometer (SCIEX, Framingham, MA, USA) operated in Sequential Window Acquisition of all Theoretical Ion Spectra (SWATH), ESI in negative mode, and the data was processed using SCIEX OS.

#### 2.4.1. PFAS Spectra

From the data processed using SCIEX OS, these had a feature identification through library search, suspect screening, and NTA feature identification. Confidence was determined using the Charbonnet scale [34], which is a confidence scale with identification criteria specific to suspect or NTA of PFAS. The results of these can be found in Tables 2 and 3. For further information on the materials and method see Dewapriya et al. [30].

The results are as follows:

**Table 2.** Blood and serum dataset with reported Charbonnet identification confidence [34].

| Level | Identification Confidence | Potential PFAS Reported |
|---|---|---|
| 1a | Confirmed by reference standard | 12 |
| 1b | Indistinguishable from reference standard | 1 |
| 2a | Probable by library spec. match | 2 |
| 3a | Positional isomer candidates | 10 |
| 3b | Fragmentation-based candidate | 4 |
| 4 | Unequivocal molecular formula | 3 |
| 5a | PFAS suspect screening exact mass match | 4 |
| 5b | Non-target PFAS exact mass of interest | 29 |

PFAS per Charbonnet ID.

**Table 3.** Blood and serum dataset with reported classifications. Library matches are spectra that were identified using the SCIEX OS (SCIEX, Framingham, MA, USA) LibraryView tool with an in-house library of 414 known PFASs, using the candidate search algorithm. The Suspected Matches were obtained via the SCIEX OS suspect screening tool. The new structures are compounds that were not identified but whose properties give them a high likelihood (3b in the Charbonnet scale) of being a PFAS.

| Classifications | Number of PFAS Detected |
|---|---|
| Library Match | 15 |
| Suspect Match | 12 |
| New Structures | 3 |
| Unknowns | 34 |

PFAS per classification.

Although the algorithm processed all compounds, only spectra with a Charbonnet scale of level 1a and 1b are used to mitigate the chances of false positives.

#### 2.4.2. Non-PFAS Compounds

From the compound list obtained from SCIEX OS, the compounds detected as PFAS or potentially PFAS were excluded. 77 compounds were selected at random. The fragmentation of these compounds was then extracted from SCIEX-OS and further screened for PFAS, four of these compounds could not be ruled out as being PFAS and were removed from the possible candidates. The remaining 73 were treated as not being PFAS.

### *2.5. Training the Model*

#### 2.5.1. Classifier Selection

The LightGBM classifier was chosen for its ability to handle complex relationships efficiently [47]. The comparison with XGBoost and CatBoost that led to this choice is provided in Text S5.

#### 2.5.2. Creating the Cross-Validation Datasets

The model uses a grouped 10-fold CV with a three-way (train/validation/test) split. To obtain this CV, the dataset was first partitioned into ten folds at the compound level. Each compound can have multiple associated reference spectra; however, all spectra belonging to the same compound were restricted to a single fold, ensuring that spectra from the same compound never appear in more than one-fold. The folds were then arranged following a standard ten-fold CV scheme. For each iteration, eight folds were used for training the model, one-fold for validation, and the remaining fold for testing. The procedure was carried out ten times, rotating the validation and test folds so that each fold was used once as the validation set and once as the test set. Critically, each model's test fold was strictly excluded from its own training and hyperparameter tuning, ensuring no data leakage within any individual model. Collectively, the validation and test sets span the entire dataset. The rationale for this setup is given in Text S6.

Each spectrum is assigned a weight calculated by dividing one by the total number of spectra for the corresponding compound. This weighting ensures that compounds with limited representation are given equal importance, compared to those with more spectra (e.g., compounds with only 30 spectra versus those with over 100).

#### 2.5.3. Parameter Tuning

There were 5 different hyperparameters (see Table 4) that were varied and 10 iterations for the cross validation. A limited grid search was applied to find the optimal parameters; this resulted in 7560 trained models.

**Table 4.** Hyperparameters and constants of data and model.

| Name | Description |
|---|---|
| Bin width | The size of the bins used to group the fragments on the m/z scale. |
| Number of leaves | Maximum number of leaves in one tree. |
| Minimum data in leaf | Minimum number of observations in a leaf node. |
| Learning rate | Determines the step size in the direction of steepest descent of the loss function for each iteration. |
| Number of estimators | Maximum number of trees to be built (the model can be stopped early by "early stopping round" if little to no increase in log loss is obtained by the last n number of trees). |
| Early Stopping Rounds | Stops the training early if the log loss of validation dataset doesn't improve in the last n trees. |
| Minimum fragment bin count considered | Specifies the minimum number of fragments required for a bin to be considered. |
| Minimum count considered | The minimum number of fragments present in a bin for that bin to be included in the model. |
| Is OCFPC | Only considers fragments which have at least one parent compound of the class of interest (PFAS-related). |
| Is SGHL | Removes fragments that are only present in a single class. |
| Metric | The evaluation Metric used on the validation set. |
| Class weight | When set to "balanced" the model automatically assigns weights to each class inversely proportional to their frequency in the input data. |
| Objective | The type of regression or classification the model needs to perform. |
| Max depth | The number of sequential splits allowed. |

Hyperparameters of model.

#### 2.5.4. Model Training

Within each CV iteration, models were trained using different hyperparameter combinations and evaluated on the corresponding validation fold. For each CV iteration, the hyperparameter set achieving the highest validation F1-score was selected, and the model trained with these parameters was used to generate predictions for the test fold. The test performance reported for the model corresponds to the average of the test results across the ten iterations.

The final hyperparameter set of the final model is obtained from the set of hyperparameters that produces the highest validation score over the ten iterations. The final model is trained on all the data using this selected set of hyperparameters.

2.5.5. Hyperparameter Description

The hyperparameters evaluated for model assessment can be categorized into two groups: those directly influencing the structure of the input data fed into the model, and those stemming from the classifier's hyperparameters. The parameters for each category can be observed in Table 4.

For additional information regarding LightGBM parameters, see the LightGBM documentation [48].

### *2.6. Metrics Used*

The metric used to assess the model was the F1 score on the validation set. Further evaluation metrics were reported, these are the F1 score, PFAS-related Recall, Balanced Accuracy, and the precision-recall. These metrics were investigated when considering the spectra as a singular sample, and the compound as a singular sample.

## 3. Results

### *3.1. Final Parameters*

Table 5 gives the final parameters for the model, they are separated into "Data Parameters" for the parameters that affect the dataset the model will train on, and the "Model Parameters" are the parameters that directly affect the model.

