Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 196

Novel Application of Machine Learning Techniques for Rapid Online Source Apportionment of Aerosol Mass Spectrometer Datasets

MANISH SHRIVASTAVA

PARITOSH PANDE, JOHN E. SHILLING, ALLA ZELENYUK, QI ZHANG, QI CHEN, NGA LEE NG, YUE ZHANG, MASAYUKI TAKEYUCHI, THEODORA NAH, QUAZI Z. RASOOL, YUWEI ZHANG, BIN ZHAO, YING LIU

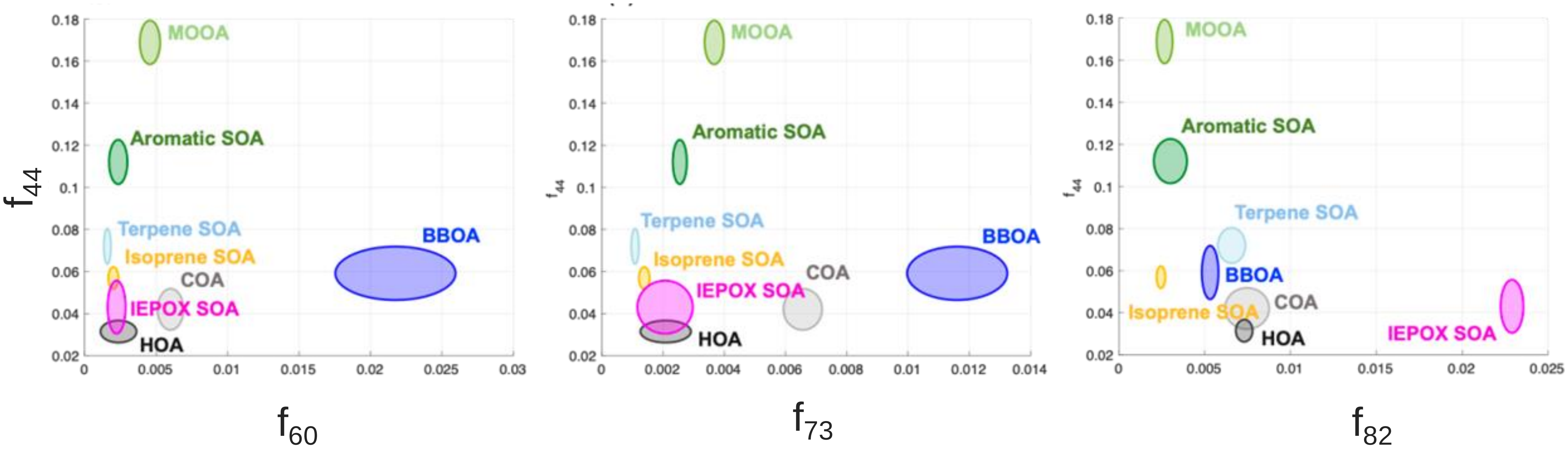
ASR
Atmospheric
System Research

Acknowledgements: Supported by the U.S. Department of Energy's Office of Science Biological and Environmental Research Early Career Award and ASR

Motivation

- Long-term measurements of aerosol mass spectrometer data require fast and online source-apportionment of organic aerosols
- PMF assumes a global profile fit for each source that does not vary in time
- PMF is very time consuming, needs the whole time series a-priori and involves substantial user judgement
- We train a machine learning (ML) algorithm to rapidly identify source profiles
- ML training uses knowledge from laboratory and field measurements of OA profiles
- ML can be applied to single OA spectrum for online source apportionment
- It can be applied to rapidly identify changing emissions, source profiles

