PFAS Prophet - Prioritising Per- and Polyfluoroalkyl Substances Using Mass and MS - MS Spectra Supporting information

Text

Text S1. Synthetic measured m/z of parent ion rational

Justification for the resolution choice (R = 30,000)

By using the theoretical peak width as the mass error, the synthetic m/z generation accommodates data acquired at lower resolution. A well-calibrated HRMS instrument can determine the peak centre with a precision significantly better than the FWHM itself, since centroid accuracy scales with the signal-to-noise ratio and the number of data points across the peak.

the Gaussian error assumption

The assumption of a Gaussian error distribution is supported by empirical studies of HRMS mass accuracy. Orbitrap mass errors have been shown to follow an approximately Gaussian distribution with nearly all values within ±2 ppm using lock mass calibration[1], while 5 ppm accuracy is achieved with >95% probability at a resolution of 30,000 on an Orbitrap analyser[1]. Modern QTOF instruments routinely achieve sub-5 ppm accuracy. In practice, the synthetic error term used here is substantially larger than these reported instrument accuracies, providing a conservative upper bound that accommodates variation across instrument types, calibration states, and operating conditions.

Text S2 Homologous series

Rationale for Kendrick Mass Defect (KMD)

The synthetic measured m/z of the parent ion encodes the exact mass of an individual compound and is therefore highly compound-specific. A model trained on raw m/z values would be forced to memorise each PFAS reference compound as a distinct point in m/z space, which generalises poorly given the limited number of PFAS reference spectra. The KMD provides a coordinate transformation in which homologous compounds — those differing only in repeat units of a chosen building block — collapse onto near-horizontal lines in the (nominal mass, KMD) plane. This converts a memorisation problem into a pattern-recognition problem and is the foundation of homologue-series detection in non-target PFAS screening[2,3].

Equations

$$r(x,d) \coloneqq \frac{\lfloor x \times 10^{d} + 0.5 \rfloor}{10^{d}}$$

$$MM(em) = em - \mathrm{p}$$

$$KM(em,\mathrm{C}) = MM \times \frac{r(C,0)}{C}$$

$$KMD(em,\mathrm{C}) = r(r(KM(em,\mathrm{C}),0) - KM(em,\mathrm{C}),3)$$

For each compound, the following *KMD* were obtained.

$KMD(\mathrm{em},CF)$
$KMD(\mathrm{em},CF_2)$
$KMD(\mathrm{em},CF_2O)$
$KMD(\mathrm{em},C_2F_4O)$
$KMD(\mathrm{em},\mathrm{CF})$
$KMD(\mathrm{em},C_2H_2F_2)$
$KMD(\mathrm{em},C_3H_6O)$
$KMD(\mathrm{em},\mathrm{CN})$
$KMD(\mathrm{em},\mathrm{CO})$
$KMD(\mathrm{em},CH_2)$
$KMD(\mathrm{em},CCl)$
$KMD(\mathrm{em},CS)$

Where:

Abbreviations:

$r(x,d) : rounding\ x\ to\ d\ decimal\ places$
$KM : Kendric\ \ Mass$
$KMD : Kendrick\ \ Mass\ \ Defect$
C : $Repeating\ unit\ compound$ mass
p : $1.0073\ Da\ (Proton\ Mass)$
$MM : Measured\ \ Mass$

Constants:

Repeating unit compound mass:

CF: $30.9984Da$
CF_2: $49.9968Da$
CF_2O: $65.9917Da$
C_2F_4O: $115.9885Da$
$\mathrm{C}_2H_2F_2$: $64.0125Da$
C_3F_6O: $165.9853Da$
CO:$27.9949Da$
CN: $26.0031Da$
CCI: $138.9045Da$
CS: $43.9721Da$
CH_2: $14.0156Da$

variables:

$em : synthetic\ measured\ mass$

x, d *variables for rounding function where* x *is the number to be rounded and* d *is the significant digit wanted.*

Choice of repeat units

Twelve repeat units were used: seven fluorinated (CF, CF_2, CF_2O, C_2F_4O, CF, $C_2H_2F_2$, C_3F_6O) chosen because they correspond to the dominant homologous building blocks in environmental PFAS - perfluoroalkyl chains, ether-linked perfluoro segments, and partially fluorinated motifs; and five non-fluorinated (CN, CO, CH_2, CCl, CS) chosen to give the model a contrastive view of non-PFAS homologue families that share mass-defect behaviour with PFAS in narrow regions of m/z space. Including both families lets the model use KMD as a discriminative rather than purely descriptive feature: PFAS-like KMD signatures are informative only when

contrasted with the KMD signatures of common non-PFAS organics.

Text S3: Oversampling rational

Why oversampling is needed.

The *MassBank* dataset currently contains only 177 distinct PFAS-related compounds (after filtering), with some having very few mass spectral reference data for representation, and a few having only one spectrum. As such, the representation of the synthetic measured m/z range of these PFAS-related compounds is very limited for the training set, and the model will have a limited understanding of the underlying error distribution.

Text S4: Binning Rationale

Why dimensionality reduction was not used.

With potentially 1.8 million distinct fragment entries to select as input variables, the number of candidate features far exceeds the number of observations, making the data unsuitable for most ML models due to the curse of dimensionality***[4]***. Three standard dimensionality reduction methods were considered. PCA[5] convolutes the variable inputs, which makes them abstruse, requires an extra transformation of the HRMS data to be suited for the model, makes interpretability challenging, and makes no use of the response variable. LASSO[6] and COMBSS[7] only select a much smaller subset of variables based on linear transformations, which can potentially lose valuable information if the correlations are non-linear. For these reasons an alternative approach – binning – was selected.

binning.

Each fragment represents an m/z measurement of a compound. Ideally, all measurements of the same fragment would be represented by the same fragment measurement; however, due to measurement error, fragment measurements of the same compound may vary, and thus the variability of the distinct measurements comes not only from distinct fragments but also from error in the measurements. Utilising distinct measurements of each fragment would yield suboptimal variable inputs since it is unlikely that the same compound is consistently represented by the same fragment measurement. A more ideal solution is therefore to find a method to form groups in which all fragments are chemically identical.

Bin width as a hyperparameter

Estimating the optimal bin size is difficult because the fragments have distinct error distributions due to their differing m/z, resolution, and instrument settings. Finding a "best fit" bin given the scope of the problem is complex, as there are multiple reasons why the algorithm might prefer larger or smaller bins than expected. It is therefore preferable to treat bin width as a hyperparameter and let the model determine the optimal value based on validation set performance.

Minimum occurrence threshold

Each distinct bin is used as an input for the model. The model needs multiple occurrences of fragments inside a bin for the bin to be included as an explanatory variable; this threshold is set as a hyperparameter.

Asymmetric treatment of PFAS-related fragments

Only fragments associated with at least one parent compound in a PFAS-related class were retained. Although this parameter potentially lowers overall prediction accuracy, it makes the model more general by preventing it from predicting class solely on the presence of a fragment. This is an important step as the model has limited representation of the PFAS fragmentation space; without this filter, uncharacterised PFAS would likely contain fragments in bins that the model only saw as non-PFAS during training. The inverse (removing fragments only seen in PFAS-related but not non-PFAS-related compounds) was not performed because the PFAS-related class only represents 177 compounds in the dataset whereas the non-PFAS class has over 50,000 compounds; the likelihood that a fragment seen only in PFAS is not actually correlated with PFAS is therefore much smaller. This biases the model slightly toward false positives, which is preferable as it also makes the model better at detecting new PFAS.