**Table 5.** Optimal parameters for the dataset that is used for the model, and optimal parameters of the LightGBM model. Not all combinations of the variables were tested as it was noticed that some values were performing poorly regardless of the values of the other variables and thus were not further considered. Other variables were tested but not listed as they were sub-optimal.

| **Data Parameters** | | |
|---|---|---|
| **Name** | **Final Value** | **Values Evaluated** |
| Fixed resolution | True | True |
| Bootstrap PFAS | True | True |
| Minimum exponent | 3 | 3 |
| Minimum loss | −10 | −10 |
| CV splits | 10 | 10 |
| Minimum count considered | 3 | 3 |
| Only consider fragments whose parents belongs to searched class | True | True |
| Group fragments that are only present in one class | True | True |
| Bin width | 0.10 | 0.07, 0.08, 0.09, 0.10, 0.11, 0.12, 0.13 |
| **Model Parameters** | | |
| **Name** | **Value** | **Variables** |
| Number of leaves | 10 | 10, 20, 30 |
| Class weight | balanced | Balanced |
| Number of estimators | 100,000 | 100,000 |
| Metric | Binary log loss | Binary log loss |
| Objective | Binary | Binary |
| Depth | 20 | 20 |
| Minimum data in leaf | 10 | 10, 20, 30 |
| Learning rate | 0.15 | 0.05, 0.08, 0.10, 0.15 |
| Early stopping rounds | 10 | 3, 5, 10 |

Optimal parameters of model.

### *3.2. Test Sets*

The metric used to estimate the model strength and model is the F1-score [49] of the PFAS class, which is used over other scores as it manages class imbalance well. Further metrics are reported, these are the macro F1-score, Balanced accuracy, Precision, and Recall. The metrics are further distinguished by adding a weight factor: The first is by spectra, meaning that each distinct spectrum has equal weight. This is because some PFAS-related spectra were duplicated to create more variance in the synthetic measured m/z measurement. The other is by SMILES, meaning that each distinct compound has equal weight. This was done as some compounds had hundreds of spectra for their representation, while others had very few, and this could bias the results.

As estimated via CV, the optimized models have an F1 macro score of 0.843 (by spectrum) or 0.928 (by compound). Tables 6–8 are the results from the test sets.

Test set results

**Table 6.** Overall results of the test set: F1 Macro, Precision, Balanced Accuracy, and Macro Precision-Recall Area Under the Curve. Values are reported as mean ± half-width of the 95% confidence interval, estimated by percentile bootstrap (2000 resamples of the test set).

| | F1 Macro | Macro Precision | Bal. Acc | Macro PR-AUC |
|---|---|---|---|---|
| By Spectrum | 0.843 ± 0.009 | 0.912 ± 0.010 | 0.795 ± 0.010 | 0.859 ± 0.009 |
| By Compound | 0.928 ± 0.021 | 0.964 ± 0.020 | 0.897 ± 0.030 | 0.956 ± 0.017 |

Overall

**Table 7.** Binary confusion matrix and performance metrics for the test set, using the following abbreviations: TP (True Positive), FP (False Positive), FN (False Negative), TN (True Negative), Prec (Precision), Rec (Recall), and PR-AUC (Precision-Recall Area Under the Curve). Besides for TP, FP, FN, and TN, values are reported as mean ± half-width of the 95% confidence interval, estimated by percentile bootstrap (2000 resamples of the test set).

| **By Spectrum** | **Prec** | **Rec** | **F1** | **PR-AUC** | **TP** | **FP** | **FN** | **TN** |
|---|---|---|---|---|---|---|---|---|
| Not PFAS-related | 0.982 ± 0.001 | 0.995 ± 0.001 | 0.989 ± 0.001 | 0.996 ± 0.001 | 45,563 | 826 | 227 | 1215 |
| PFAS-related | 0.843 ± 0.019 | 0.595 ± 0.021 | 0.698 ± 0.017 | 0.722 ± 0.018 | 1215 | 227 | 826 | 45,563 |
| **By Compound** | **Prec** | **Rec** | **F1** | **PR-AUC** | **TP** | **FP** | **FN** | **TN** |
| Not PFAS-related | 0.993 ± 0.002 | 0.998 ± 0.001 | 0.996 ± 0.001 | 1.000 ± 0.000 | 5396 | 36 | 10 | 141 |
| PFAS-related | 0.934 ± 0.040 | 0.797 ± 0.060 | 0.860 ± 0.040 | 0.913 ± 0.034 | 141 | 10 | 36 | 5396 |

By binomial Classes.

The model-predicted probabilities reported in Table 8 are valid within this evaluation context, where the class prior is fixed. However, when applied to new environmental samples where the class distribution will differ, the output should be interpreted as a prioritization score rather than a calibrated probability.

**Table 8.** Test set metrics for distinct PFAS-related classes. Columns show count, correct predictions, average model-predicted probability, and recall for both spectrum-level and compound-level evaluation.

| Class | Spectra Count | Spectra Correctly Classified | Spectra Average PFAS Probability Reported | Spectra Recall | Compound Count | Compound Correctly Classified | Compound Average PFAS Probability Reported | Compound Recall |
|---|---|---|---|---|---|---|---|---|
| Not PFAS Related | 45,790 | 45,563 | 0.022 | 0.995 | 5406 | 5396 | 0.017 | 0.998 |
| Side-chain aromatics | 941 | 530 | 0.570 | 0.563 | 69 | 49 | 0.690 | 0.710 |
| PFAS derivatives | 685 | 317 | 0.486 | 0.463 | 54 | 40 | 0.671 | 0.741 |
| PFAAs | 232 | 221 | 0.939 | 0.953 | 31 | 31 | 0.958 | 1.000 |
| Fluorotelomer PFAA precursors | 121 | 85 | 0.674 | 0.702 | 13 | 11 | 0.741 | 0.846 |
| Other aliphatics | 28 | 28 | 0.847 | 1.000 | 5 | 5 | 0.886 | 1.000 |
| FASA based PFAA precursors | 26 | 26 | 0.968 | 1.000 | 4 | 4 | 0.976 | 1.000 |
| Perfluoro PFAA precursors | 8 | 8 | 1.000 | 1.000 | 1 | 1 | 1.000 | 1.000 |

Results by distinct classes.

### *3.3. External Test Set*

#### 3.3.1. Positive Samples

The model was further tested with an external test set. It is important to note that the model is not comparing with true PFAS compounds but with compounds identified on the Charbonnet Identification Level as PFAS. Thus, the higher the Charbonnet identification level is, the less likely the compound identified is a PFAS compound. The external test set overview can be observed in Table 9. For evaluation purposes, only spectra with a Charbonnet Identification Level of 1a and 1b are used and can be observed in Figure 3. The results of the remaining spectra can be observed in Tables S1 and S2.

**Table 9.** External test set results per Charbonnet identification level, along with recall statistic per identification Confidence level.

| Charbonnet Identification Level | PFAS Reported | PFAS Detected | Recall |
|---|---|---|---|
| 1a | 12 | 11 | 0.92 |
| 1b | 1 | 1 | 1 |

External test set confidence level statistics.