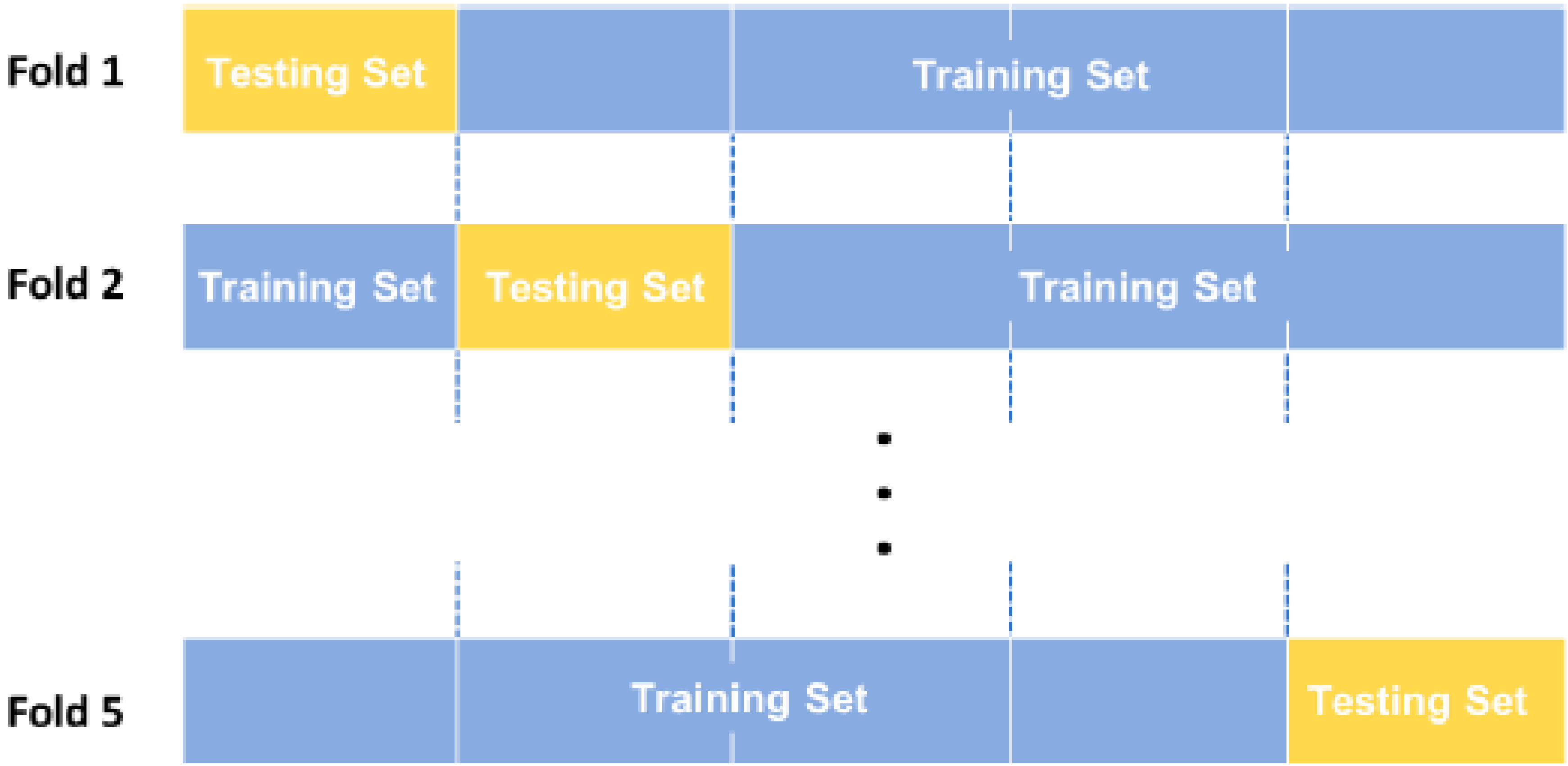
AMS data: Key mass to charge ratio markers can be used to identify different SOA sources



Pande, Shrivastava et al. 2022

- Trained a logistic regression classifier to identify 4 well characterized laboratory spectra (isoprene SOA, IEPOX-SOA, aromatic SOA and monoterpene SOA), and PMF derived factor spectra related to HOA, COA, BBOA and MO-OOA

5-fold cross validation used to test generalizability of model to identify SOA mass spectra from ground truth lab & field data