Text S5 : Rational for model selection

LightGBM is a decision tree-based gradient boosting framework. While XGBoost[8] also had potential, LightGBM was chosen as its main strength is its fast model-building capabilities which is particularly crucial given the need to tune numerous hyperparameters and preprocess the data extensively without loss in model strength[8]. CatBoost was also tested but found to be extremely slow when confronted with a large number of input variables[9].

Text S6 : Cross-Validation (CV) Datasets

The scheme described in the Methods is a grouped 10-fold cross-validation with a rotating three-way split. At each of 10 iterations, one-fold is the test set, one is the validation set used for hyperparameter selection, and the remaining eight are the training set. The test fold is excluded from both training and hyperparameter tuning at its iteration.

The scheme is not fully nested. In a fully nested 10-fold CV, the validation fold is rotated within each outer iteration and hyperparameter performance averaged across those inner rotations, requiring approximately 90 model fits per hyperparameter configuration instead of 10. Hyperparameter selection in our scheme depends on a single validation fold per iteration rather than an averaged inner estimate. The selected hyperparameters are therefore noisier than those a fully nested search would produce, and the configuration chosen at a given iteration may differ from the optimum identified under nested CV. Reported test-fold performance reflects models trained under this hyperparameter-selection procedure.

The grouping constraint requires that all spectra from a given compound remain in a single fold, since spectra from the same compound appearing in both training and test folds would produce optimistic estimates relative to performance on unseen compounds. A rotated inner loop on top of the grouped 10-fold structure would increase the hyperparameter grid search from 7,560 to approximately 68,000 trained models.

Text S7 : Discussion – Model - Input variable comparisons

KMD was utilized as it provides better generalization, capturing a larger number of instances within the PFAS class compared to *m/z* values. Given the limited diversity of distinct compounds available, KMD proves highly

generalizable across the compounds of interest [10]. In contrast, employing *m/z* would have likely led to compound-specific outcomes, potentially resulting in poor performance.

The model exhibits a strong preference towards KMD (as observed from the high gain and split proportions from the model), which aligns with its focus on specific KMD ranges for PFAS-related values. By exclusively utilizing KMD values, the model can effectively rule out numerous compounds as PFAS-related. However, fragment values remain crucial for discerning PFAS-related compounds. This is evident when examining individual test set results for distinct spectrum sets belonging to the same compound. The spectrum that has been oversampled exhibit minimal variation between samples (typically less than 3% difference in prioritisation score of being PFAS) only attributable to variations in pseudo-measured *m/z* calculations (which is reflected in KMD) as it is this the only variable value that is distinct between them. However, there is substantial variation between sets of different spectra of the same compound, often exceeding 34% difference in reported probability. This discrepancy is extremely unlikely to come from the differences in pseudo-measured *m/z*, as the error terms are drawn from the same distribution. Therefore, the variation can only be explained by differences in fragments. Observing the example trees of the model, all 4 trees start with a fragment feature rather than a KMD feature, which indicates that for the training set given for those trees, the fragment value was more informative for the trees than the KMD values in deciding if the spectra are PFAS or not.

Text S8: Discussion - Limitations – Data quality

Most of the spectra that did not report resolution were still included in the dataset. Manual inspection of these entries indicated that most are from HRMS instruments, and those suspected to be of low resolution were excluded. Because the mantissa-based filter cannot fully distinguish HRMS from low-resolution data (low-resolution instruments can also report m/z values to several decimal places), it is likely that some low-resolution spectra were retained. However, the bin width used during training (0.09 m/z) is substantially larger than the worst-case fragment standard deviation expected even at reduced resolution (≈0.028 m/z at m/z 1,000 and R = 15,000; see Fixed resolution below), so any residual low-resolution entries are largely absorbed by the binning tolerance. Some of the spectra used appear to have minimal second stage of mass spectrometry (MS2) curation. Additionally, there was no measurement of the MS_1 which necessitated an estimation of measured *m/z*. Since a fixed resolution was applied during *m/z* estimation, the resolution of the fragments remained unchanged, presenting another distinction between the training data and NTA data. It is also important to note that a resolution of 30,000 was used for the estimated spectrum mass, however, the reported fragment measurement was unchanged and likely to have been obtained using a different resolution.

Text S9: Discussion - Limitations - continued

Synthetic measured^{i} m/z

The synthetic measured m/z used in this study is derived by adding a normally distributed error term to the exact mass, based on the theoretical mass accuracy at a fixed resolution of 30,000. While empirical studies confirm that HRMS mass errors are approximately Gaussian and that modern instruments routinely achieve

accuracies well within the bounds of our synthetic error term, this approach does not capture the full complexity of real instrument error. In practice, mass accuracy is influenced by signal intensity, calibration drift, ion suppression, and matrix effects, factors that can introduce systematic biases not represented by a simple Gaussian model. Ideally, the synthetic error would be calibrated against empirical error distributions obtained from diverse instruments under varying operating conditions. However, this would require extensive characterisation across multiple instrument platforms, ionisation modes, and sample matrices, effectively constituting a separate validation study. By instead using a deliberately conservative error term that exceeds typical reported instrument accuracies, the model is designed to accommodate a broad range of instrument performance. The external test set, which was acquired on a QTOF instrument at a resolution of 25,000, demonstrates that this conservative approach produces reliable predictions on real-world measured data despite the idealised error model.

Fixed resolution

When computing the error term to derive the synthetic measured m/z, utilizing the resolution reported in the database per spectra was considered against using a fixed resolution of 30,000. Several rationales supported the selection of the latter approach:

With over two-thirds of the data lacking reported resolution, assuming a resolution becomes necessary. Upon inspection, most of these datasets suggested higher resolutions (30000+), while those suspected to be of low resolution were removed.

Because there are no reported measurements of the compound *m/z*, the resolution of the spectra only informs the estimated error of the fragments. Given the considerably larger size of the bins (0.09 *m/z*) relative to the likely error range of the fragments, this minimalizes the influence of errors. Even in the worst-case scenario, with an *m/z* of 1,000 at a resolution of 15000, the standard deviation of the measurement is of 0.028 *m/z*, which, while not insignificant compared to the bin size, applies to only a small fraction of fragments.

Adopting a fixed resolution of 30,000 establishes a clearer parameter for its application, restricting its use to spectra with resolutions of 30,000 or higher. However, this could be lowered to 25,000 as the external test set performed very well even though it only had a resolution of 25,000.

Bin width

The optimal bin width appears to fall within the range of 0.01 Da to 0.1 Da. Within this range, certain values, such as 0.02 Da and 0.04 Da, were evidently suboptimal, while others, like 0.03 Da, 0.06 Da, and 0.09 Da, performed significantly better. This observation suggests that specific bin values may better represent crucial fragments.

Interestingly, the model exhibited a preference for bin widths larger than expected from the fragment resolution. Although larger bins reduce the likelihood of splitting identical fragments into different bins, it also increases the likelihood of many bins containing distinct fragments. Consequently, the model cannot effectively utilize the distinction between fragments in the same bin. However, looking at some of the example trees, the utility of most fragments arises only after KMD has been utilized. Therefore, for two distinct fragment values to

benefit the model, they must originate from compounds with distinct classes with similar enough KMD values that the model cannot distinguish them from KMD values alone and thus needs to use fragments. Given that the scenario above has a low likelihood, the model only needs to separate these crucial fragments and thus is a possible reason larger bins are preferred.

