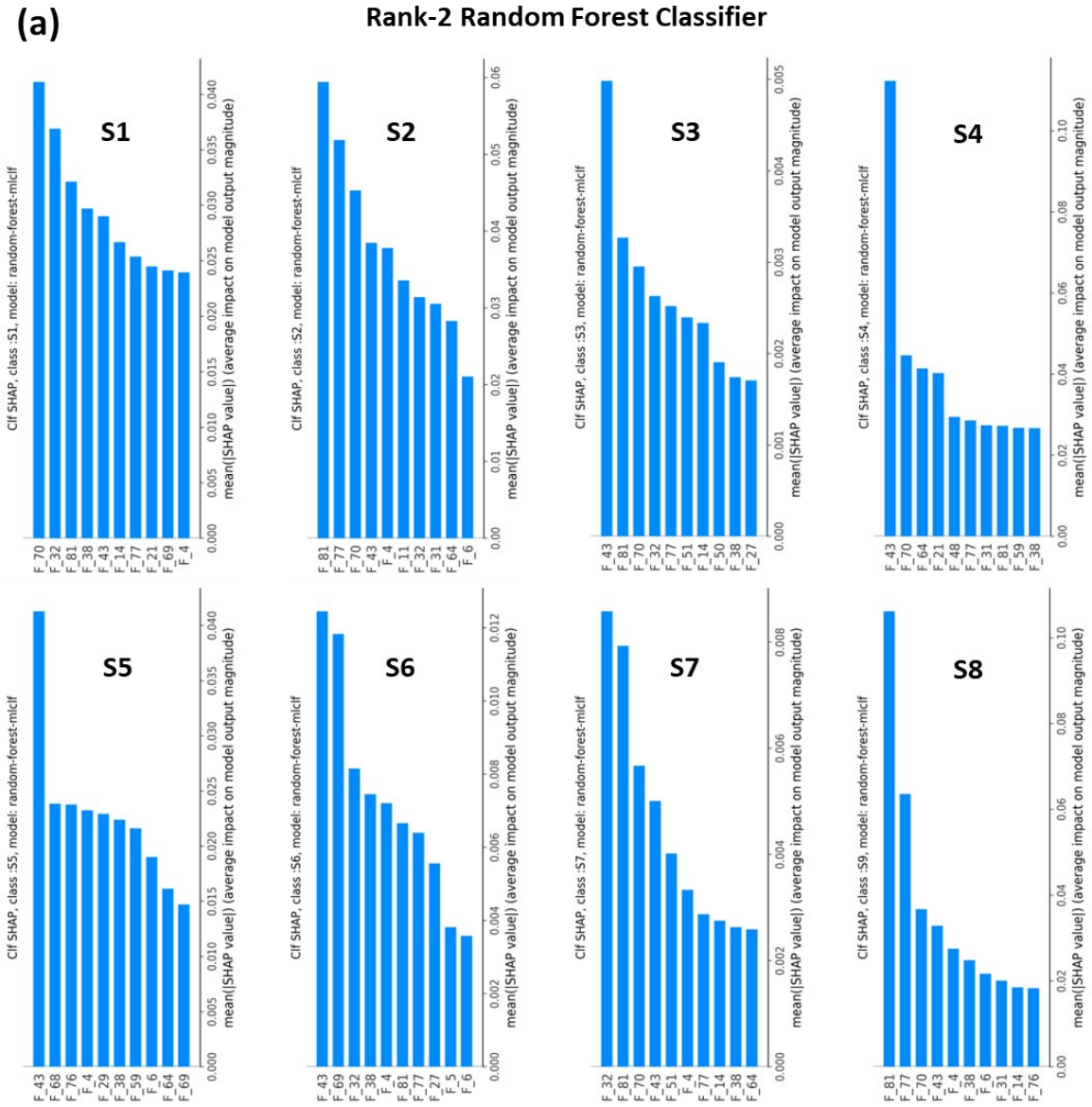
SUPPLEMENTARY INFORMATION

Stochastic Generalization Models Learn to Comprehensively Detect Volatile Organic Compounds Associated with Foodborne Pathogens via Raman Spectroscopy

Bohong Zhang[1,*], Anand K Nambisan[1], Abhishek Prakash Hungund[1], Xavier Jones[2], Qingbo Yang[2,*], and Jie Huang[1,*]

[1]*Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, Missouri, 65409, USA*

[2]*Cooperative Research, College of Agriculture, Environmental and Human Sciences, Lincoln University of Missouri, Jefferson City, MO 65101, USA*