### 3.3.2. External Test Set Class Statistics

Of the 13 compounds with confidence from 1a to 1b, 11 were detected as PFAS. Of the 12 compounds suspected as PFAS-related, all were detected as PFAS-related. All the three new structures were detected as PFAS-related with a high level of confidence (Figure 3; see Table S1 for suspected and new structure compounds). The unknown list, that had a confidence level ranging from 4 to 5b had 6 that suggested were PFAS-related, and 22 that it suggested were not PFAS-related (Table S2).

| Blood Sample | Serum Sample | Feature m/z | MS2 fragments detected | Charbonnet Identification Level | Model Predicted probability |
|---|---|---|---|---|---|
| New structures | | | | | PFAS related |
| ✓ | ✓ | 248.9461 | Very low intensity | 1a | 0.165 |
| ✓ | ✓ | 298.9427 | 63.9624, 77.9654, 118.9921, 218.9820, 297.9606 | 1a | 0.281 |
| ✓ | ✓ | 348.9396 | 79.9566, 98.9557, 348.9388 | 1a | 0.522 |
| ✓ | ✓ | 397.9525 | 77.9653, 397. 9525 | 1a | 0.848 |
| ✓ | ✓ | 398.9357 | 79.9570, 98.9556, 118.9925, 168.9888, 229.9479, 398.9356 | 1a | 0.862 |
| ✓ | ✓ | 412.9667 | 79.9582, 98.9560, 168.9904 | 1a | 0.981 |
| ✓ | ✓ | 448.9334 | 79.9569, 98.9555, 118.9921, 168.9884, 229.9454, | 1a | 0.992 |
| ✓ | ✓ | 462.9632 | 118.9919, 168.9893, 218.9853, 268.9810, 418.9727 | 1a | 0.975 |
| ✓ | ✓ | 498.9302 | 79.9578, 98.9561, 118.9939, 168.9939, 218.9855, 418.9730, 462.9654 | 1a | 0.990 |
| ✓ | ✓ | 512.9601 | 118.9924, 168.9898, 218.9857, 268.9826, 318.9799, 468.9681 | 1a | 0.968 |
| ✓ | ✓ | 548.9270 | 79.9574, 98.9565, 129.9532, 179.9506, 229.9474, 279.9445, 329.9411, 429.9436 | 1a | 0.902 |
| ✓ | ✓ | 562.9565 | 118.9922, 168.9891, 218.9856, 268.9825, 318.9825, 418.9783, 518.9665 | 1a | 0.975 |
| ✓ | ✓ | 514.9001 | 79.9571, 98.9554, 134.9867, 184.9829, 168.9827, 279.9459 | 1b | 0.760 |

**Figure 3.** External test set results of confirmed PFAS compounds (Charbonnet confidence 1a–1b), with identification confidence, RT, KMD, Library/Suspect acronyms, and fragments reported. (Results of external test set of PFAS Spectra).

Suspected and potential PFAS compounds are reported in Tables S1 and S2 of the Supporting Information.

### 3.3.3. Negative Samples

The algorithm was used to process the non-PFAS compounds of the external dataset. Of the 73 compounds, 71 were identified as not being PFAS. The two remaining compounds were further evaluated by an independent analyst, one is deemed to be highly likely to be a PFAS and the other also had the potential to be a PFAS. However, because this was initially identified by the model, changing the results of these for the evaluation would create data leakage and bias and thus will be considered as a True negative in the evaluation of the model.

### 3.3.4. External Test Set Statistics

The 13 compounds with a confidence of 1b or higher on the Charbonnet scale were selected as PFAS-related spectra. The 73 compounds that were independently screened are used as non-PFAS-related spectra. These spectra were then evaluated by the model, see Table 10 for results:

**Table 10.** Confusion matrix of the external test set and Precision, Recall, and F1 score of PFAS-Related class and Not PFAS-Related class. Where "TP" are PFAS correctly identified, "FP" are not PFAS compounds wrongly identified as PFAS, "FN" are PFAS compounds wrongly identified as not PFAS compounds, and lastly, "TN" are not PFAS compounds correctly identified.

| Precision, Recall and F1 Results of Classes | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **TP** | **FP** | **FN** | **TN** | **Precision** | **Recall** | **F1** |
| PFAS-Related | 11 | 2 | 2 | 71 | 0.846 | 0.846 | 0.846 |
| Not PFAS-related | 71 | 2 | 2 | 11 | 0.973 | 0.973 | 0.973 |

External test set statistics.

### *3.4. Model Investigation*

#### 3.4.1. PFAS Model-Predicted Probability by Compounds

Figure 4 illustrates the reported PFAS model-predicted probability by compound of the model's test set. It is important to note that the *y* axis is scaled logarithmically. The majority of non-PFAS compounds are correctly identified as being not PFAS with a very high degree of confidence. The same can be said for PFAS-related compounds but to a lesser extent. It is also evident that the model struggles more with properly identifying some PFAS-related compounds. This is also reflected in Table 8, Figures S5 and S6, which suggests that most of the PFAS compounds that have a low model-predicted probability score come from "side-chain aromatics", "derivatives", and "fluorotelomer PFAA precursors". It is also verified that all PFAS that were not detected come from these three classes. The reported PFAS model-predicted probability by Spectra can be observed in Figures S4 and S5.

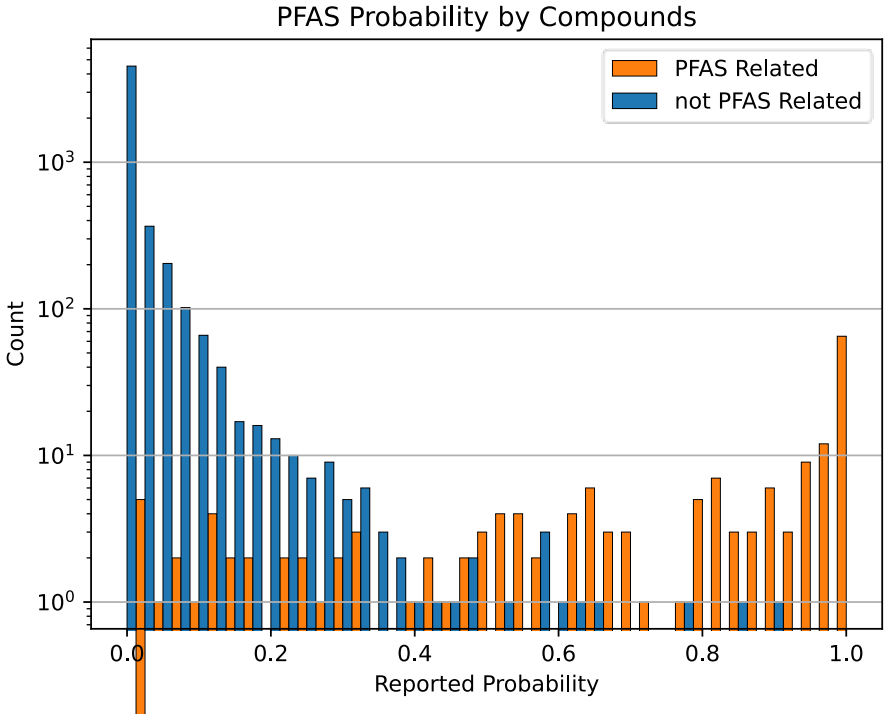


**Figure 4.** Histogram of the Test set model reported probabilities of classes per compounds. Bins are of size 0.025.

#### 3.4.2. Input Variable Comparison

The input variables importance (Figure 5, Figure S7) provides insight into what the model is doing. It was expected that the gain of the KMD of CF, $CF_2$, $C_2F_4O$, $CF_2O$, $C_2H_2F_2$, and $C_3F_6O$ values would have more importance compared to the rest of the input variables given that PFAS have $CF_2$ as repeating unit. However, the KMD of CH, CCI, CN, and CS are more important for the model. It is suspected that this is because these KMD values easily rule out a spectra of being a PFAS.