While the error variance of the neutral loss is higher due to being the sum of the variance of the error of the fragment and error of the pseudo-measured *m/z*, expert opinion suggests that neutral losses may be more beneficial to the algorithm than fragments. This is because different PFAS tend to share common neutral losses that are distinct from non-PFAS losses. Initially it was considered to only use neutral loss values and disregard fragments, as combining both would result in a wider table, complicating the task of discerning signal from noise. However, testing indicated that the algorithm performs best when provided with both fragment and neutral loss lists.

Model Training

Employing a ML model with training, validation, and test sets involves maximizing the model's potential by conducting extensive testing and comparing various ideas to determine the optimal configurations that maximize the validation metric. However, this task is complex as it involves not only optimizing the model parameters but also other data-dependent hyperparameters. We chose to employ a nested 10-fold cross-validation to maximize the utilization of the limited data, necessitating the training of 10 distinct models for each parameter set.

Given the number of hyperparameters to tune, a model that can converge in a timely manner is important as it will allow for a more refined parameter selection. Furthermore, a strategic approach for parameter selection, along with an efficient model is crucial for parameter optimization. To better understand why such an approach was needed, if all 14 of the hyperparameters were binary, and it takes roughly two minutes for a model to be built, then there are $2(\text{binary})^{14(\text{models})}$x10(CV iterations) models to be trained, it would take in total 227.5 days of computing to evaluate all possible combinations. Given that most of the parameters are not binary, with some having up to ten distinct values, the actual number is orders of magnitude higher.

Consequently, conducting a proper grid search on the parameters proved too demanding and thus only a limited grid search was performed.

Despite these acknowledged challenges, a model constructed using this dataset remains valuable for analysing data obtained directly from NTS, as evidenced by its performance on the external test set.

Text S10: Discussion - Potential future development – continued

Losses and Fragments

The model heavily relies on the utilization of neutral losses and fragments from the spectra. However, the current method of classifying fragments through binning is inherently simplistic and prone to errors. Bins can easily split identical fragments into distinct bins or bin multiple distinct fragments together, leading to inaccuracies. Improving the bin classification method is crucial for enhancing the model's performance.

To address this issue, more complex methods could be investigated, such as an adapted Mean-Shift algorithm

with a variable radius dependent on compound *m/z*. Alternatives include allowing bins to overlap slightly by the estimated error at that range, ensuring that compounds with ambiguous weights are still represented accurately, or adjusting bin widths based on the estimated standard deviation of the compound *m/z*. For instance, larger *m/z* with higher standard deviations could have larger bins.

Other considerations would be obtaining KMDs of the losses and incorporating a column containing the lowest KMD or separating the losses and fragments into distinct bins. However, this would require a cautious approach as subtracting fragment *m/z* from parent *m/z* may not always yield a neutral loss.

Use of Synthetic data

The use of synthetic spectral data was not considered at first as it was suspected that synthetic data is still not accurate enough to represent what would be obtained from actual data. However, if this assumption is wrong, then if synthetic data were to be exclusively used, many of these limitations could be resolved. Furthermore, data preparation for the synthetic data would is more challenging as there are no measured data, and the dataset is considerably larger.

Data sparsity

The CFM-ID [11] dataset represents a collection of 700,000 distinct compounds, many of which are PFAS. This very high level of distinct compounds would solve the data sparsity issue that the experimental dataset experienced.

Exact fragments and losses

Because data from CFM-ID has exact measurements for fragments and parent compound, there is no need to develop an algorithm to bin likely measurements together. However, given that the algorithm will receive measurements containing error, an algorithm would be needed to indicate to the model how likely it was that a measured fragment would be in a given bin.

Resolution as a parameter

Because with experimental data the fragments are measured, and their exact mass is unknown, these cannot be varied in the same way the synthetic measured mass was. However, with synthetic data, an error term can be provided to all measurements that are resolution dependent. As a result, not only would the resulting estimated fragment error be more accurate to the resolution than the experimental data was, it also would allow use of the resolution as a parameter. This would allow resolution-specific ML algorithms, which would likely result in more accurate models.

Collision energy as a parameter

For the ML algorithm using experimental data, the spectra were obtained through methods that used different collision energies. Synthetic data can produce fragmentation patterns that are obtained from a distinct collision energy. This has the advantage of allowing the collision energy to be used as a parameter as well, which in turn, will likely result in more accurate models.

Figures

Figure S1

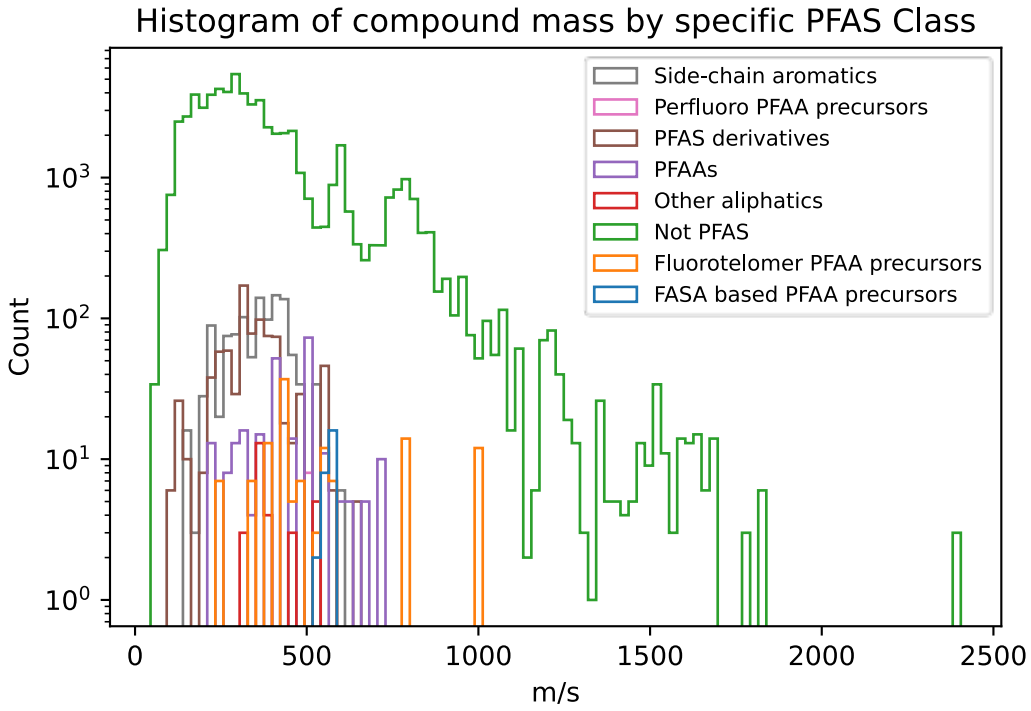


Figure S1: Histogram of compound m/z per class, grouped by specific PFAS classes and non-PFAS class, after removal of suspected low-quality spectra.