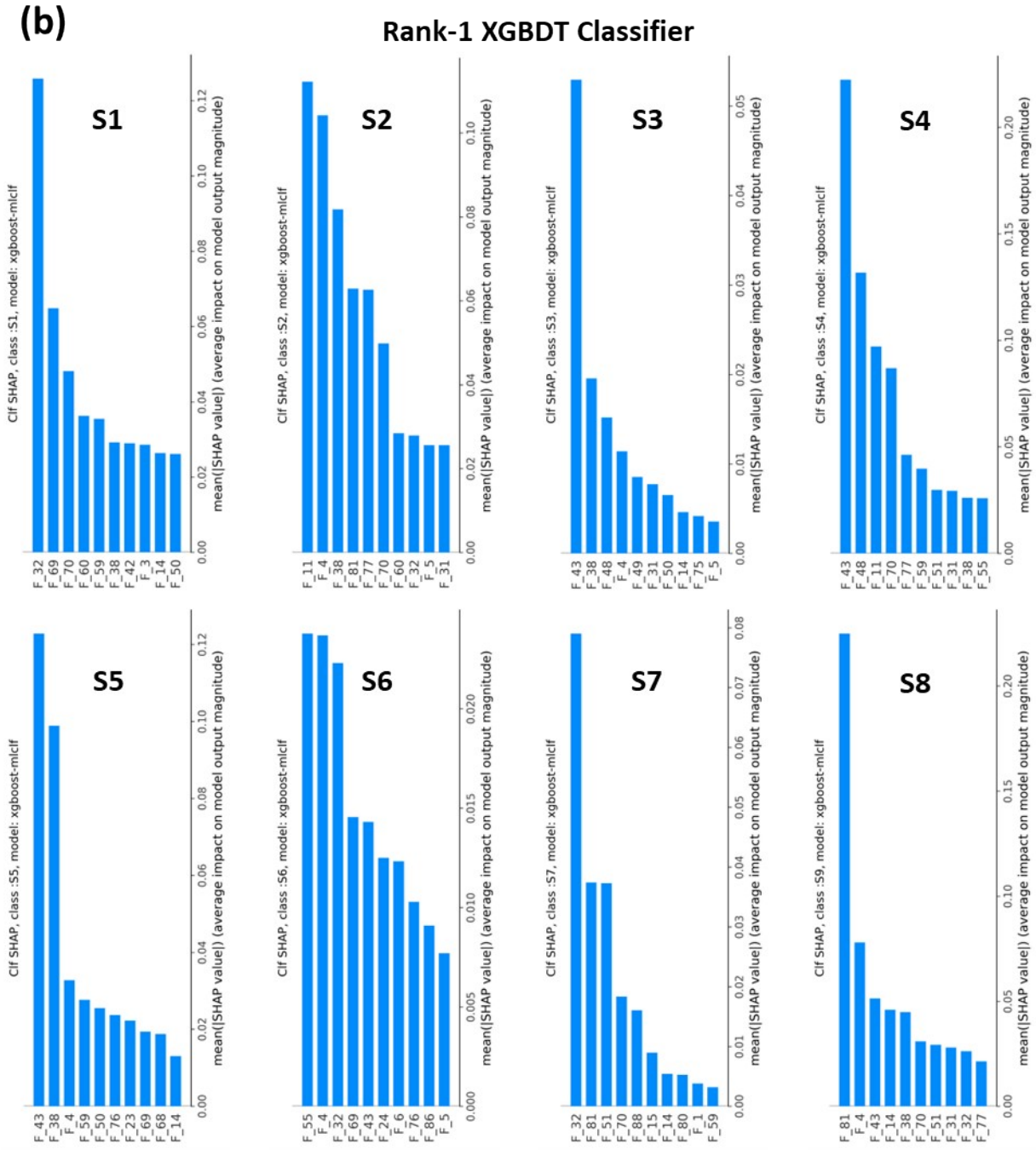


Figure S1. Feature importance plots for (a) Rank-2 Random Forest and (b) Rank-1 XGBDT.

In **Table S1**, a variety of features are used to capture different characteristics of the signal. Features in this context include statistical metrics (e.g., mean, standard deviation, skewness), spectral peaks (e.g., intensity, position, width) derived from the wavenumber (Raman shift) domain, and frequency-domain characteristics (e.g., Fourier transforms, power spectral density). Specifically, the signal's frequency-domain features capture the power and logarithmic magnitude across different frequency scales, providing a deeper understanding of the signal's characteristics. These include the maximum, total, average, variance, and peak of the power and scaled logarithmic

magnitude. The skewness and kurtosis of the power and scaled logarithmic magnitude in the frequency domain are also calculated. These features might be correlated, particularly those related to power and magnitude. Additional complex features such as spectral roll-off, signal envelope features, Mel-Frequency Cepstrum Coefficients (MFCCs), and wavelet-transform-based features are also extracted. Spectral roll-off represents the frequency below which a given percentage of the total spectral energy is contained, useful for distinguishing between harmonic and non-harmonic content. Signal envelope features can include extracting the envelope of the signal using techniques like the Hilbert transform. MFCCs are coefficients that represent audio and are widely used in speech and audio processing. Wavelet transforms can also be used to extract features. Histogram bins are calculated for different feature sets, providing another view of the distribution of these features. The count of the values of those features within the range defined by the bins could also show correlation depending on the distribution of your data. In summary, these features provide a comprehensive representation of the signal, capturing its various characteristics in both the wavenumber and frequency domains. However, potential correlations between these features should be considered before the model pipeline training.

Table S1. Feature alias/definition table. This table contains the list of aliases used to represent each of the features used for training the classifiers. The aliases are assigned to each of the feature for ease of representation and to optimize the feature importance plotting. This table contains all the 96 extracted features for the model pipeline training. The correlation between each of these features is computed and analyzed in detail before getting to the model pipeline training.