Fold 1 Testing Set Training Set

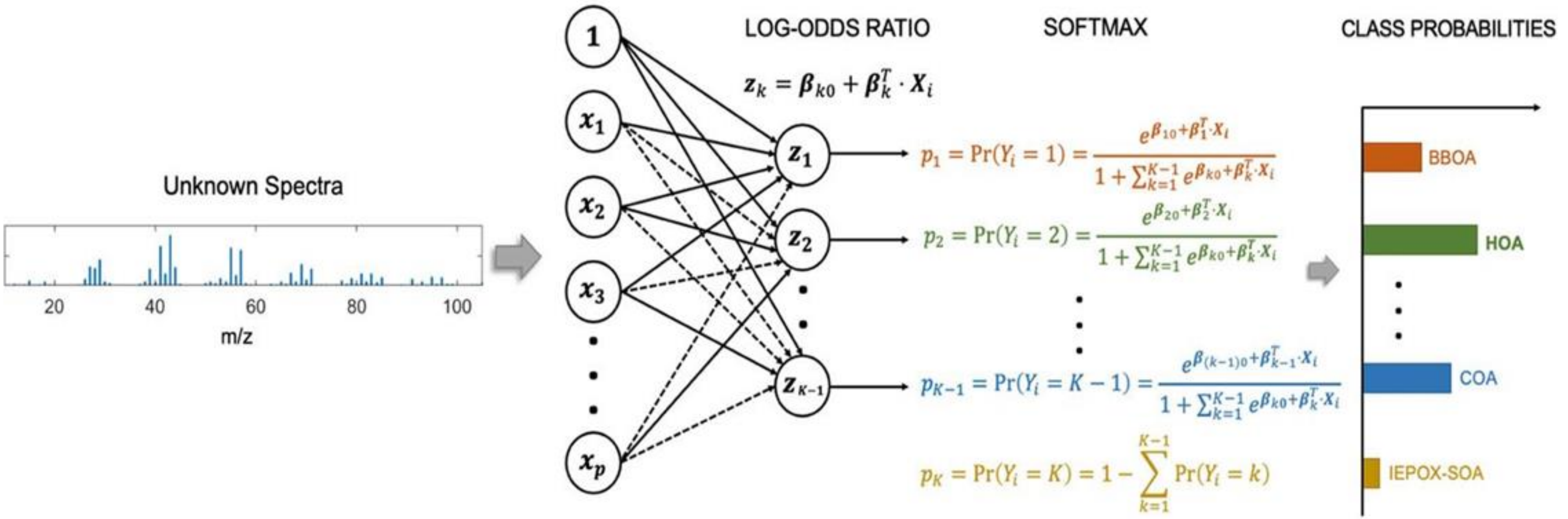
Fold 2 Training Set Testing Set Training Set

Fold 5 Training Set Testing Set

Pande, Shrivastava et al. 2022

- When training data are sparse, a 5-fold cross validation repeated 10 times helped increase the ability to train the classifier on well characterized laboratory AMS data with high signal to noise ratio

Multinomial logistic regression classifier assigns probabilities that a given mass spectra belongs to different sources



Unknown Spectra
20
40
60
80
100
m/z
1
x_1
x_2
x_3
x_p
z_1
z_2
z_{K-1}
LOG-ODDS RATIO
$z_k = \beta_{k0} + \beta_k^T \cdot X_i$
SOFTMAX
$p_1 = \Pr(Y_i = 1) = \frac{e^{\beta_{10}+\beta_1^T \cdot X_i}}{1+\sum_{k=1}^{K-1} e^{\beta_{k0}+\beta_k^T \cdot X_i}}$
$p_2 = \Pr(Y_i = 2) = \frac{e^{\beta_{20}+\beta_2^T \cdot X_i}}{1+\sum_{k=1}^{K-1} e^{\beta_{k0}+\beta_k^T \cdot X_i}}$
$p_{K-1} = \Pr(Y_i = K-1) = \frac{e^{\beta_{(k-1)0}+\beta_{k-1}^T \cdot X_i}}{1+\sum_{k=1}^{K-1} e^{\beta_{k0}+\beta_k^T \cdot X_i}}$
$p_K = \Pr(Y_i = K) = 1 - \sum_{k=1}^{K-1} \Pr(Y_i = k)$
CLASS PROBABILITIES
BBOA
HOA
COA
IEPOX-SOA

Pande, Shrivastava et al. 2022

- Used ground truth mass spectra from laboratory and field measurements to train the classifier in identifying probabilities for assigning mass spectra to sources

Classification probabilities are used to predict mass fractional abundances: Non-linear boosted regression

STEP 1

Generate N Mixed Sample Spectra M using averaged spectra of the K constituent OA species $S_j, j = 1, 2, \cdots, K$ obtained from source spectra samples

$$M \equiv \left\{\sum_{j=1}^{K} \alpha_j^i \cdot S_j\right\}_{i=1}^{i=N}$$ such that: $\{(\alpha_1^i, \alpha_2^i, \cdots, \alpha_K^i)\}_{i=1}^{i=N} \sim$ Dirichlet distribution, *i.e.* $\sum_{j=1}^{K} \alpha_j^i = 1$