Figure S2

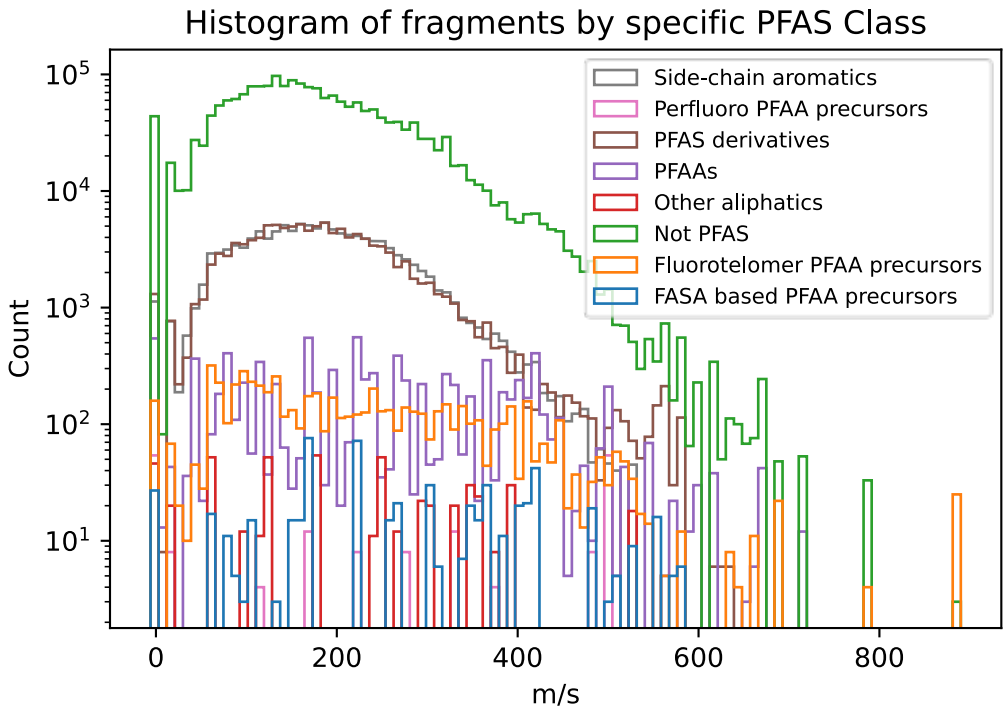


Figure S2: Histogram of the collection of fragments and losses of the dataset, grouped by specific PFAS classes and non-PFAS class, after removal of suspected low-quality spectra and removal of fragments which do not appear in any PFAS spectra.

Figure S3

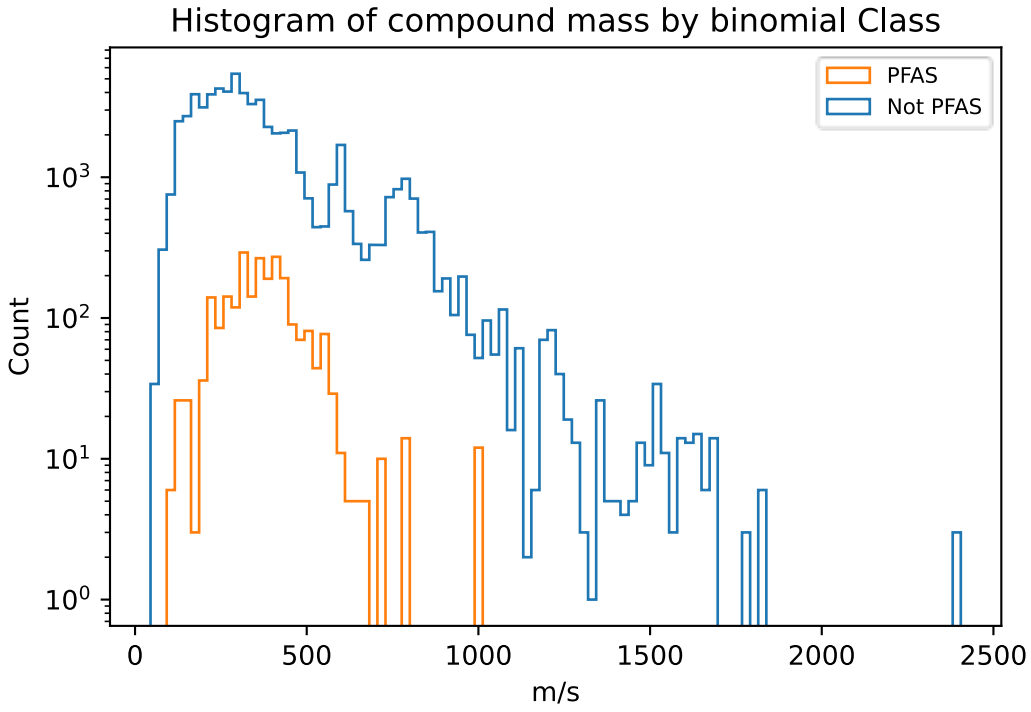


Figure S3: Histogram of compound m/z per binomial class, grouped by PFAS and non-PFAS class, after removal of suspected low-quality spectra.

Figure S4

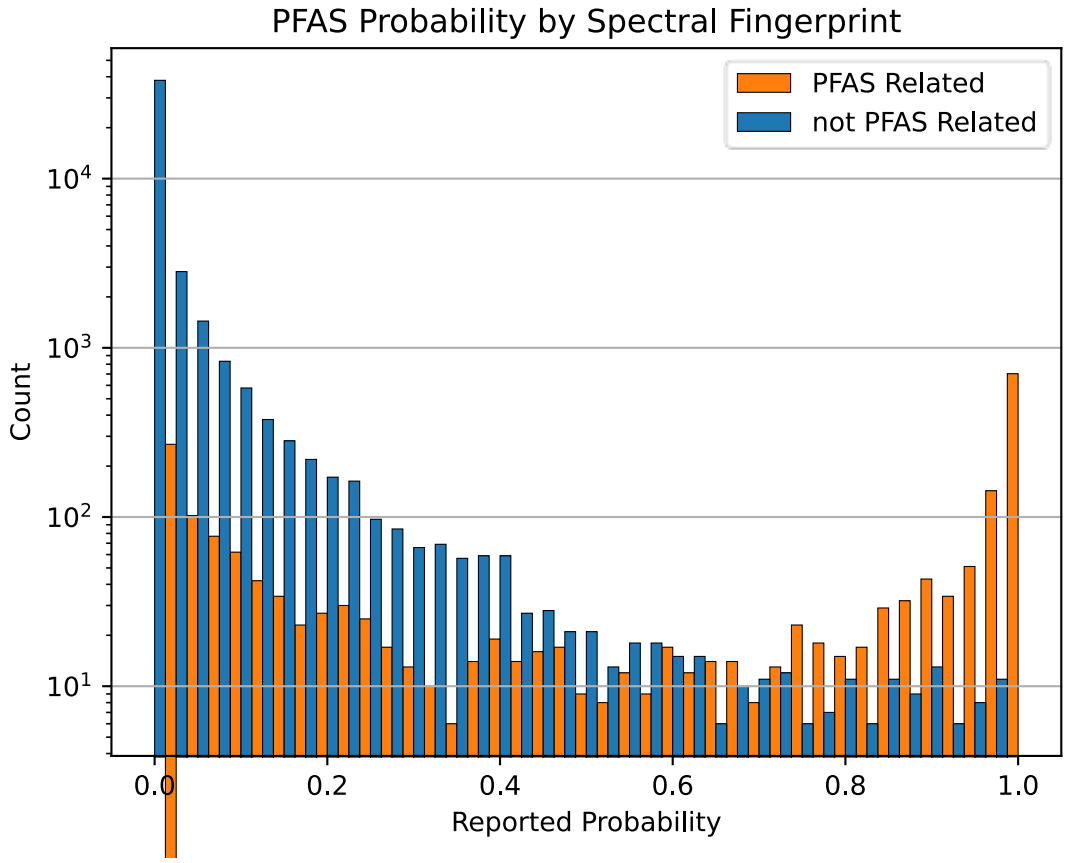


Figure S4 : Histogram of the Test set model reported probabilities of classes per spectrum. Bins are of size 0.025.