(a)

Feature	Definition	Alias
FREQ_SCALEDLOGMAG_TOP10_LOC_0	0th set of Top 10 peak locations of Scaled log -mag in frequency domain	F_0
PSD_23	23rd Power spectral density	F_1
PSD_87	87th Power spectral density	F_2
RS_IQR	Interquartile range of the signal in wavenumber (Raman shift) domain, a measure of statistical dispersion	F_3
WAVELET_LEVEL_2_MEAN	Mean of the level-2 wavelet transforms	F_4
PSD_20	20th Power spectral density	F_5
PSD_15	15th Power spectral density	F_6
PSD_70	70th Power spectral density	F_7
MFCC_9	9th Mel-Frequency Cepstrum Coefficients	F_8
PSD_64	64th Power spectral density	F_9
FREQ_SCALEDLOGMAG_BOT10_PKS_2	2nd set of Bottom 10 peak magnitudes of Scaled log -mag in frequency domain	F_10
PSD_13	13th Power spectral density	F_11
RS_CREST_FACTOR	Ratio of the peak value to the root mean square (RMS) value over a range wavenumbers	F_12
RS_AUTOCORRELATION	Autocorrelation of the signal, measuring the linear relationship between lagged values of the signal	F_13
RS_MEDIAN	Median value of the wavenumber (Raman shift) domain signal	F_14
FREQ_HIST_MAG_3	Magnitude of Histogram in 3rd bin of frequency domain signal	F_15
SPECTRAL_FLATNESS	Measure of the "flatness" or "tonality" of spectrum of the signal. A higher value indicates a more noise -like signal.	F_16
PSD_25	25th Power spectral density	F_17
PSD_65	65th Power spectral density	F_18
PSD_47	47th Power spectral density	F_19
MFCC_11	11th Mel-Frequency Cepstrum Coefficients	F_20
WAVELET_LEVEL_3_STD	Standard deviation of the level -3 wavelet transforms	F_21
MFCC_17	17th Mel-Frequency Cepstrum Coefficients	F_22
WN_TOP5_PKS_3	3rd set of wavenumbers associated with top 5 peaks	F_23
PSD_32	32nd Power spectral density	F_24
PSD_17	17th Power spectral density	F_25
MFCC_18	18th Mel-Frequency Cepstrum Coefficients	F_26
WAVELET_LEVEL_5_STD	Standard deviation of the level -5 wavelet transforms	F_27
FREQ_SCALEDLOGMAG_BOT10_PKS_3	3rd set of Bottom 10 peak magnitudes of Scaled log -mag in frequency domain	F_28
FREQ_MEAN_SCALEDLOGMAG	Average scaled logarithmic magnitude in the frequency domain	F_29
PSD_76	76th Power spectral density	F_30
PSD_7	7th Power spectral density	F_31
WAVELET_LEVEL_1_MEAN	Mean of the level-1 wavelet transforms	F_32
FREQ_SCALEDLOGMAG_TOP10_LOC_7	7th set of Top 10 peak locations of Scaled log -mag in frequency domain	F_33
PSD_69	69th Power spectral density	F_34
HIST_2	2nd bin of Histogram	F_35
WN_TOP5_LOC_3	3rd set of wavenumber indices associated with top 5 peaks	F_36
FREQ_HIST_MAG_1	Magnitude of Histogram in 1st bin of frequency domain signal	F_37
WAVELET_LEVEL_3_MEAN	Mean of the level-3 wavelet transforms	F_38
WN_TOP5_LOC_1	1st set of wavenumber indices associated with top 5 peaks	F_39
FREQ_SCALEDLOGMAG_TOP10_LOC_3	3rd set of Top 10 peak locations of Scaled log -mag in frequency domain	F_40
MFCC_5	5th Mel-Frequency Cepstrum Coefficients	F_41
PSD_66	66th Power spectral density	F_42
PSD_102	102nd Power spectral density	F_43
PSD_34	34th Power spectral density	F_44
PSD_28	28th Power spectral density	F_45
PSD_72	72nd Power spectral density	F_46

(b)