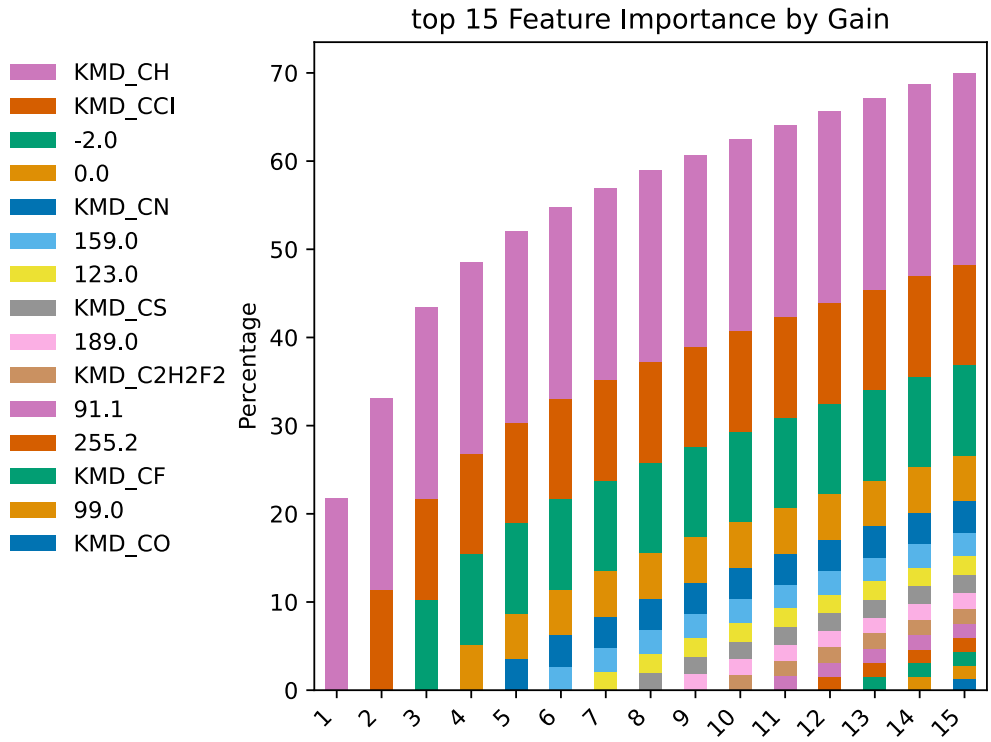


**Figure 5.** Top 15 important Input variables by Gain of the model. The *y* axis of the gain is the percentage of each input variable gain relative of the total gain, where gain is the reduction in training loss that results from adding a split point. The input variables are color coded according to the legend. The *x* axis ordered by starting with the most important input variable, and sequentially adding the next most important input variables. The model has 2137 distinct input variables to select from. The input variables vary from the differing KMD values, and the fragment (neutral losses and fragments) bins.

#### 3.4.3. Sample Trees

The first four trees were investigated to gain further insight in how the model is making its decisions. Typically, in a gradient boosted model, the first trees have a higher impact on the overall model's performance as the subsequent trees are built upon the earlier trees. This can be seen as each subsequent tree has less gain from the former one, and the leaf values have less weight (see "Gain Values" and "leaf Value" in legend in Figures S8–S13).

## 4. Discussion

### *4.1. Plots*

#### 4.1.1. Histogram Plots

The histogram plots (Figure 4, Figure S4) show the test set distributions of the model's reported PFAS-related probabilities separated by PFAS-related and not PFAS-related classes. On both classes, the distribution seems to be highly skewed (when considering the vertical axis is scaled logarithmically) to the right for non-PFAS-related spectra and left for PFAS-related spectra, thus the model is confident in most of its predictions. When comparing the model-predicted probability distributions by compounds to those by spectrum, we see that using compounds as explanatory variables provides much better separation between the PFAS-related and not PFAS-related classes.

#### 4.1.2. Tree Investigation

Although the trees individually are easy to follow, the overall model is hard to interpret accurately. It is noticeable from the four example trees that the model is leveraging the input variables in a complex way. Furthermore, the trees make use of many distinct input variables and given that the top 15 input variables only account for approximately 73% of the total model gain (see Figure 5). The top 86 input variables account for 90% of the model gain and are used often well down each tree. This illustrates that the model makes effective use of multiple input variables.

### *4.2. Model*

#### 4.2.1. Usage and Comparison

For PFAS specific experiments that are DDA and that have three or more experimental blanks, FluoroMatch2 has strong advantages over PFAS Prophet, namely that FluoroMatch2 is much more informative about the matches it finds such as being able to annotate the confidence using the Schymanski scale [50]. PFAScreen also requires specific pre-processing conditions, such as DDA, centroided spectra, and one collision energy source. By design, PFAS Prophet does not process HRMS files, it uses the Separation of Concerns (SoC) principle. It can however process the data from algorithms which are specifically made to process HRMS files, such as the CompCreate [37] algorithm which can process both DIA and DDA from mzXML data. By using the SoC principle, CompCreate can process any type of vendor file as the msConvert [51] algorithm is able to convert any vendor file to the mzXML format. More recently, Meekel et al. [52] demonstrated a conceptually related strategy of using machine learning on MS/MS spectra to prioritize non-target features carrying structural alerts for potential hazard, with models trained on raw fragments and neutral losses and evaluated on aromatic amines and organophosphorus moieties, reporting recall values of 0.65 for organophosphorus structural alerts and 0.58 for aromatic amines on their test set. PFAS Prophet shares the broader objective of using ML on HRMS data to prioritize features of concern, but differs in target chemistry (PFAS-related compounds rather than structural-alert classes), in input representation (componentized MS1/MS2 features from SAFD and CompCreate rather than raw fragments and neutral losses), and in compatibility with both DDA and DIA acquisition modes.

#### 4.2.2. Model Performance

The model has very differing performances depending on the class of PFAS it is detecting. It is particularly strong at detecting PFAAs, other aliphatics, FASA based PFAA precursors, and perfluoro PFAA precursors. When grouped by compound, these classes showed a perfect recall and very high model-predicted probability. It is important to note that both FASA based PFAA precursors and perfluoro PFAA precursors have only 4 and 1 compounds respectively for representation. The model struggled with both PFAS derivatives and side-chain aromatics, it was only able to achieve a spectral recall of 0.463 and 0.563 respectively, and a compound Recall of 0.741 and 0.710 respectively. Of the 5406 distinct non-PFAS-related compounds, only 10 non-PFAS were identified as PFAS-related, which represents a 99.8% accuracy. However, there are 36 PFAS (20 Side-chain aromatics, 14 PFAS derivatives, and 2 Fluorotelomer PFAA precursors) that were not correctly identified.