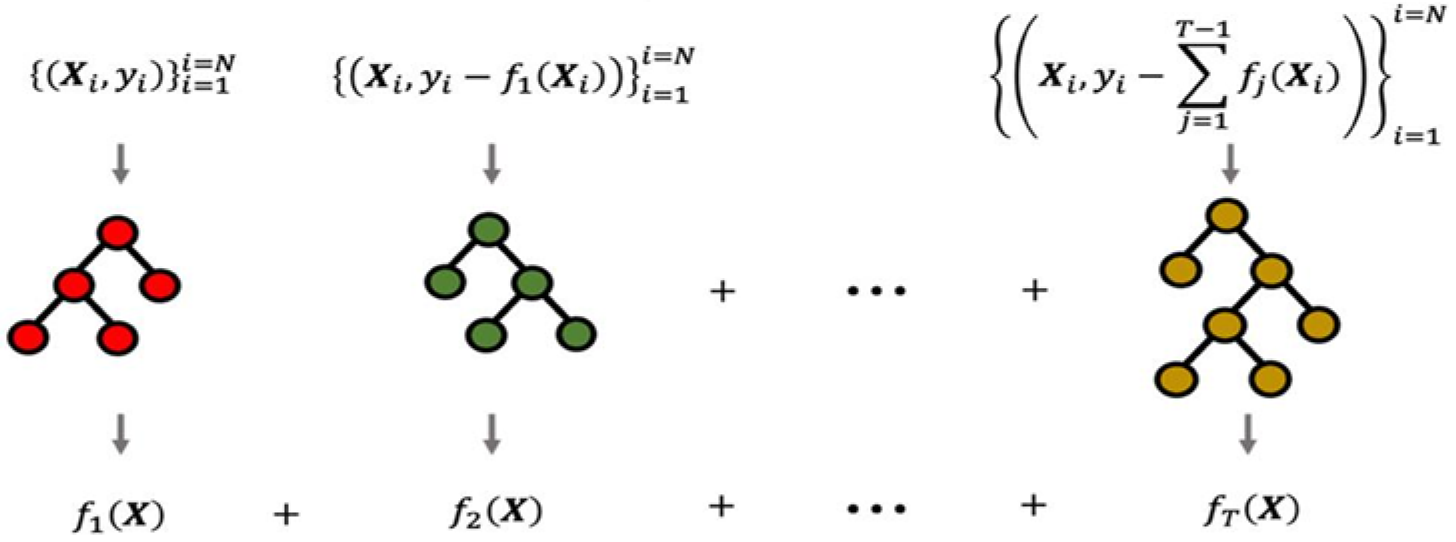
STEP 2

Predict class probabilities $\wp$ using the multinomial logistic classifier trained on source OA spectra

$$\wp \equiv \{(p_1^i, p_2^i, \cdots, p_K^i)\}_{i=1}^{i=N}$$

STEP 3

Train K boosted regression models to predict fractional contributions $y_i = p_k^i, k = 1,2, \cdots, K$ of each constituent OA species from predicted probabilities $\boldsymbol{X}_i = [\alpha_1^i, \alpha_2^i, \cdots, \alpha_K^i]$