MFCC_13	13th Mel-Frequency Cepstrum Coefficients	F_47
PSD_96	96th Power spectral density	F_48
WN_TOP5_PKS_2	2nd set of wavenumbers associated with top 5 peaks	F_49
RS_ENTROPY	Entropy of the signal's probability distribution, representing the randomness or unpredictability.	F_50
PSD_101	101st Power spectral density	F_51
HIST_3	3rd bin of Histogram	F_52
FREQ_SCALEDLOGMAG_TOP10_PKS_4	4th set of Top 10 peak magnitudes of Scaled log -mag in frequency domain	F_53
FREQ_SCALEDLOGMAG_BOT10_LOC_5	5th set of Bottom 10 peak locations of Scaled log -mag in frequency domain	F_54
PSD_81	81st Power spectral density	F_55
FREQ_SCALEDLOGMAG_TOP10_PKS_5	5th set of Top 10 peak magnitudes of Scaled log -mag in frequency domain	F_56
MFCC_10	10th Mel-Frequency Cepstrum Coefficients	F_57
WAVELET_LEVEL_5_MEAN	Mean of the level-5 wavelet transforms	F_58
PSD_14	14th Power spectral density	F_59
PSD_26	26th Power spectral density	F_60
PSD_31	31st Power spectral density	F_61
HIST_1	1st bin of Histogram	F_62
PSD_42	42nd Power spectral density	F_63
WN_TOP5_PKS_1	1st set of wavenumbers associated with top 5 peaks	F_64
WN_TOP5_LOC_4	4th set of wavenumber indices associated with top 5 peaks	F_65
PSD_75	75th Power spectral density	F_66
MFCC_1	1st Mel-Frequency Cepstrum Coefficients	F_67
FREQ_VAR_SCALEDLOGMAG	variance of the scaled logarithmic magnitude in the frequency domain	F_68
WAVELET_LEVEL_4_MEAN	Mean of the level-4 wavelet transforms	F_69
PSD_97	97th Power spectral density	F_70
MFCC_6	6th Mel-Frequency Cepstrum Coefficients	F_71
MFCC_12	12th Mel-Frequency Cepstrum Coefficients	F_72
WN_TOP5_PKS_0	0th set of wavenumbers associated with top 5 peaks	F_73
PSD_27	27th Power spectral density	F_74
MFCC_15	15th Mel-Frequency Cepstrum Coefficients	F_75
FREQ_SKEW_POWER	skewness of the power in the frequency domain	F_76
PSD_100	100th Power spectral density	F_77
PSD_68	68th Power spectral density	F_78
PSD_61	61st Power spectral density	F_79
FREQ_HIST_MAG_2	Magnitude of Histogram in 2nd bin of frequency domain signal	F_80
PSD_99	99th Power spectral density	F_81
WN_TOP5_LOC_2	2nd set of wavenumber indices associated with top 5 peaks	F_82
PSD_62	62nd Power spectral density	F_83
MFCC_19	19th Mel-Frequency Cepstrum Coefficients	F_84
MFCC_8	8th Mel-Frequency Cepstrum Coefficients	F_85
MFCC_14	14th Mel-Frequency Cepstrum Coefficients	F_86
MFCC_16	16th Mel-Frequency Cepstrum Coefficients	F_87
SPECTRAL_ROLLOFF	frequency below which a given percentage (e.g., 85%) of the total spectral energy is contained	F_88
WN_TOP5_LOC_0	0th set of wavenumber indices associated with top 5 peaks	F_89
MFCC_7	7th Mel-Frequency Cepstrum Coefficients	F_90
MFCC_2	2nd Mel-Frequency Cepstrum Coefficients	F_91
MFCC_0	0th Mel-Frequency Cepstrum Coefficients	F_92
MFCC_3	3rd Mel-Frequency Cepstrum Coefficients	F_93

For all the features given in the **Figure S1 and S2**, the correlations are calculated w.r.t. one another and represented in a matrix form. The features showing lower correlations are represented by white or lighter shade of red and the features showing higher correlations are represented as darker shade of red. The purpose of this calculation is to determine the features having a high level of correlation. Those features can affect the model pipeline performance due to the multi-collinearity causing inaccuracies. Such highly correlated features are dropped during the preprocessing before the pipeline training step. The threshold to drop features based on correlation analysis is set to 0.95.