The model seems to perform better when comparing per compound results to per spectra results. This is mainly due to many of the spectra being misclassified are often the minority of a compound and likely have had a higher error term associated. There are 79 PFAS compound that had at least one spectrum misclassified, of those, 43 of these were identified as PFAS compounds. There were 153 FN spectra that ended up belonging to a compound that was identified correctly, and 117 TP spectra that ended up belonging to a compound that was misidentified.

Because PFAS are known to have common characteristic fragments [53], it is suspected that the model would depend on these for its predictions. The model was able to detect PFAA compounds with high recall and high model-predicted probability. It is suspected that PFAA compounds would produce common characteristic fragments which would make it easier for the model to identify these fragments of interest, however, the model detected compounds such as side-chain aromatics, PFAS derivatives, and fluorotelomer PFAA precursors with low recall and low reported probability, it is suspected that these compounds fragment in a more unpredictable pattern, thus making it harder for the model to discriminate these between PFAS-related spectra, and non-PFAS-related spectra. Furthermore, PFAS derivatives can contain spectra which are not considered PFAS. Some of these spectra can come from compounds which contain two or fewer fluorine.

Observing the input variable importance (see Figure 5 and Figure S7), the KMD whose repeating units have fluorine in them discriminate better than those that do not, and 4 distinct fragment bins are held at higher importance than the KMD with $CF_2$ as a repeating unit.

#### 4.2.3. External Test Set

When observing the F1 score, the model performs better at identifying PFAS-related compounds on the external test set compared to the database test set, and slightly worse at identifying not PFAS-related compounds. This is likely because the measured m/z of the external test set have less error compared to the CV test set. Another factor could be that the PFAS compounds in the external test set may have differed in distribution from those in the database test set. Specifically, the external test set might have had a lower proportion of PFAS derivatives or

side-chain aromatics, which are classes the model struggled with the most. This difference in compound composition could explain the model performance on the external test set.

A discussion of the input variables can be found in Text S7.

### *4.3. Limitations*

#### 4.3.1. Model Usage

It is crucial to understand that while the model outputs the predicted probability of a class based on the given data, this shouldn't be directly interpreted as such when operational. The distribution of the data used to train the model will inevitably differ, and sometimes significantly, from the distribution of the data in each sample. Hence, it is more appropriate to view the model's output as a score indicating the likelihood that a compound is PFAS-related or not.

Additionally, given that the model used a balanced class weight (giving it a 50/50 prior distribution), the model will inherently exhibit a bias towards identifying PFAS-related compounds that are more like those it encountered during training. Another limitation of the model, common to complex ML models, is its limited interpretability. Understanding precisely why the model makes a particular decision is highly challenging, However, models such as gradient boosting do allow one to quantify the relative importance of the various input variables in distinguishing between the classes, as seen in Figure 5, and Figure S7.

#### 4.3.2. Omission of Retention Time

Retention index and retention time were not considered as input parameters as the database did not have any retention information. Furthermore, retention time is experiment dependent, and thus the parameter is not generalizable [54]. Retention indices could be considered, but these were also not reported.

#### 4.3.3. Data Sparsity

Several challenges had to be met to develop a successful model. The main one was data sparsity. After removing the unusable data (~27% of the original data), only 2041 out of 47,831 were PFAS-related. Additionally, there are 5583 distinct compounds, with only 177 of them being related to PFAS, and most of these belong to 3 sub-classes (side-chain aromatics, PFAS derivatives and PFAAs) with 20 belonging to the other 4 sub-classes (fluorotelomer PFAA precursors, other aliphatics, FASA based PFAA precursors, perfluoro PFAA precursors). The limited representation of PFAS-related compounds and classes is potentially the biggest constraint as the aim of the model is to develop a generalized algorithm applicable for all PFAS-related compounds.

#### 4.3.4. Data Quality

Our aim in gathering data for training the model was to replicate the range of entries the model could encounter from NTA data. The MassBank repository data is of varying quality. For example, not all characteristics of spectra are recorded (e.g., Resolution), and some spectra are of differing resolution and acquired under different conditions, and many entries contain noise incorrectly recorded as reference spectra. For this, the data underwent some certain transformations to approximate what could be found from componentization of HRMS data. However, certain segments of the database may still deviate from authentic data representation. Striking a balance between data integrity and quality is essential: excluding all non-suitable entries would yield insufficient training data, while retaining too many could introduce superfluous noise or induce erroneous model expectations. The limitations of data quality are further discussed in Text S8.

The limitations regarding the synthetic measured m/z, the fixed resolution, the bin width, and training the model is further discussed in Text S9.

### *4.4. Potential Future Development*

#### 4.4.1. Data Quality and Quantity

For its datasets, the model had access to only 177 distinct PFAS compounds, representing approximately 0.0025% of known and theoretical PFAS. These compounds belonged to seven distinct classes and 25 subclasses, whereas the PFAS-map algorithm encompasses nine classes and 45 subclasses. Moreover, the model only had access to theoretical measurements of the parent ion, faced inconsistencies in resolution between the parent compounds and fragments, and encountered many non-curated spectra.

Despite these constraints, the model demonstrated significant efficacy when tested against actual data. It is suspected that more and better-quality data could deliver better results if retrained in this above but also allow for the inclusion of additional input variables such as compound m/z which could further benefit the algorithm.

#### 4.4.2. Generalizability

Currently, the model functions solely as a binary classifier due to the sparse nature of the PFAS data. However, with an adequate abundance of PFAS spectra, there is potential to transform the model into a multi-class classifier, in order to know the subclass of the detected PFAS-related compound.

Furthermore, the model's preparation process could be replicated to classify other types of chemicals of emerging concern, such as bisphenols, phthalates and halogenated flame retardants. Although the data source would likely remain similar, optimizing certain model parameters and data parameters is necessary to adapt to the characteristics of the new chemical classes.

Further potential future development is discussed in Text S10.