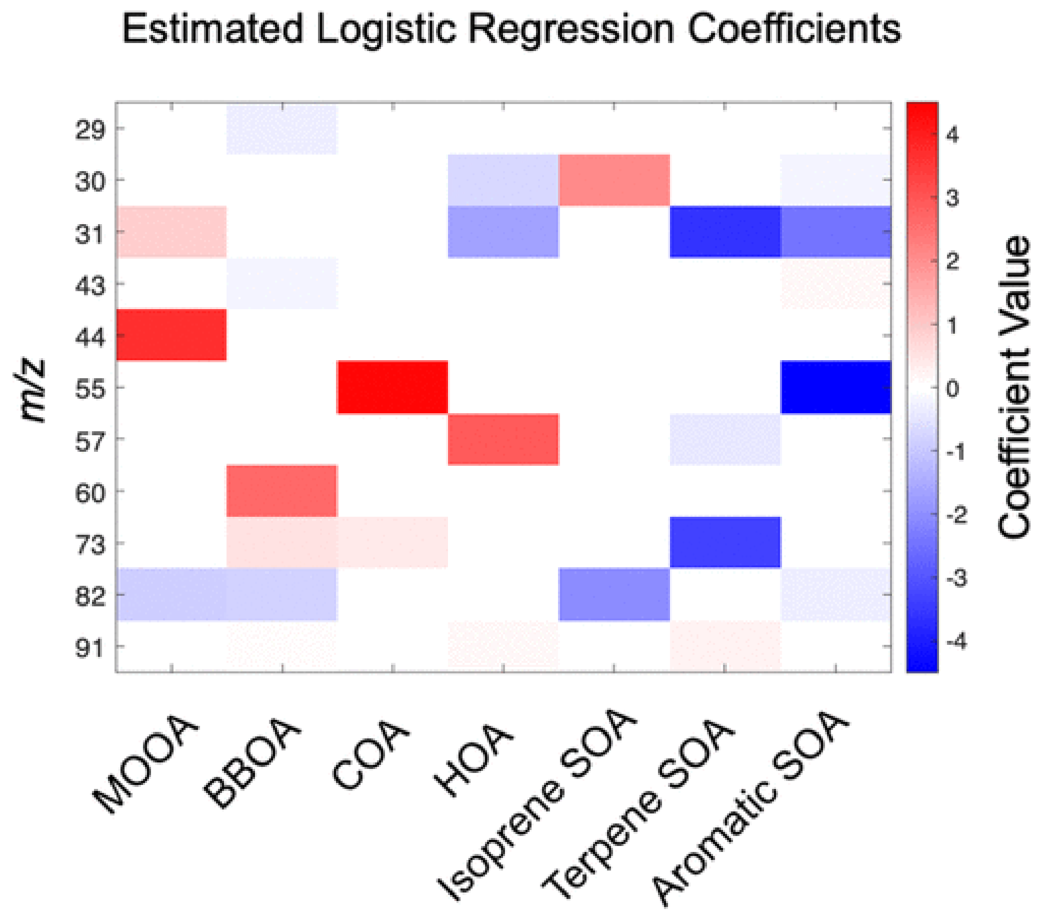
$\{(\boldsymbol{X}_i, y_i)\}_{i=1}^{i=N}$

$\{(\boldsymbol{X}_i, y_i - f_1(\boldsymbol{X}_i))\}_{i=1}^{i=N}$

$\left\{\left(\boldsymbol{X}_i, y_i - \sum_{j=1}^{T-1} f_j(\boldsymbol{X}_i)\right)\right\}_{i=1}^{i=N}$

$$f(\boldsymbol{X}) = f_1(\boldsymbol{X}) + f_2(\boldsymbol{X}) + \cdots + f_T(\boldsymbol{X})$$

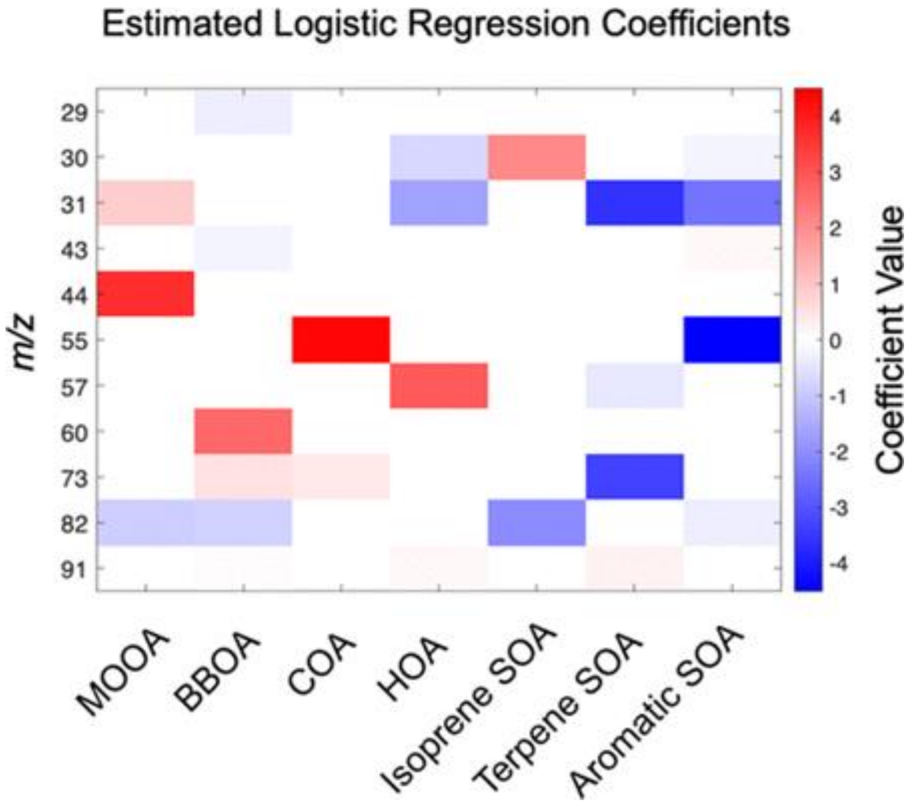
Logistic regression determined weights identify key molecular markers that distinguish different sources