Figure S5

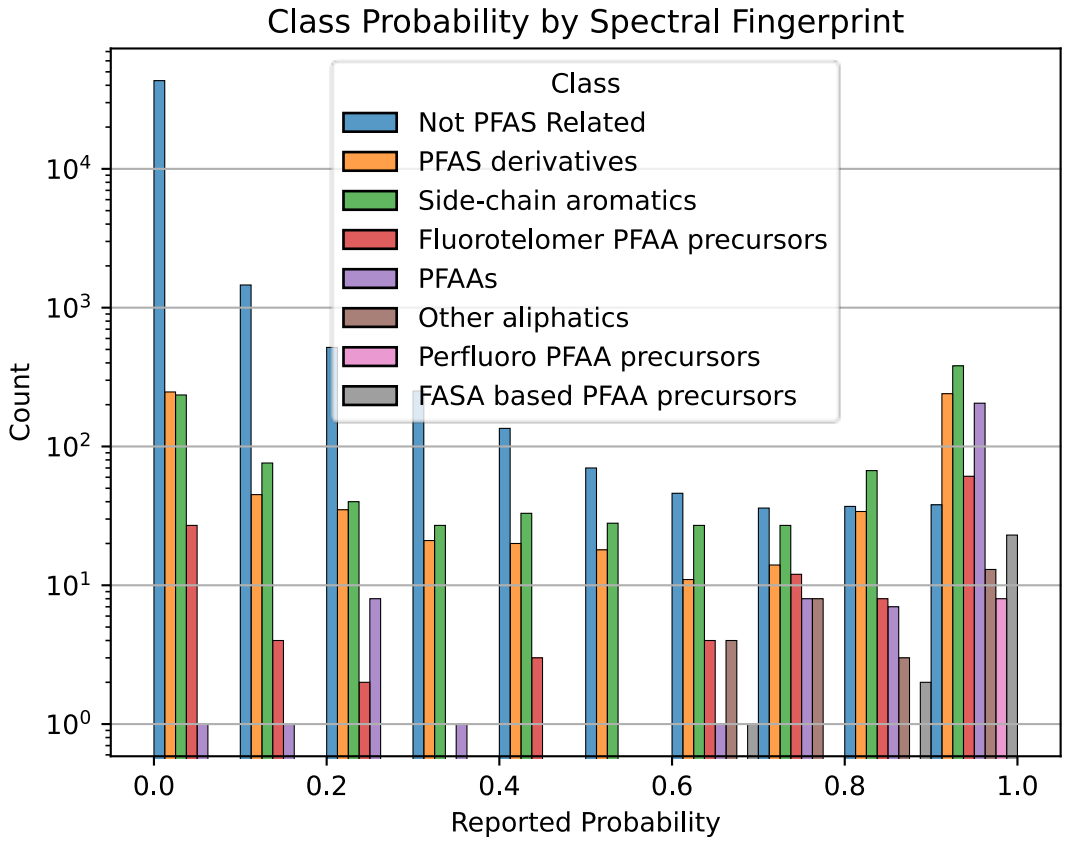


Figure S5: Histogram of the spectrum Test set model reported probabilities by Class. Bins are of size 0.1

Figure S6

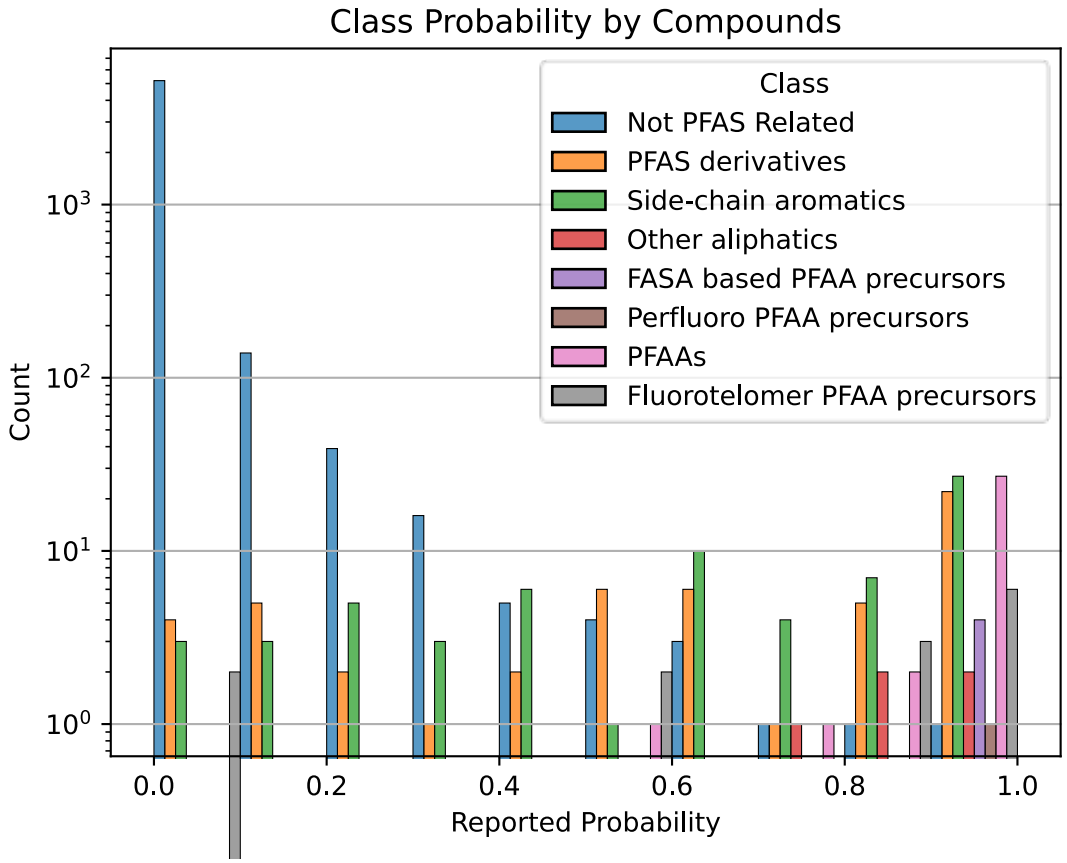


Figure S6: Histogram of the Compounds Test set model reported probabilities by Class. Bins are of size of 0.1.