## Supplementary Materials

The additional data and information can be downloaded at: https://media.sciltp.com/articles/others/260810143940075 3/ECCS-26020125-SI.pdf. Text S1: Synthetic measured m/z of parent ion rational. Text S2: Homologous series. Text S3: Oversampling rational. Text S4: Binning Rationale. Text S5: Rational for model selection. Text S6: Cross-Validation (CV) Datasets. Text S7: Discussion—Model—Input variable comparisons. Text S8: Discussion—Limitations—Data quality. Text S9: Discussion—Limitations—continued. Text S10: Discussion—Potential future development—continued. Figure S1: Histogram of compound m/z per class, grouped by specific PFAS classes and non-PFAS class, after removal of suspected low-quality spectra. Figure S2: Histogram of the collection of fragments and losses of the dataset, grouped by specific PFAS classes and non-PFAS class, after removal of suspected low-quality spectra and removal of fragments which do not appear in any PFAS spectra. Figure S3: Histogram of compound m/z per binomial class, grouped by PFAS and non-PFAS class, after removal of suspected low-quality spectra. Figure S4: Histogram of the Test set model reported probabilities of classes per spectrum. Bins are of size 0.025. Figure S5: Histogram of the spectrum Test set model reported probabilities by Class. Bins are of size 0.1. Figure S6: Histogram of the Compounds Test set model reported probabilities by Class. Bins are of size of 0.1. Figure S7: Top 15 important input variables by split of the model. The *y* scale of the split count is the percentage of each input split relative of the total split. Figure S8: part 1 of first tree of the model. As this is the first tree of the ensemble, the further a node is away from the root node, the more its value tend to differs from the root node. Another aspect is that it is setting a baseline, it is the only tree whose root node value is other than 0. Figure S9: second Tree, this tree compared to the first makes more discerning decisions on which spectrum is a PFAS and which isn't as it has negative values on its leaves. Figure S10: Third Tree. Figure S11: part one of the fourth tree. Table S1: External test set results of suspected PFAS spectra with identification confidence, RT, KMD, Library/Suspect acronyms, and fragments reported. Table S2: External test set results of potential PFAS compounds with low identification confidence. Identification confidence, RT, KMD, Library/Suspect acronyms, and fragments reported. References [55–63] have been cited in the Supplementary Materials.

## Author Contributions

M.F.F.: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft, writing—review & editing, visualization, supervision, project administration; I.A.W.: methodology, formal analysis, writing—review & editing, supervision, project administration; S.S.: conceptualization, methodology, writing—review & editing, supervision; P.D.: investigation, resources, writing—review & editing; J.O.: investigation, writing—original draft, writing—review & editing, supervision; K.T.: conceptualization, resources, writing—review & editing, funding acquisition. All authors have read and agreed to the published version of the manuscript.

## Funding

The author acknowledges support from the National Health and Medical Research Council (NHMRC) through the grant APP1185347. J. O'Brien is the recipient of a National Health and Medical Research Council (NHMRC) Investigator Grant (2009209). The Queensland Alliance for Environmental Health Sciences, The University of Queensland, gratefully acknowledges the financial support of Queensland Health.

## Institutional Review Board Statement

All samples used in the external test set were collected, stored, and analyzed in accordance with the ethical guidelines approved by the University of Queensland (UQ) Ethical Clearance (2013000317, 2018001790 and 2014000614).

## Informed Consent Statement

Not applicable.

## Data Availability Statement

The PFAS Prophet machine learning model and the scripts required to perform prioritization on HRMS data are publicly available on GitHub at https://github.com/Mathieu-Feraud/PFASProphet (archived at https://doi.org/10.5281/zenodo.18251103).

## Conflicts of Interest

The authors declare no conflict of interest.

## Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used Github Copilot for model refinement, and Claude for reviewing and editing the paper. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

## Glossary of Terms, Abbreviations, and Symbols

*Abbreviations*

| Abbreviation | Definition |
|---|---|
| AFFF | Aqueous Film Forming Foam |
| CV | Cross-Validation |
| DDA | Data Dependent Acquisition |
| DIA | Data Independent Acquisition |
| ESI | Electrospray Ionisation |
| F1 score | Harmonic mean of precision and recall |
| FASA | Fluorotelomer Alkyl Sulfonamides |
| FN | False Negative |
| FP | False Positive |
| FWHM | Full Width at Half Maximum |
| HRMS | High-Resolution Mass Spectrometry |
| KM | Kendrick Mass |
| KMD | Kendrick Mass Defect |
| ML | Machine Learning |
| MM | Measured Mass |
| MS1 | First Stage of Mass Spectrometry |
| MS2 | Second Stage of Mass Spectrometry |
| mzXML | Open file format for storing and exchanging mass spectrometry data |
| NTA | Non-Target Analysis |
| NTS | Non-Targeted Screening |
| OECD | Organisation for Economic Co-operation and Development |
| PFAA | Perfluoroalkyl Acids |
| PFAS | Per- and Polyfluoroalkyl Substances |
| PFOA | Perfluorooctanoic Acid |
| PFOS | Perfluorooctane Sulfonic Acid |
| PR-AUC | Precision-Recall Area Under the Curve |
| QTOF | Quadrupole Time-of-Flight |
| SAFD | Self-Adjusting Feature Detection |
| SMILES | Simplified Molecular Input Line Entry System |
| SoC | Separation of Concerns |
| SWATH | Sequential Window Acquisition of all Theoretical Ion Spectra |
| TN | True Negative |
| TP | True Positive |

*Glossary of Terms*

| Term | Definition |
|---|---|
| Feature | A signal from a HRMS chromatogram |
| MassBank | Public repository of mass spectra |
| Spectrum | In mass spectrometry, the chemical information in terms of m/z, intensity, and fragmentation of a single ion |

*Symbols*

| Symbol | Definition |
|---|---|
| $C$ | Repeating unit compound mass |
| $Da$ | Unit of mass |
| $em$ | Synthetic measured mass |
| $p$ | Proton mass |
| $\sigma$ | Standard deviation |
| $\mathcal{N}$ | Gaussian Distribution |

