Electronic Supplementary Material (ESI) for Environmental Science: Advances.

Supporting Information for:

Multi-Class Machine Learning Classification of PFAS in Environmental Water Samples: A Blinded Test of Performance on Unknowns

Tohren C. G. Kibbey, Denis M. O'Carroll, Andrew Safulko, and Greg Coyle

Supporting Information is split into the following seven files:

1. **SI - Training Dataset Sources.pdf**
   Detailed descriptions of all data sources in the training dataset, with links to original data.

2. **SI - List of components used as features.pdf**
   A list of the 30 PFAS components used as features for machine learning analyses

3. **SI - Full Table 3 - All data for Airport 2.pdf**
   A full version of Table 3 showing all classification results for Airport 2. The version in the paper only shows the first 45 samples for conciseness.

4. **SI - Test_Set_Unknowns_from_BC.csv**
   CSV file sent from BC to OU, UNSW researchers containing the unknowns for classification

5. **SI - RF Classification results sent to BC by OU and UNSW.csv**
   CSV file containing the classification results, as sent to BC by OU and UNSW researchers. This is the *original*, *unedited* file sent to BC. Only the filename has been changed for inclusion in the SI section. Note that internal terminology is slightly different in this file than in the paper, as some column headings were changed in the paper for clarity. Most notably, in this file, the top three classes are labeled S1, S2, S3 instead of C_1, C_2, C_3. There is a key in the spreadsheet indicating the meanings.

6. **SI - Key with site descriptions from BC (Plot ID added).xlsx**
   This is the site description information about the unknowns sent by BC to OU and UNSW researchers after classification had been completed. The file here is sorted in the same order as the unknowns (item 4 above). Note that Plot IDs were added to this file afterwards for this paper.

7. **SI - All plots from unknown samples.zip**
   Plots of unknown sample beta values shown with the closest match from each of the top three identified classes for each unknown. (Note: No plot is included for unknown T-46, because all components were below detection limits, so that sample contained no information to allow classification.)