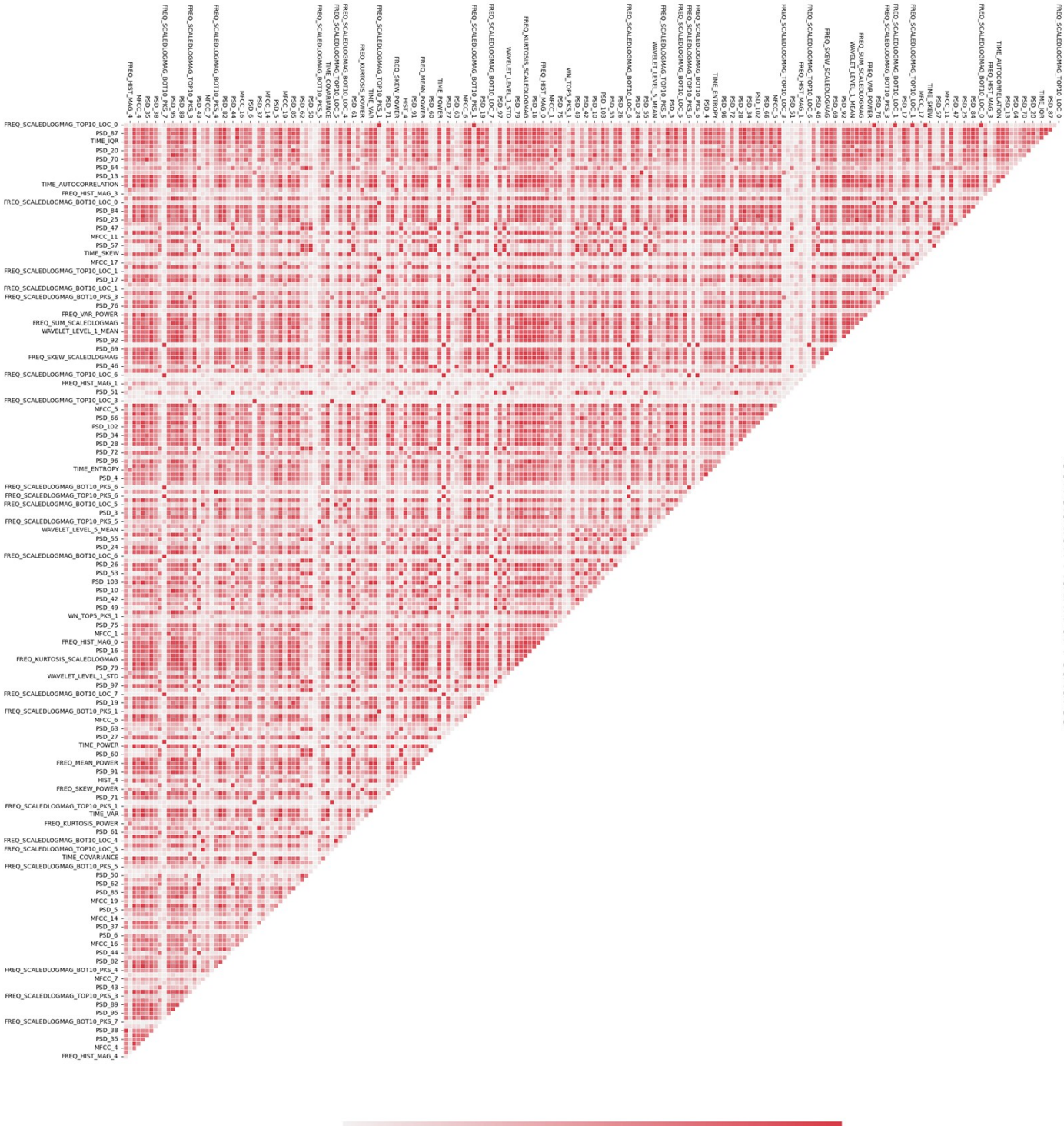


Figure S2. Feature correlation matrix.