## References


1. Costa, G.; Sartori, S.; Consonni, D. Thirty years of medical surveillance in perfluorooctanoic acid production workers. *J. Occup. Environ. Med.* **2009**, *51*, 364–372. https://doi.org/10.1097/JOM.0b013e3181965d80.
2. Eriksen, K.T.; Raaschou-Nielsen, O.; McLaughlin, J.K.; et al. Association between plasma PFOA and PFOS levels and total cholesterol in a middle-aged Danish population. *PLoS ONE* **2013**, *8*, e56969. https://doi.org/10.1371/journal.pone.0056969.
3. Sakr, C.J.; Kreckmann, K.H.; Green, J.W.; et al. Cross-sectional study of lipids and liver enzymes related to a serum biomarker of exposure (ammonium perfluorooctanoate or APFO) as part of a general health survey in a cohort of occupationally exposed workers. *J. Occup. Environ. Med.* **2007**, *49*, 1086–1096. https://doi.org/10.1097/JOM.0b013e318156eca3.
4. Sakr, C.J.; Leonard, R.C.; Kreckmann, K.H.; et al. Longitudinal study of serum lipids and liver enzymes in workers with occupational exposure to ammonium perfluorooctanoate. *J. Occup. Environ. Med.* **2007**, *49*, 872–879. https://doi.org/10.1097/JOM.0b013e318124a93f.
5. Skuladottir, M.; Ramel, A.; Rytter, D.; et al. Examining confounding by diet in the association between perfluoroalkyl acids and serum cholesterol in pregnancy. *Environ. Res.* **2015**, *143*, 33–38. https://doi.org/10.1016/j.envres.2015.09.001.
6. Steenland, K.; Tinker, S.; Frisbee, S.; et al. Association of perfluorooctanoic acid and perfluorooctane sulfonate with serum lipids among adults living near a chemical plant. *Am. J. Epidemiol.* **2009**, *170*, 1268–1278. https://doi.org/10.1093/aje/kwp279.
7. Winquist, A.; Steenland, K. Modeled PFOA exposure and coronary artery disease, hypertension, and high cholesterol in community and worker cohorts. *Environ. Health Perspect.* **2014**, *122*, 1299–1305. https://doi.org/10.1289/ehp.1307943.
8. Frisbee, S.J.; Shankar, A.; Knox, S.S.; et al. Perfluorooctanoic acid, perfluorooctanesulfonate, and serum lipids in children and adolescents: Results from the C8 Health Project. *Arch. Pediatr. Adolesc. Med.* **2010**, *164*, 860–869. https://doi.org/10.1001/archpediatrics.2010.163.
9. Geiger, S.D.; Xiao, J.; Ducatman, A.; et al. The association between PFOA, PFOS and serum lipid levels in adolescents. *Chemosphere* **2014**, *98*, 78–83. https://doi.org/10.1016/j.chemosphere.2013.10.005.
10. Zeng, X.W.; Qian, Z.; Emo, B.; et al. Association of polyfluoroalkyl chemical exposure with serum lipids in children. *Sci. Total Environ.* **2015**, *512*, 364–370. https://doi.org/10.1016/j.scitotenv.2015.01.042.
11. Fu, Y.; Wang, T.; Fu, Q.; et al. Associations between serum concentrations of perfluoroalkyl acids and serum lipid levels in a Chinese population. *Ecotoxicol. Environ. Saf.* **2014**, *106*, 246–252. https://doi.org/10.1016/j.ecoenv.2014.04.039.
12. Nelson, J.W.; Hatch, E.E.; Webster, T.F. Exposure to polyfluoroalkyl chemicals and cholesterol, body weight, and insulin resistance in the general U.S. population. *Environ. Health Perspect.* **2010**, *118*, 197–202. https://doi.org/10.1289/ehp.0901165.
13. Steenland, K.; Woskie, S. Cohort mortality study of workers exposed to perfluorooctanoic acid. *Am. J. Epidemiol.* **2012**, *176*, 909–917. https://doi.org/10.1093/aje/kws171.
14. Leonard, R.C.; Kreckmann, K.H.; Sakr, C.J.; et al. Retrospective cohort mortality study of workers in a polymer production plant including a reference population of regional workers. *Ann. Epidemiol.* **2008**, *18*, 15–22. https://doi.org/10.1016/j.annepidem.2007.06.011.
15. Barry, V.; Winquist, A.; Steenland, K. Perfluorooctanoic acid (PFOA) exposures and incident cancers among adults living near a chemical plant. *Environ. Health Perspect.* **2013**, *121*, 1313–1318. https://doi.org/10.1289/ehp.1306615.
16. Vieira, V.M.; Hoffman, K.; Shin, H.M.; et al. Perfluorooctanoic acid exposure and cancer outcomes in a contaminated community: A geographic analysis. *Environ. Health Perspect.* **2013**, *121*, 318–323. https://doi.org/10.1289/ehp.1205829.
17. Kirk, M.; Smurthwaite, K.; Bräunig, J.; et al. *The PFAS Health Study: Systematic Literature Review*; The Australian National University: Canberra, ACT, Australia, 2018.
18. Mogensen, U.B.; Grandjean, P.; Heilmann, C.; et al. Structural equation modeling of immunotoxicity associated with exposure to perfluorinated alkylates. *Environ. Health* **2015**, *14*, 47. https://doi.org/10.1186/s12940-015-0032-9.

19. Grandjean, P.; Heilmann, C.; Weihe, P.; et al. Serum vaccine antibody concentrations in adolescents exposed to perfluorinated compounds. *Environ. Health Perspect.* **2017**, *125*, 077018. https://doi.org/10.1289/ehp275.
20. Kielsen, K.; Shamim, Z.; Ryder, L.P.; et al. Antibody response to booster vaccination with tetanus and diphtheria in adults exposed to perfluorinated alkylates. *J. Immunotoxicol.* **2016**, *13*, 270–273. https://doi.org/10.3109/1547691x.2015.1067259.
21. NTP. *NTP Monograph: Immunotoxicity Associated with Exposure to Perfluorooctanoic Acid (PFOA) or Perfluorooctane Sulfonate (PFOS)*; U.S. Department of Health and Human Services, Office of Health Assessment and Translation: Research Triangle Park, NC, USA, 2016.
22. Nahar, K.; Zulkarnain, N.A.; Niven, R.K. A review of analytical methods and technologies for monitoring per- and polyfluoroalkyl substances (PFAS) in water. *Water* **2023**, *15*, 3577. https://doi.org/10.3390/w15203577.
23. Joerss, H.; Menger, F. The complex 'PFAS world'—How recent discoveries and novel screening tools reinforce existing concerns. *Curr. Opin. Green Sustain. Chem.* **2023**, *40*, 100775. https://doi.org/10.1016/j.cogsc.2023.100775.
24. Brunn, H.; Arnold, G.; Körner, W.; et al. PFAS: Forever chemicals—Persistent, bioaccumulative and mobile. Reviewing the status and the need for their phase out and remediation of contaminated sites. *Environ. Sci. Eur.* **2023**, *35*, 20. https://doi.org/10.1186/s12302-023-00721-8.
25. Williams, A.J.; Grulke, C.M.; Edwards, J.; et al. The CompTox chemistry dashboard: A community data resource for environmental chemistry. *J. Cheminf.* **2017**, *9*, 61. https://doi.org/10.1186/s13321-017-0247-6.
26. Schymanski, E.L.; Zhang, J.; Thiessen, P.A.; et al. Per- and polyfluoroalkyl substances (PFAS) in PubChem: 7 million and growing. *Environ. Sci. Technol.* **2023**, *57,* 16918–16928. https://doi.org/10.1021/acs.est.3c04855.
27. EPA. *PFAS Analytical Methods Development and Sampling Research*; EPA: Washington, DC, USA, 2025.
28. Koelmel, J.P.; Stelben, P.; McDonough, C.A.; et al. FluoroMatch 2.0—Making automated and comprehensive non-targeted PFAS annotation a reality. *Anal. Bioanal. Chem.* **2022**, *414*, 1201–1215. https://doi.org/10.1007/s00216-021-03392-7.
29. Nguyen, T.M.H.; Bräunig, J.; Kookana, R.S.; et al. Assessment of mobilization potential of per- and polyfluoroalkyl substances for soil remediation. *Environ. Sci. Technol.* **2022**, *56*, 10030–10041. https://doi.org/10.1021/acs.est.2c00401.
30. Dewapriya, P.; Nilsson, S.; Ghorbani Gorji, S.; et al. Novel per- and polyfluoroalkyl substances discovered in cattle exposed to AFFF-impacted groundwater. *Environ. Sci. Technol.* **2023**, *57*, 13635–13645. https://doi.org/10.1021/acs.est.3c03852.
31. Ghorbani Gorji, S.; Gómez Ramos, M.J.; Dewapriya, P.; et al. New PFASs identified in AFFF impacted groundwater by passive sampling and nontarget analysis. *Environ. Sci. Technol.* **2024**, *58*, 1690–1699. https://doi.org/10.1021/acs.est.3c06591.
32. Barzen-Hanson, K.A.; Roberts, S.C.; Choyke, S.; et al. Discovery of 40 classes of per- and polyfluoroalkyl substances in historical aqueous film-forming foams (AFFFs) and AFFF-impacted groundwater. *Environ. Sci. Technol.* **2017**, *51*, 2047–2057. https://doi.org/10.1021/acs.est.6b05843.
33. Bugsel, B.; Zweigle, J.; Zwiener, C. Nontarget screening strategies for PFAS prioritization and identification by high resolution mass spectrometry: A review. *Trends Environ. Anal. Chem.* **2023**, *40*, e00216. https://doi.org/10.1016/j.teac.2023.e00216.
34. Charbonnet, J.A.; McDonough, C.A.; Xiao, F.; et al. Communicating confidence of per- and polyfluoroalkyl substance identification via high-resolution mass spectrometry. *Environ. Sci. Technol. Lett.* **2022**, *9*, 473–481. https://doi.org/10.1021/acs.estlett.2c00206.
35. Zweigle, J.; Bugsel, B.; Fabregat-Palau, J.; et al. PFΔScreen—An open-source tool for automated PFAS feature prioritization in non-target HRMS data. *Anal. Bioanal. Chem.* **2024**, *416*, 349–362. https://doi.org/10.1007/s00216-023-05070-2.
36. Samanipour, S.; O'Brien, J.W.; Reid, M.J.; et al. Self adjusting algorithm for the nontargeted feature detection of high resolution mass spectrometry coupled with liquid chromatography profile data. *Anal. Chem.* **2019**, *91*, 10800–10807. https://doi.org/10.1021/acs.analchem.9b02422.
37. Samanipour, S.; Reid, M.J.; Baek, K.; et al. Combining a deconvolution and a universal library search algorithm for the non-target analysis of data independent LC-HRMS spectra. *Environ. Sci. Technol.* **2018**, *52*, 4694–4701. https://doi.org/10.1021/acs.est.8b00259.
38. Horai, H.; Arita, M.; Kanaya, S.; et al. MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, *45*, 703–714. https://doi.org/10.1002/jms.1777.
39. Enders, J.R.; O'Neill, G.M.; Whitten, J.L.; et al. Understanding the electrospray ionization response factors of per- and polyfluoroalkyl substances (PFAS). *Anal. Bioanal. Chem.* **2021**, *414*, 1227–1234. https://doi.org/10.1007/s00216-021-03545-8.
40. Wolfram. Gaussian Function. Available online: https://mathworld.wolfram.com/GaussianFunction.html (accessed on 1 June 2023).
41. Kendrick, E. A mass scale based on $CH_2$ = 14.0000 for high resolution mass spectrometry of organic compounds. *Anal. Chem.* **1963**, *35*, 2146–2154. https://doi.org/10.1021/ac60206a048.
42. Merel, S. Critical assessment of the Kendrick mass defect analysis as an innovative approach to process high resolution mass spectrometry data for environmental applications. *Chemosphere* **2023**, *313*, 137443. https://doi.org/10.1016/j.chemosphere.2022.137443.
43. Liu, Y.; Pereira, A.D.S.; Martin, J.W. Discovery of C5–C17 poly- and perfluoroalkyl substances in water by in-line SPE-HPLC-Orbitrap with in-source fragmentation flagging. *Anal. Chem.* **2015**, *87*, 4260–4268. https://doi.org/10.1021/

