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

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

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

August 18, 2023

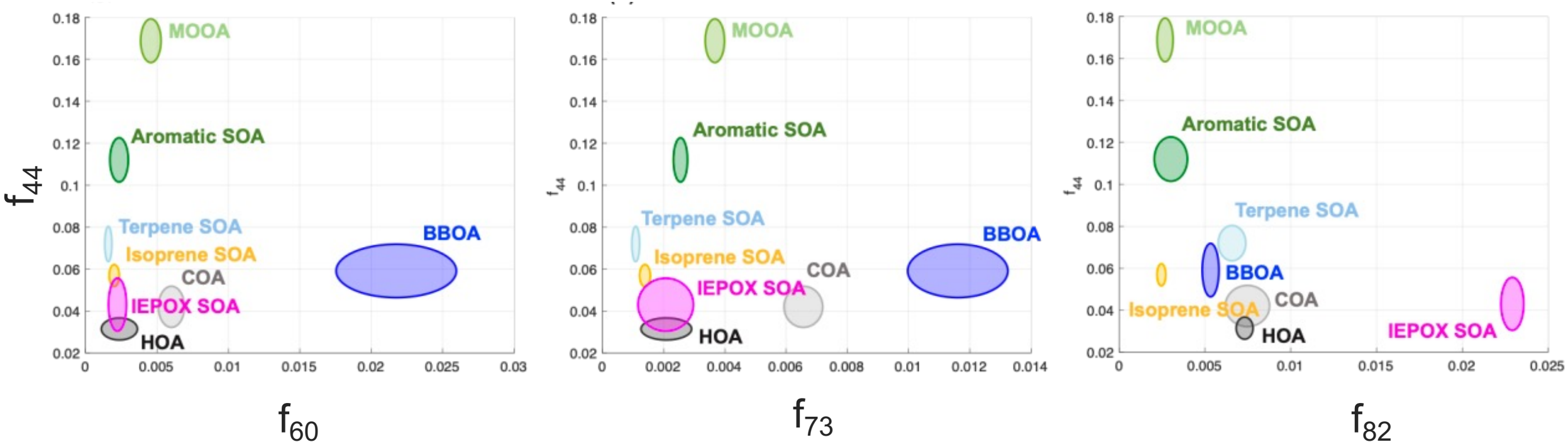
Acknowledgements: Supported by the U.S. Department of Energy's Office of Science Biological and Environmental Research Early Career Award and ASR

ASR Atmospheric System Research

Motivation

- With collection of long-term measurements of aerosol mass spectrometer data, there is a need for fast and online source-apportionment of organic aerosols
- Traditional techniques like PMF assume 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 2-step supervised machine learning algorithm to rapidly identify source profiles of single OA mass spectrum, which can be applied online as the AMS data are being collected
- Our approach does not need the entire time series mass spectra a-priori unlike PMF
- Our ML technique can be trained to rapidly identify changing emissions, source profiles etc.