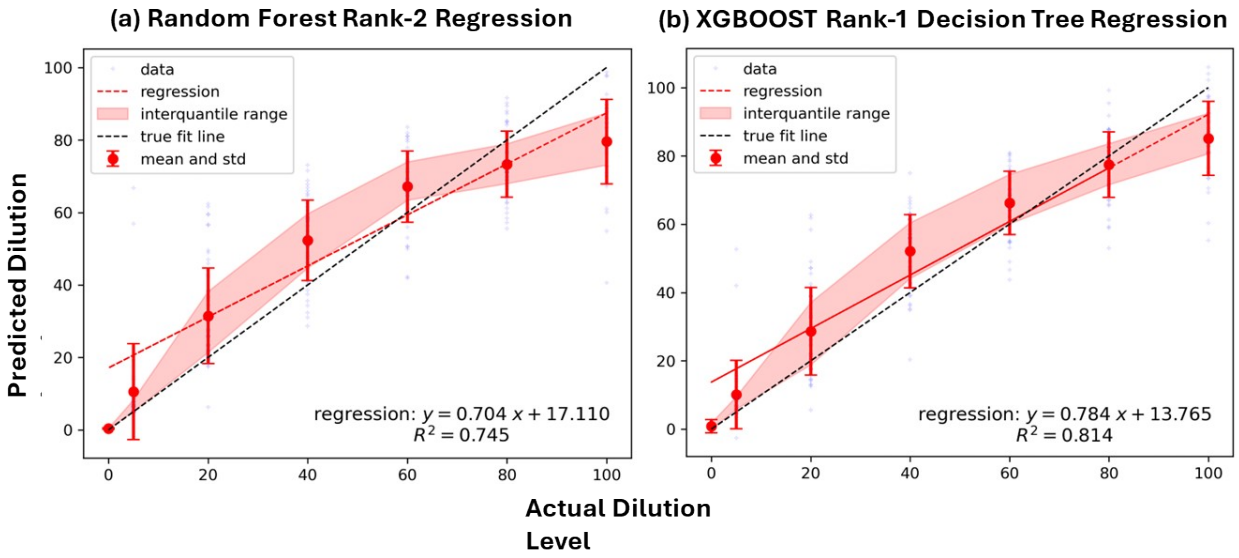
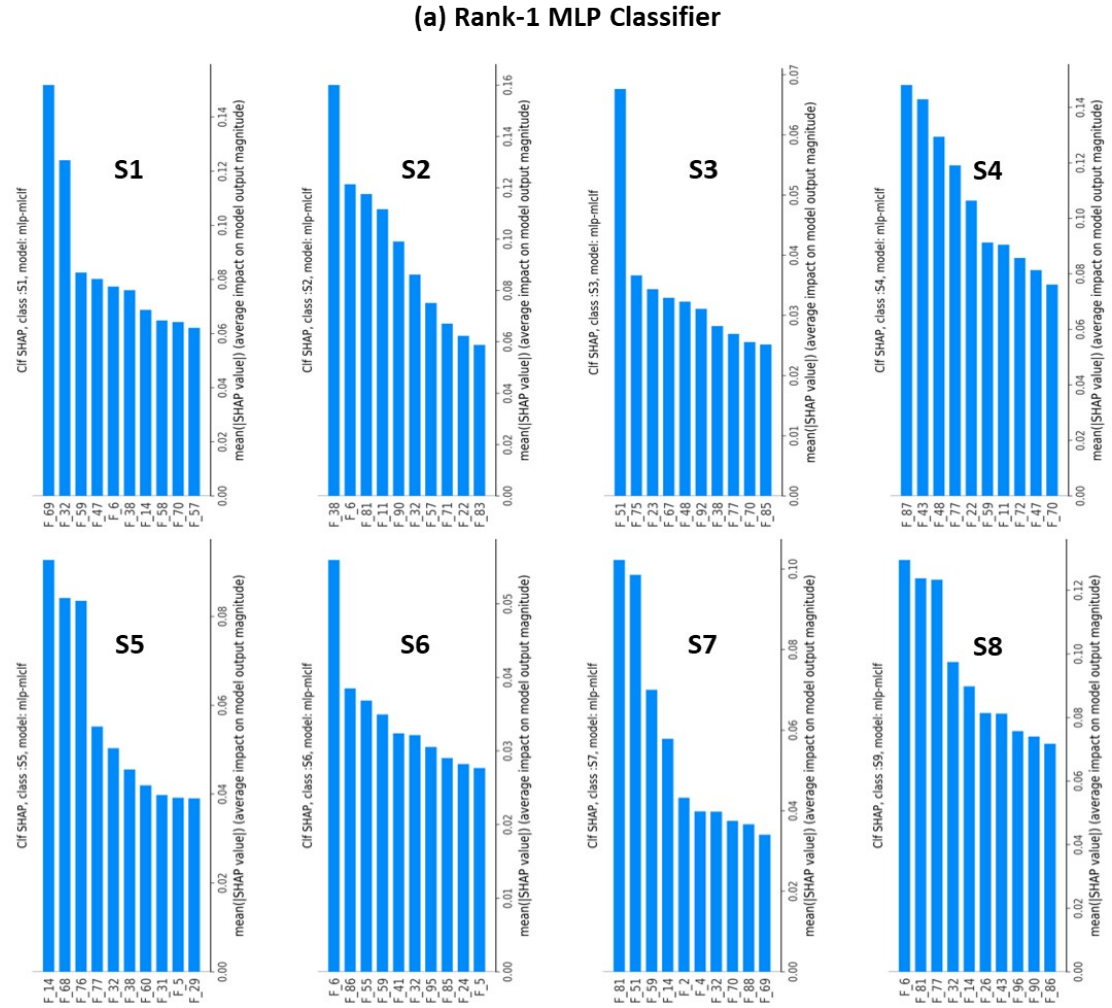
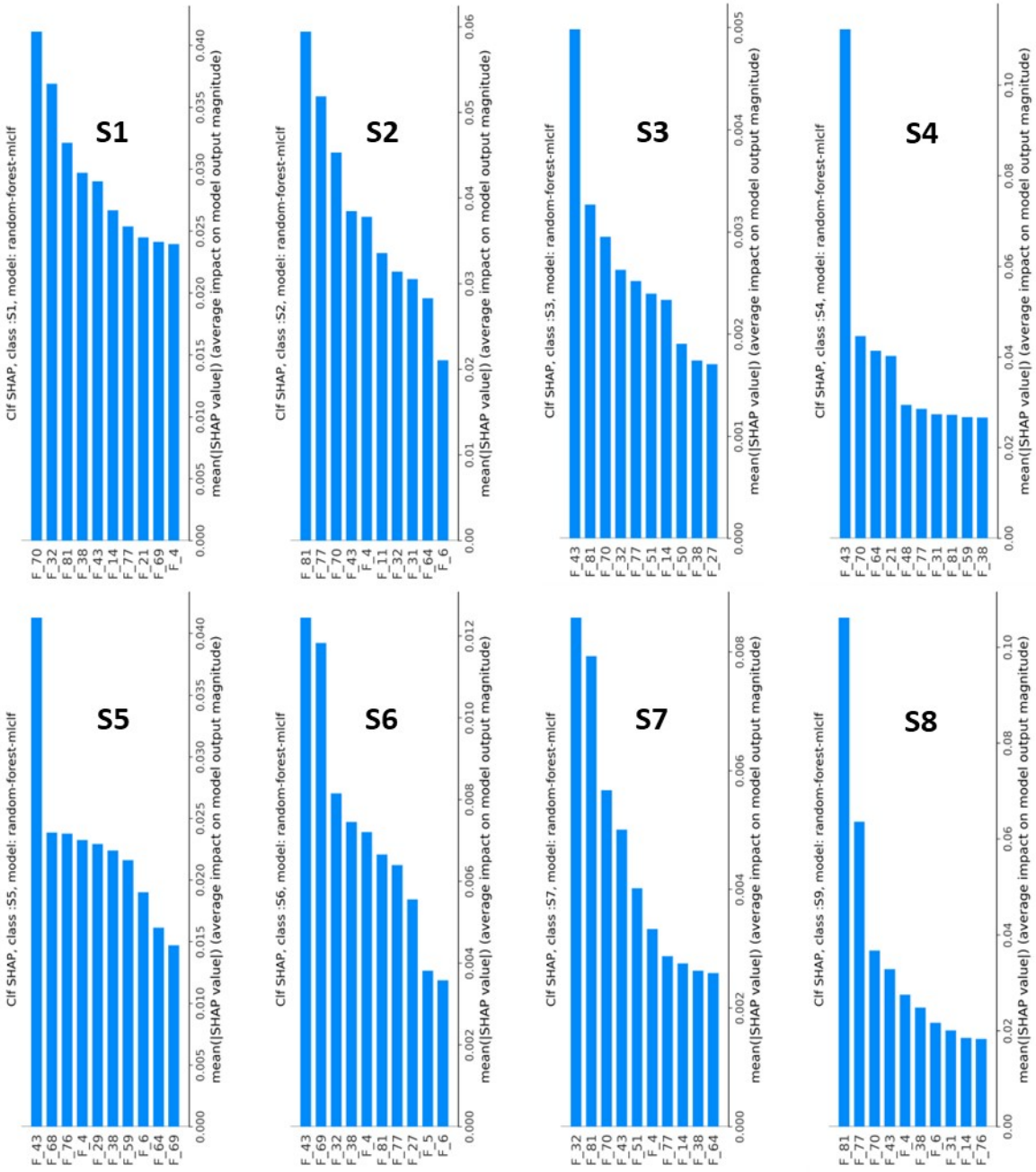