Figure S7

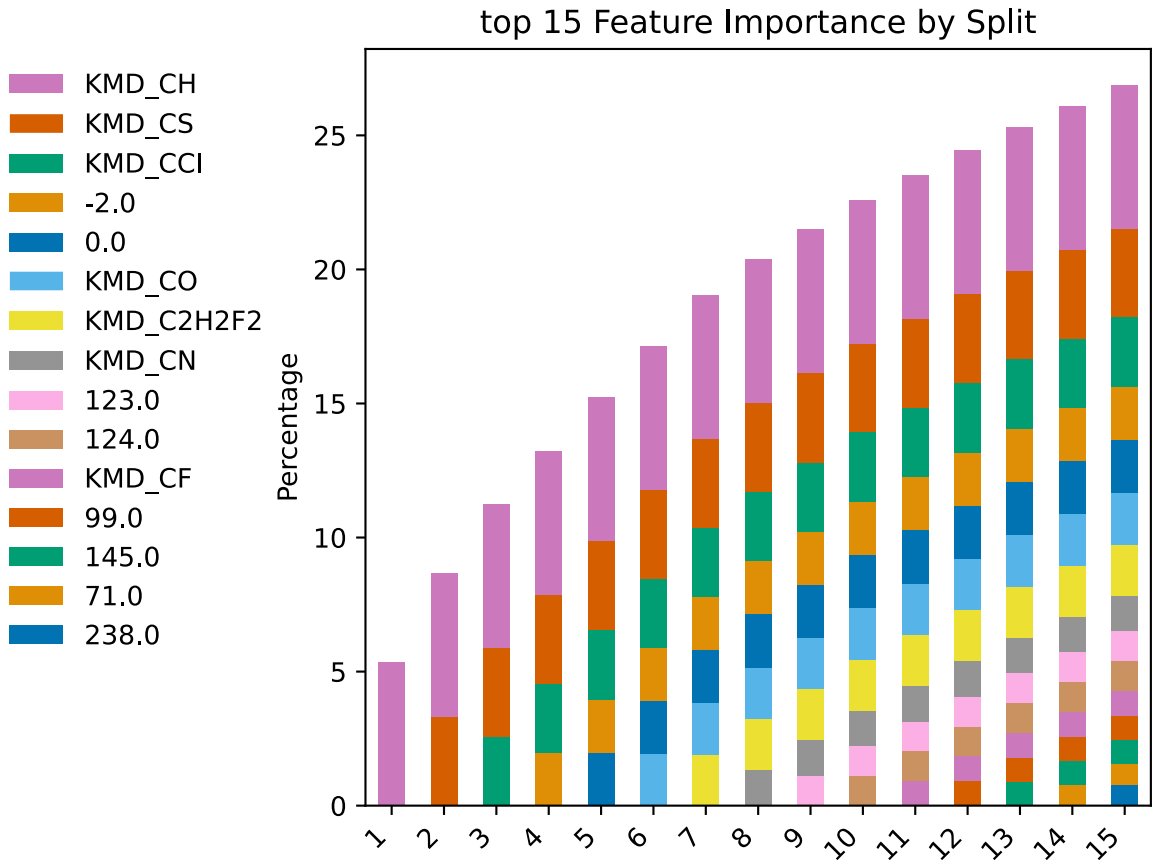
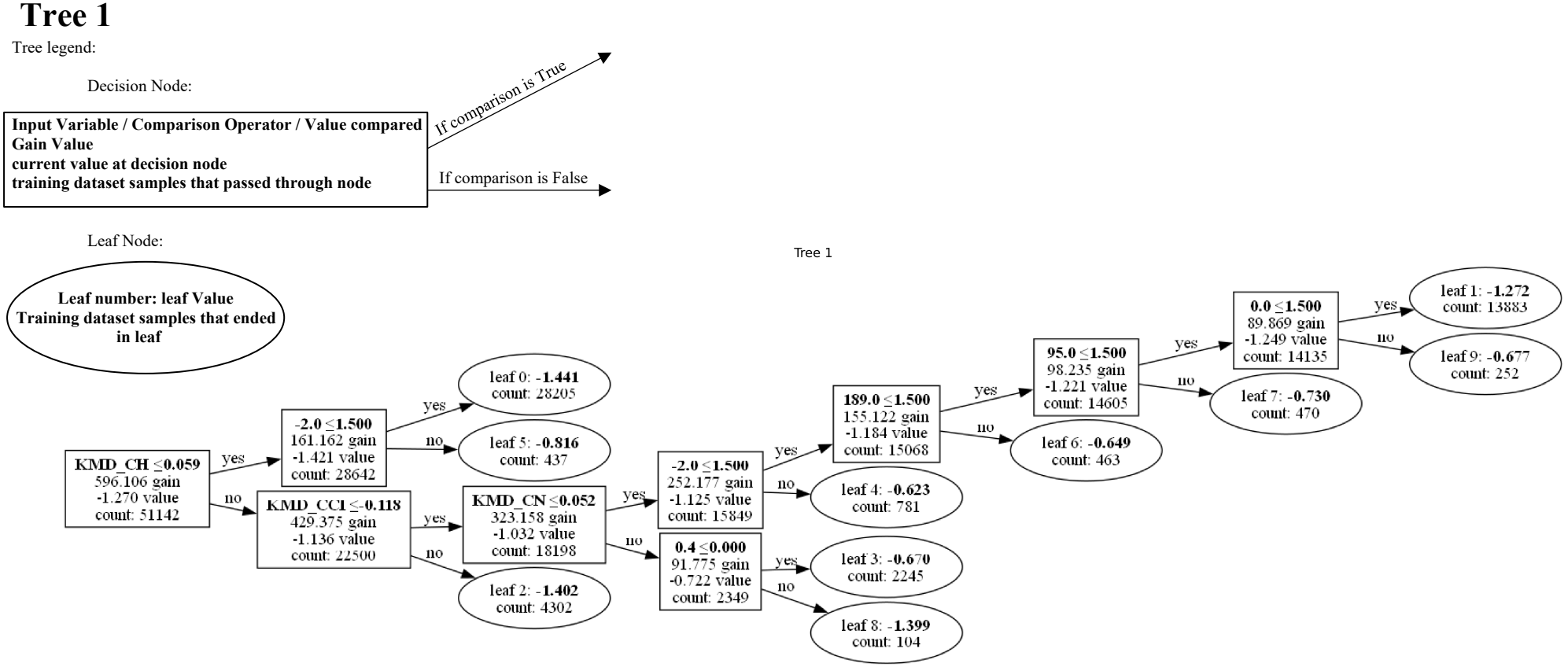


Figure S7: Top 15 important input variables by split of the model. The y scale of the split count is the percentage of each input split relative of the total split.