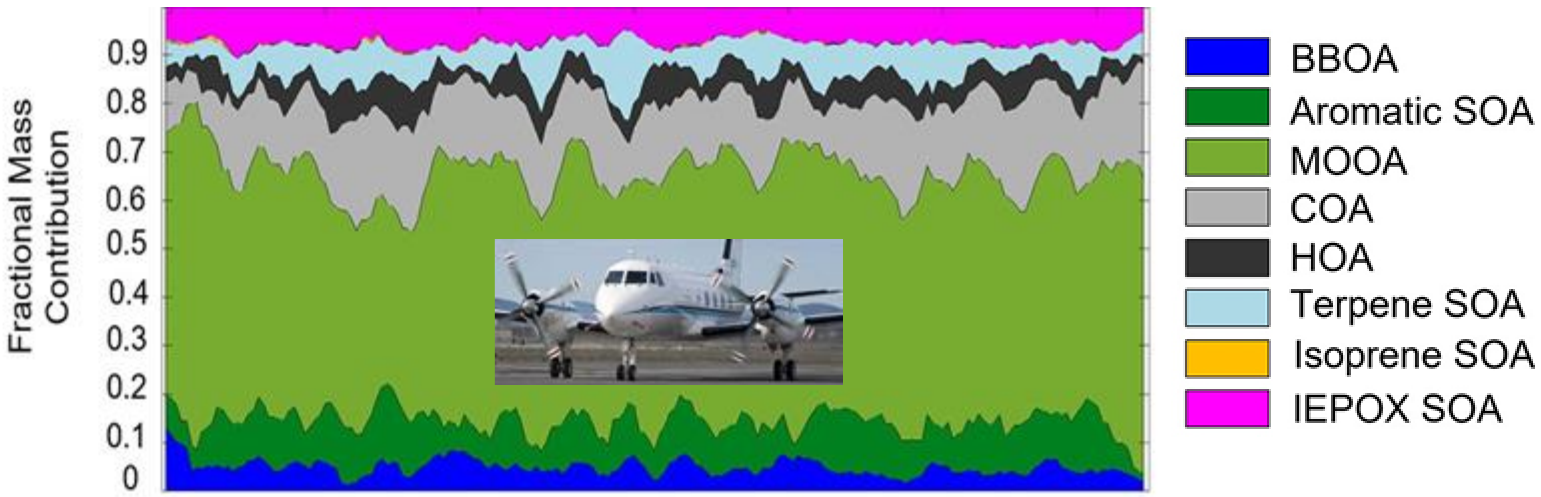


Pande, Shrivastava et al. 2022

- Weights determined by maximizing the log-odds probability of identifying OA classes
- Confirm with our domain knowledge (m/z 60 for BBOA, 44 for oxidized OA, 82 for IEPOX-SOA)
- Could be used for identifying new marker compounds for unknown source classes

Logistic regression determined weights identify key molecular markers that distinguish different sources





Pande, Shrivastava et al. 2022

- Weights determined by maximizing the log-odds probability of identifying OA classes confirms with domain knowledge of marker peaks (m/z 60 for biomass burning, 44 for MOOA, etc.)
- Aircraft data: ML model rapidly apportions mass spectra to different sources

Conclusions

- ML can perform rapid online source apportionment with ACSM & AMS data
- It can be useful when source profiles are changing rapidly
- Potential applications to long term measurements at BNF, aircraft measurements
- Once trained, the ML predictions are rapid (algebraic matrix multiplications)
- ML algorithms are freely available (python libraries)
- Much cheaper than using commercial software for data analyses