Figure S3. Regression to predict the dilution levels of all the VOCs. VOC dilution level predictions by (a) Random Forest rank-2 regression and (b) XGBDT rank-1 decision tree regression.



(b) Rank-3 Random Forest Classifier



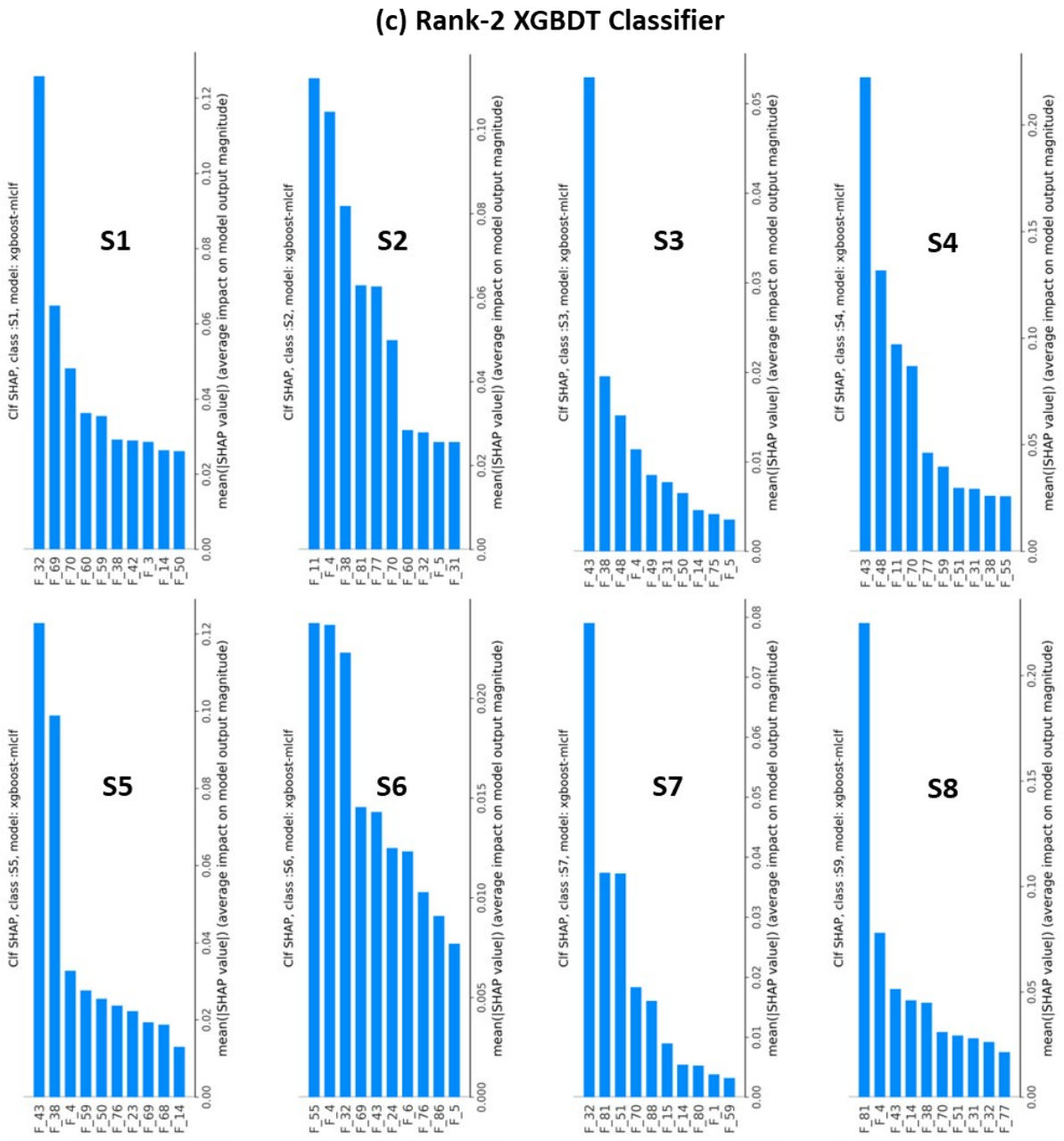


Figure S4. Feature importance plots for 3 prominent pipeline classifiers. FI plot for (a) rank-1 MLP classifier, (b) rank-3 random forest classifier, and (c) rank-2 XGBDT classifier. The plots are tabulated according to the pure VOCs' class labels S1 to S8. When these VOCs are mixed to obtain a mixture VOC, the feature importance (or mean of SHAP values) of that mixture will be a derivative of the mean SHAP value of the pure VOCs used to obtain the mixture. The feature aliases in the x-axis of the bar plots are defined in the Table. S1 in the Supplementary Information. These plots give detailed insight into the impact of certain features on the model pipeline performance and are plotted right to left with the most important feature (i.e., the feature with largest mean SHAP value) starting at the right. Higher mean SHAP values suggest a greater impact on the model's output. Features with larger bars are more influential in the model's predictions.