Figure S8



FigureS8 : 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

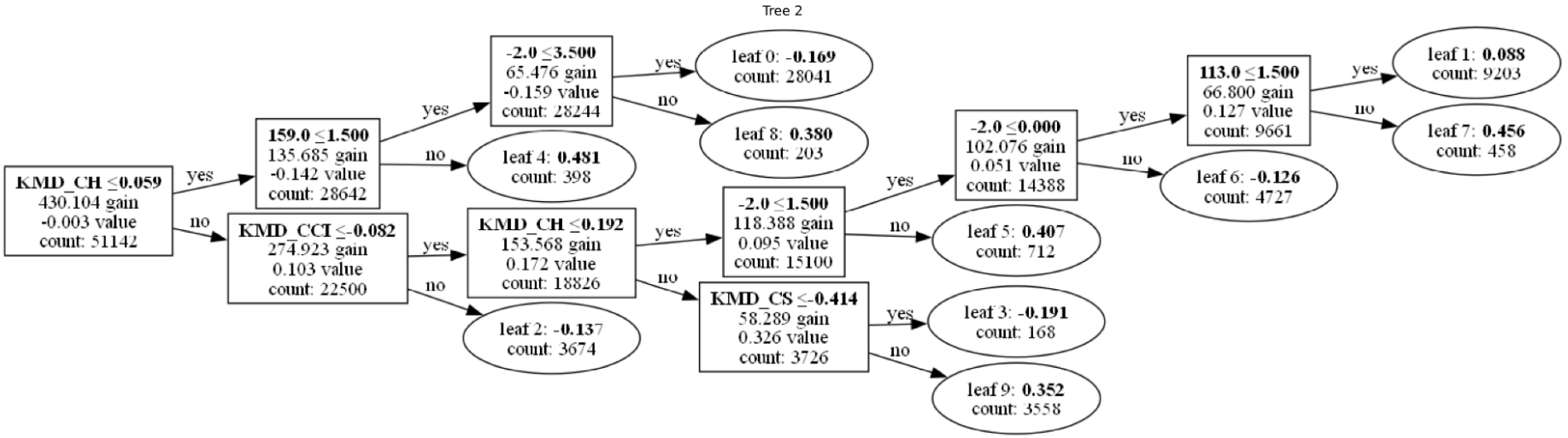


Figure S9 : second Tree, this tree compared to the first makes more discerning decisions on which spectrum is a PFAS and which isn't as it has negative values on its leaves.

Tree 3

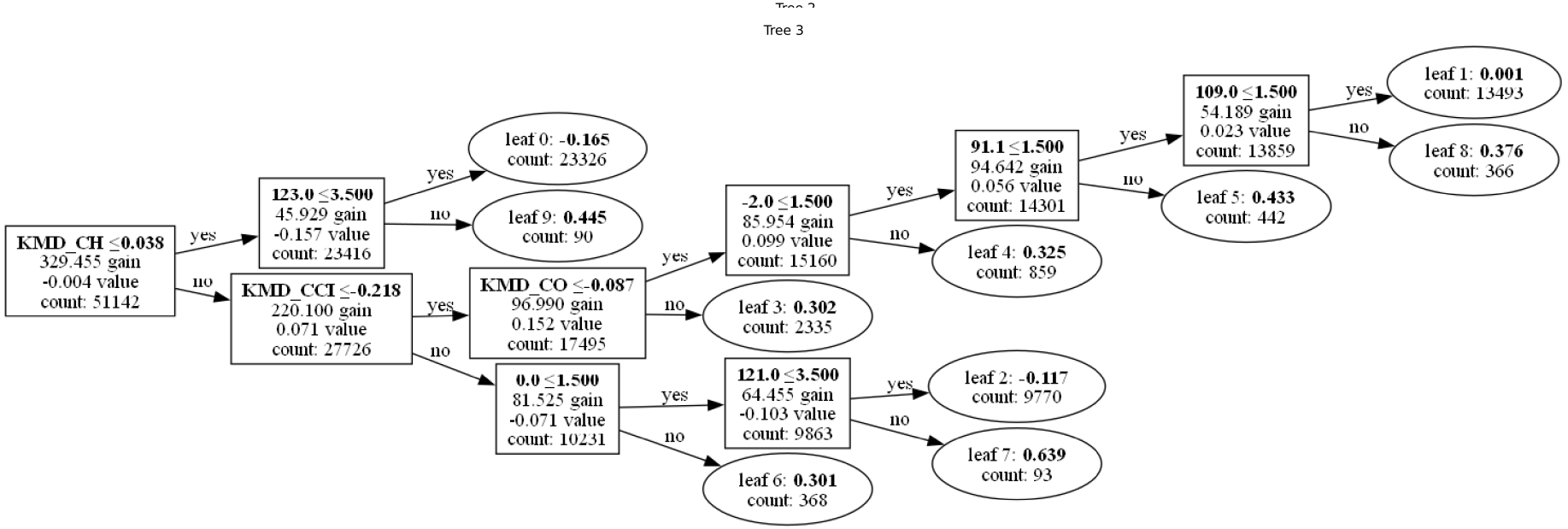


Figure S10 : Third Tree

Tree 4

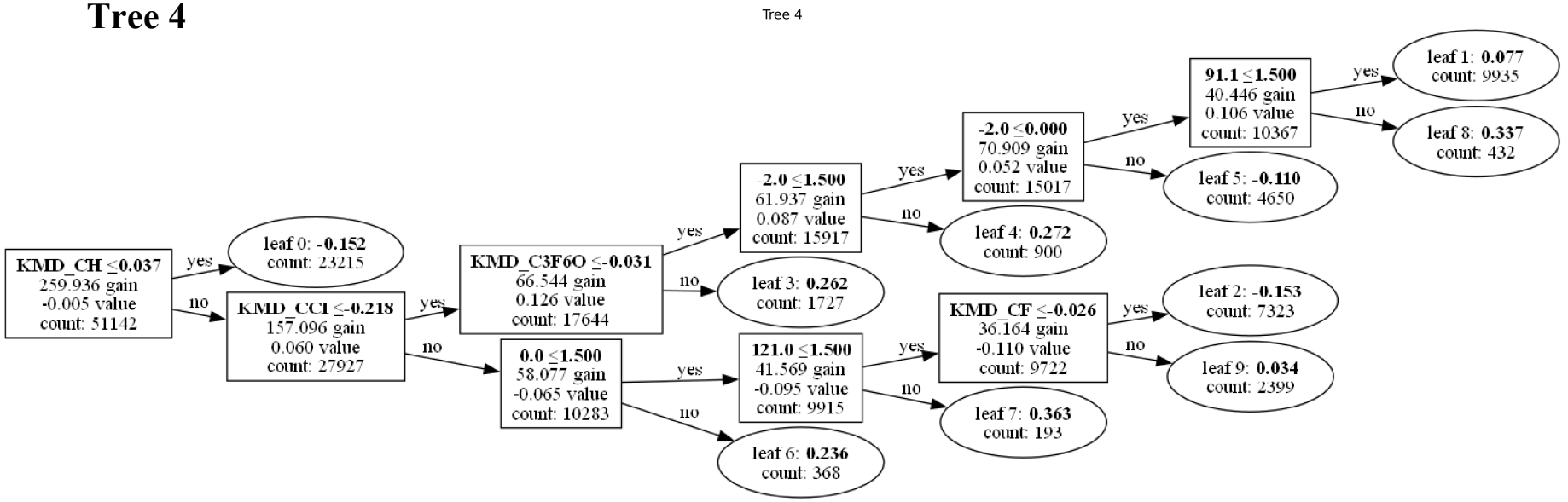


Figure S11 : Fourth tree

Tables

Table S1

Table S1 : External test set results of suspected PFAS spectra with identification confidence, RT, KMD, Library/Suspect acronyms, and fragments reported.