acs.analchem.5b00039.
44. Tang, C.; Liang, Y.; Wang, K.; et al. Comprehensive characterization of per- and polyfluoroalkyl substances in wastewater by liquid chromatography-mass spectrometry and screening algorithms. *npj Clean Water* **2023**, *6*, 6. https://doi.org/10.1038/s41545-023-00220-6.
45. Su, A.; Rajan, K. A database framework for rapid screening of structure-function relationships in PFAS chemistry. *Sci. Data* **2021**, *8*, 14. https://doi.org/10.1038/s41597-021-00798-x.
*46.* Landrum, G.; Tosco, P.; Kelley, B.; et al. *rdkit/rdkit: 2025_09_6 (Q3 2025) Release*; CERN: Genève, Switzerland, 2023. https://zenodo.org/records/18797641.
47. Ke, G.; Meng, Q.; Finley, T.; et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 3146–3154.
48. Ke, G. LightGBM Parameters. Available online: https://lightgbm.readthedocs.io/en/latest/Parameters.html (accessed on 1 June 2024).
49. Van Rijsbergen, C.J. Information Retrieval: Theory and Practice. In Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems, Oxford, UK, 4–7 September 1979.
50. Schymanski, E.L.; Jeon, J.; Gulde, R.; et al. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environ. Sci. Technol.* **2014**, *48*, 2097–2098. https://doi.org/10.1021/es5002105.
51. Chambers, M.C.; Maclean, B.; Burke, R.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30*, 918–920. https://doi.org/10.1038/nbt.2377.
52. Meekel, N.; Kruve, A.; Lamoree, M.H.; et al. Machine learning-based classification for the prioritization of potentially hazardous chemicals with structural alerts in nontarget screening. *Environ. Sci. Technol.* **2025**, *59*, 5056–5065. https://doi.org/10.1021/acs.est.4c10498.
53. Strynar, M.; McCord, J.; Newton, S.; et al. Practical application guide for the discovery of novel PFAS in environmental samples using high resolution mass spectrometry. *J. Expo. Sci. Environ. Epidemiol.* **2023**, *33*, 575–588. https://doi.org/10.1038/s41370-023-00578-2.
54. Retention Time in Chromatography and Its Role in Analysis. Available online: https://biologyinsights.com/retention-time-in-chromatography-and-its-role-in-analysis/ (accessed on 1 June 2024).
55. Olsen, J.V.; De Godoy, L.M.; Li, G.; et al. Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a c-trap. *Mol. Cell. Proteomics* **2005**, *4*, 2010–2021. https://doi.org/10.1074/mcp.T500030-MCP200.
56. DasGupta, A. *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*; Springer: New York, NY, USA, 2011. https://doi.org/10.1007/978-1-4419-9634-3.
57. Jolliffe, I.T. *Principal Component Analysis*; Springer: New York, NY, USA, 1986.
58. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B Stat. Methodol.* **1996**, *58*, 267–288.
59. Moka, S.; Liquet, B.; Zhu, H.; et al. COMBSS: Best subset selection via continuous optimization. *Stat. Comput.* **2024**, *34*, 75. https://doi.org/10.1007/s11222-024-10387-8.
60. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
61. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; et al. CatBoost: unbiased boosting with categorical features. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; Vol. 31.
62. Jia, S.; Marques Dos Santos, M.; Li, C.; et al. Recent advances in mass spectrometry analytical techniques for per- and polyfluoroalkyl substances (PFAS). *Anal. Bioanal. Chem.* **2022**, *414*, 2795–2807. https://doi.org/10.1007/s00216-022-03905-y.
63. Allen, F.; Greiner, R.; Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **2015**, *11*, 98–110. https://doi.org/10.1007/s11306-014-0676-4.