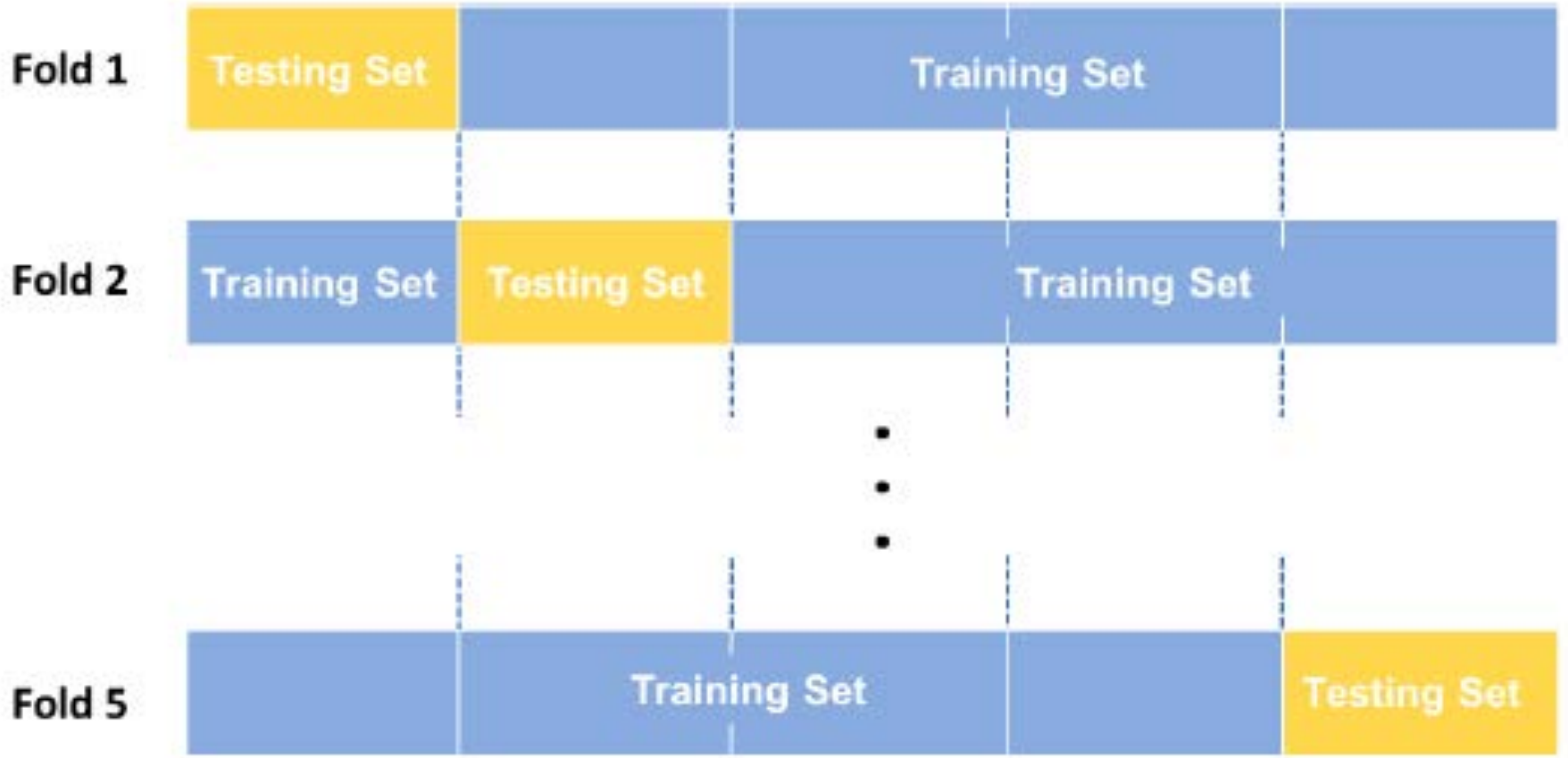
AMS data: Key mass to charge ratio markers can be used to identify different SOA sources



Pande, Shrivastava et al. 2022

- Trained the 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

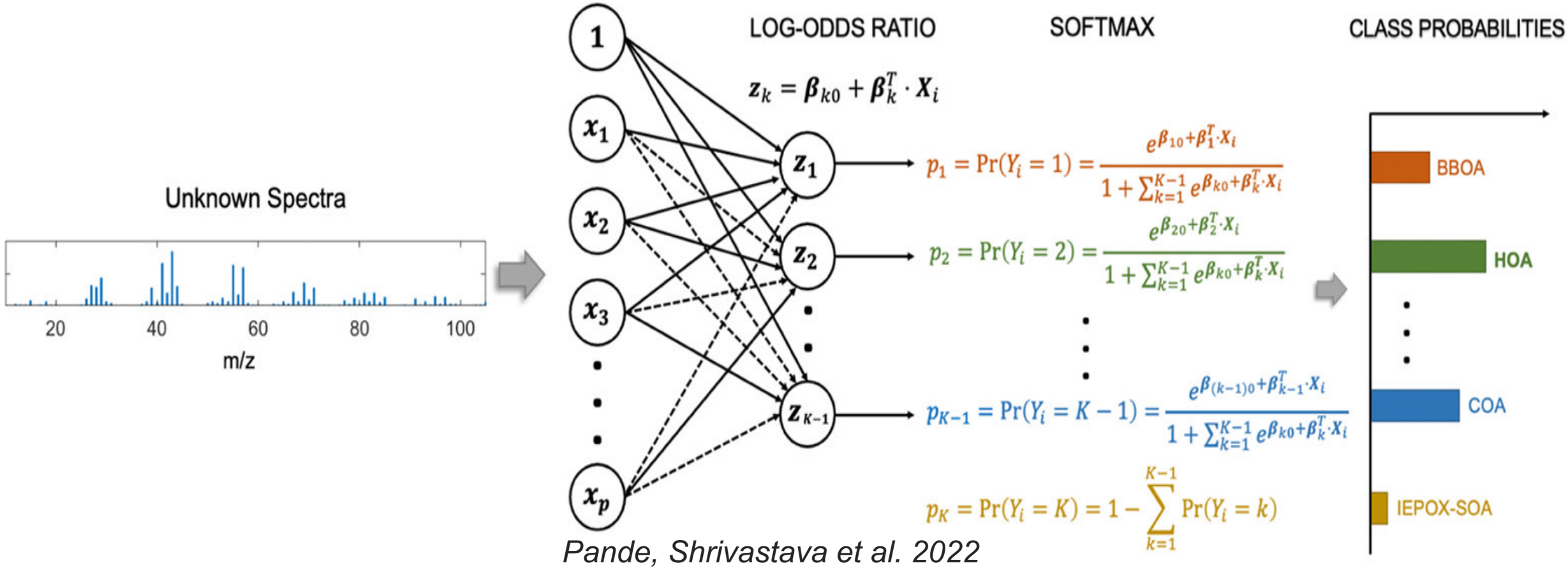
5-fold cross validation used to test generalizability of model to identify SOA mass spectra from ground truth lab & field data