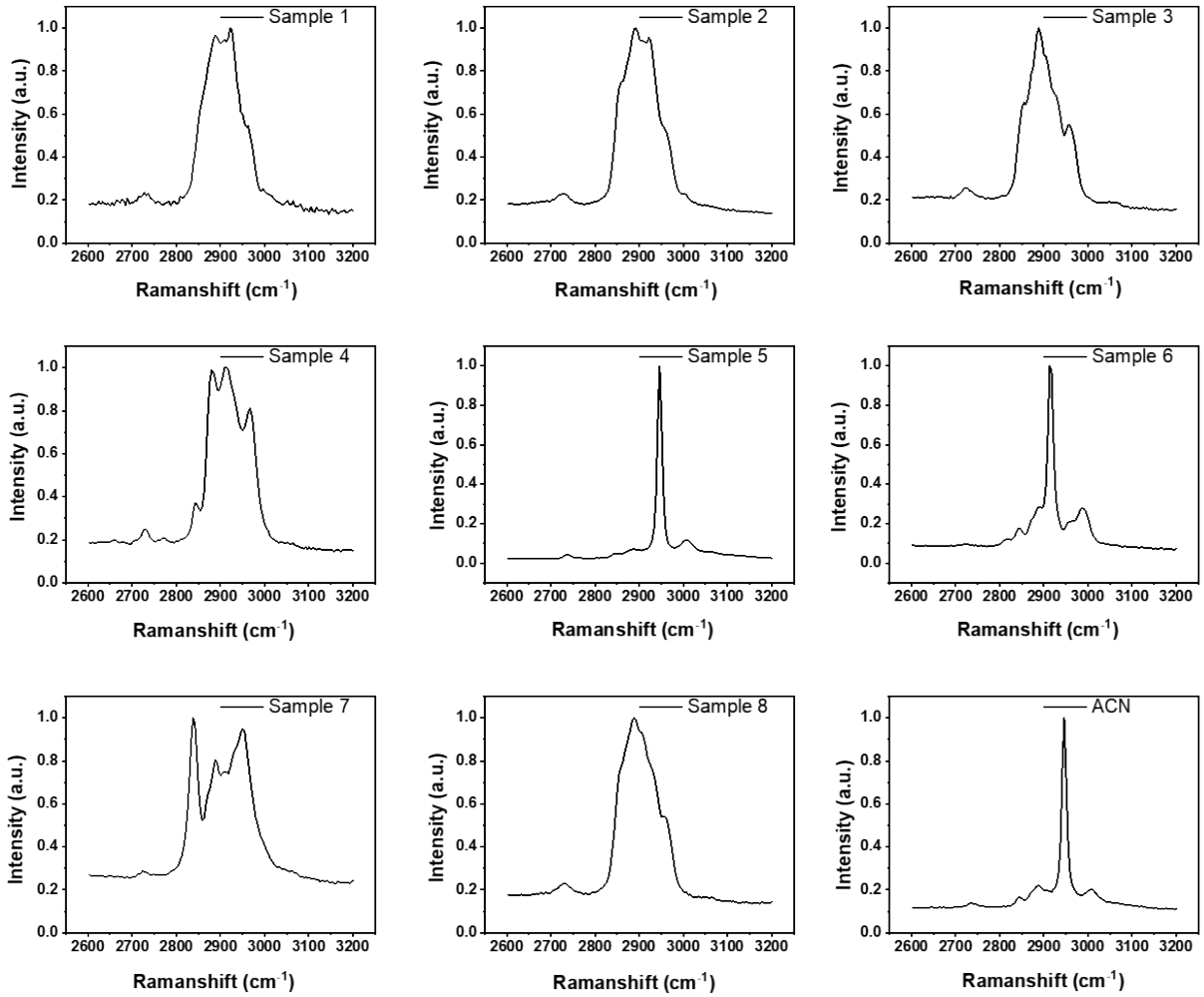


Figure S5. Raman spectra of eight pure VOCs and ACN.

Table S2. Concentration Levels of VOC Mixtures at Various Dilution Factors Expressed in Units per Milliliter (unit/mL), Percentage (%), and Parts Per Million (PPM).

Dilution Level (X)	Concentration (unit/mL)	Concentration (%)	Concentration (PPM)
0X (Pure)	1	100.00%	1,000,000
5X	0.2	20.00%	200,000
10X	0.1	10.00%	100,000
20X	0.05	5.00%	50,000
40X	0.025	2.50%	25,000
60X	0.0167	1.67%	16,667
80X	0.0125	1.25%	12,500
100X	0.01	1.00%	10,000
160X	0.0063	0.63%	6,250
200X	0.005	0.50%	5,000
240X	0.0042	0.42%	4,167
320X	0.0031	0.31%	3,125
400X	0.0025	0.25%	2,500