Blood Sample	Serum Sample	Feature m/z	MS2 fragments detected	Charbonnet Identification Level	Model Predicted probability
New structures					PFAS related
✓		247.9622	63.9623, 77.9654, 247.9612	**2a**	0.228
✓		297.9590	79.9563, 97.9592, 115.0744, 221.1902, 265.1596, 283.1917,	**2a**	0.658
Suspect match					PFAS related
✓	✓	469.9737	77.9653, 112.9861, 168.9899, 149.9850, 168.9889, 377.9565, 397.9523	**3b**	0.9467
✓	✓	460.9334	79.9575, 118.9920, 146.9863, 330.9769, 396.9667	**3a**	0.9835
✓	✓	414.9068	61.9885, 79.9571, 98.9550	**3a**	0.2851
✓	✓	414.9316	MS2 not clear, Mass	**5a**	0.9179
✓	✓	476.9283	79.9571, 98.9554, 310.9404, 360.9353	**3a**	0.9666
✓	✓	480.9393	79.9569, 98.9555, 118.9921, 180.9874	**3a**	0.981
✓	✓	510.9314	79.9581, 98.9571, 280.9561	**3a**	0.9736
✓		526.9616	80.9644, 346.91184, 486.9484, 506.9575	**3a**	0.9932
	✓	530.9353	MS2 data not clear	**3a**	0.9278
	✓	580.9325	79.9570, 98.9549, 168.9823	**3a**	0.9509
	✓	598.9235	Data not clear	**5a**	0.8636
✓	✓	612.9535	Data not clear	**5a**	0.8608
New structures					PFAS related
	✓	475.9299	77.9655, 92.9891, 156.9511, 413.2264, 475.9302	**3b**	0.9618
✓	✓	485.9686	77.9651, 118.9894, 168.9885, 397.9718, 423.9629, 485.9687	**3b**	0.7749
	✓	575.9223	77.9641, 92.9886, 156.9507,	**3b**	0.9155

Table S2

Table S2 : External test set results of potential PFAS compounds with low identification confidence. Identification confidence, RT, KMD, Library/Suspect acronyms, and fragments reported.

Blood Sample	Serum Sample	Feature m/z	MS2 fragments detected	Charbonnet Identification Level	Model Predicted probability
New structures					PFAS related
✓		204.9464	61.0078, 89.0026, 124.9800, 160.9568, 204.9460	4	0.0129
✓	✓	266.9644	79.9573, 80.9657, 107.0500, 187.0065	4	0.4267
	✓	296.9733	79.9563, 96.9592, 115.0744, 221.1902, 255.1596, 283.1917	4	0.0496
	✓	308.9437	79.9569, 80.9651, 228.9864	5a	0.3527
	✓	382.9413	82.9606, 118.9926, 168.9882, 318.9803	5a	0.7986
✓		177.0562	79.9573, 107.0499	5b	0.0075
✓		188.9862	79.9573, 91.0194, 108.0213, 109.0291, 188.9860	5b	0.0229
	✓	195.0689	79.9775, 96.9600	5b	0.0063
✓		201.0226	79.9576, 106.0427, 121.0659, 146.0615, 167.1092, 201.0228	5b	0.0122
	✓	204.0665	72.9930, 79.9575, 109.0248, 116.0505, 128.0500, 130.0661, 142.0657, 158.0610, 186.0560, 204.0655	5b	0.0041
✓	✓	213.0225	76.97, 89.0237, 133.0655, 199.1136	5b	0.0166
✓		215.0379	79.9576, 106.0426, 135.0813, 215.0385	5b	0.0117
	✓	219.9302	Fragments not clear	5b	0.0606
✓		220.9414	77.0030, 112.9801, 140.9742, 176.9514, 202.9310, 220.9420	5b	0.0318
✓		226.0179	79.9592, 131.0381, 146.0612, 226.0180	5b	0.023
✓		229.0536	78.9593, 112.9807, 149.0970, 209.0952	5b	0.4164
	✓	249.0227	Fragments not clear	5b	0.0577
	✓	249.0764	79.9572, 96.9601, 132.0040, 190.0100, 209.0854, 233.9983	5b	0.2724
	✓	256.9085	176.9512, 197.0640	5b	0.0873
	✓	260.9863	79.9576, 145.0298, 189.0923, 216.9981	5b	0.1391
✓		270.0806	79.9571, 106.9809, 124.0073, 162.0916, 270.0799	5b	0.0051
✓	✓	270.9737	79.9571, 135.0266, 159.0449, 191.0615, 270.9733	5b	0.0843
✓	✓	277.0213	79.9569, 106.0424, 133.0658, 185.9992, 213.0220, 261.9984	5b	0.0494
✓		281.0665	79.9579, 171.1020, 201.0231, 245.0451, 270.0809	5b	0.1892
✓		293.0491	79.9569, 106.0424, 185.9987, 213.0924	5b	0.3164
✓		309.0436	63.9623, 80.9654, 90.9859, 134.9752, 178.9653, 229.0875, 309.0441	5b	0.0266
	✓	379.0106	79.9560, 123.0450, 153.0184, 203.0012	5b	0.2011
✓		397.0037	79.9572, 107.0499, 187.0064	5b	0.242
	✓	422.9480	Fragments are not clear	5b	0.9803
	✓	435.0391	152.9953, 202.9992, 218.0216, 233.0457, 337.0710, 355.0843, 435.0403	5b	0.2111
✓		453.0662	79.9572, 135.0810, 215.0378, 453.0659	5b	0.3491
✓	✓	480.9399	79.9570, 80.9645, 98.9555, 118.9919, 460.9341	5b	0.9954
	✓	611.8569	MS2 data not clear	5b	0.0012
	✓	666.8565	MS2 data not clear	5b	0.0013

1. Olsen, J.V.; Godoy, L.M.F.d.; Li, G.; et al. Parts per Million Mass Accuracy on an Orbitrap Mass Spectrometer via Lock Mass Injection into a C-trap. *Molecular & Cellular Proteomics* **2005/12/01**, *4*, doi:10.1074/mcp.T500030-MCP200.
2. Kendrick, E. A Mass Scale Based on CH2 = 14.0000 for High Resolution Mass Spectrometry of Organic Compounds. *Analytical Chemistry* **May 1, 2002**, *35*, doi:10.1021/ac60206a048.
3. 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, doi:https://doi.org/10.1016/j.chemosphere.2022.137443.
4. DasGupta, A. Probability for Statistics and Machine Learning. *Springer Texts in Statistics* **2011**, doi:10.1007/978-1-4419-9634-3.
5. Jolliffe, I.T. *Principal Component Analysis*; Springer: New York, NY, 1986.
6. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **1996**, *58*, 267-288.
7. Moka, S.; Liquet, B.; Zhu, H.; et al. COMBSS: best subset selection via continuous optimization. *Statistics and Computing 2024 34:2* **2024**, *34*, doi:10.1007/s11222-024-10387-8.
8. Tianqi, C.; Carlos, G. XGBoost: A Scalable Tree Boosting System. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016.
9. Liudmila, P.; Gleb, G.; Aleksandr, V.; et al. CatBoost: unbiased boosting with categorical features. In Proceedings of the Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 2018.
10. Jia, S.; Marques Dos Santos, M.; Li, C.; et al. Recent advances in mass spectrometry analytical techniques for per- and polyfluoroalkyl substances (PFAS). *Analytical and Bioanalytical Chemistry* **2022**, *414*, 2795-2807, doi:10.1007/s00216-022-03905-y.
11. Allen, F.; Greiner, R.; Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **2015**, *11*, 98-110, doi:10.1007/s11306-014-0676-4.

---

i