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



Pande, Shrivastava et al. 2022

- The weights determined during training with L1 norm helped to focus the classifier on the most useful features and assign zero weights to marker spectra (m/z) that were not useful

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

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

STEP 2

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

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

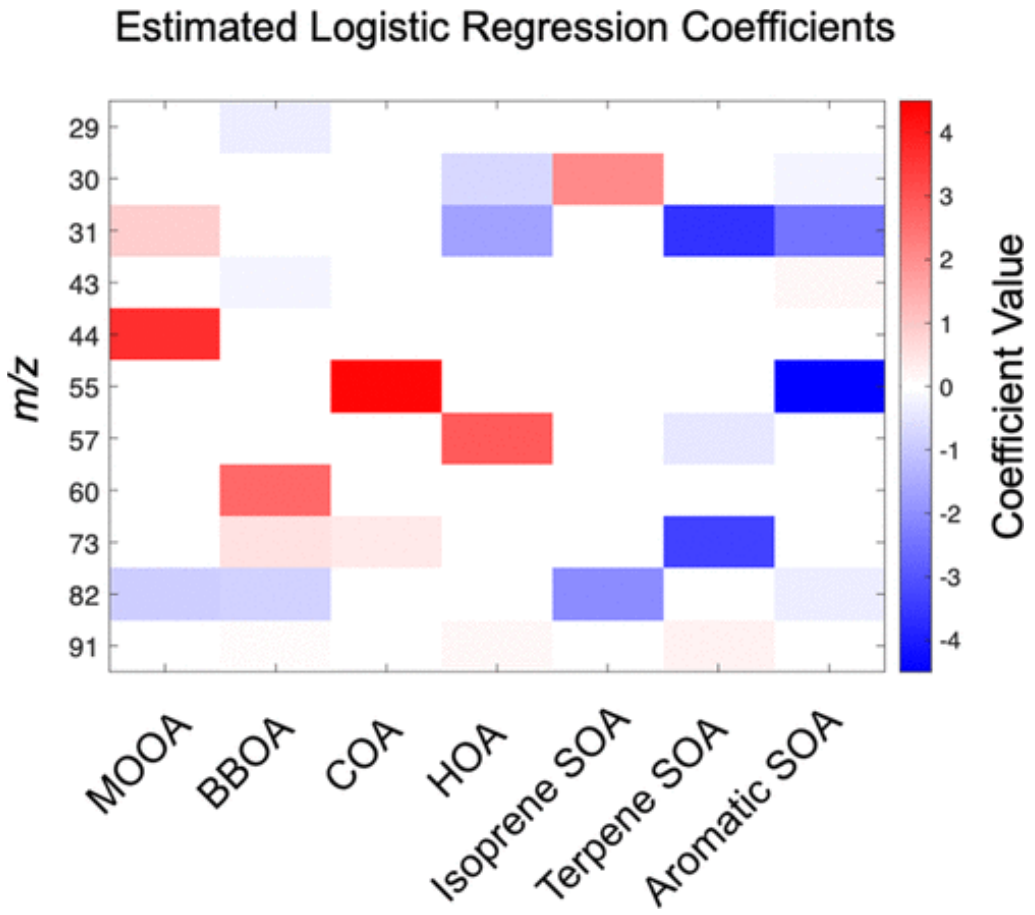
STEP 3

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 $\boldsymbol{X}_i = [\alpha_1^i, \alpha_2^i, \cdots, \alpha_K^i]$

$$\{(\boldsymbol{X}_i, y_i)\}_{i=1}^{i=N} \qquad \{(\boldsymbol{X}_i, y_i - f_1(\boldsymbol{X}_i))\}_{i=1}^{i=N} \qquad \left\{\left(\boldsymbol{X}_i, y_i - \sum_{j=1}^{T-1} f_j(\boldsymbol{X}_i)\right)\right\}_{i=1}^{i=N}$$

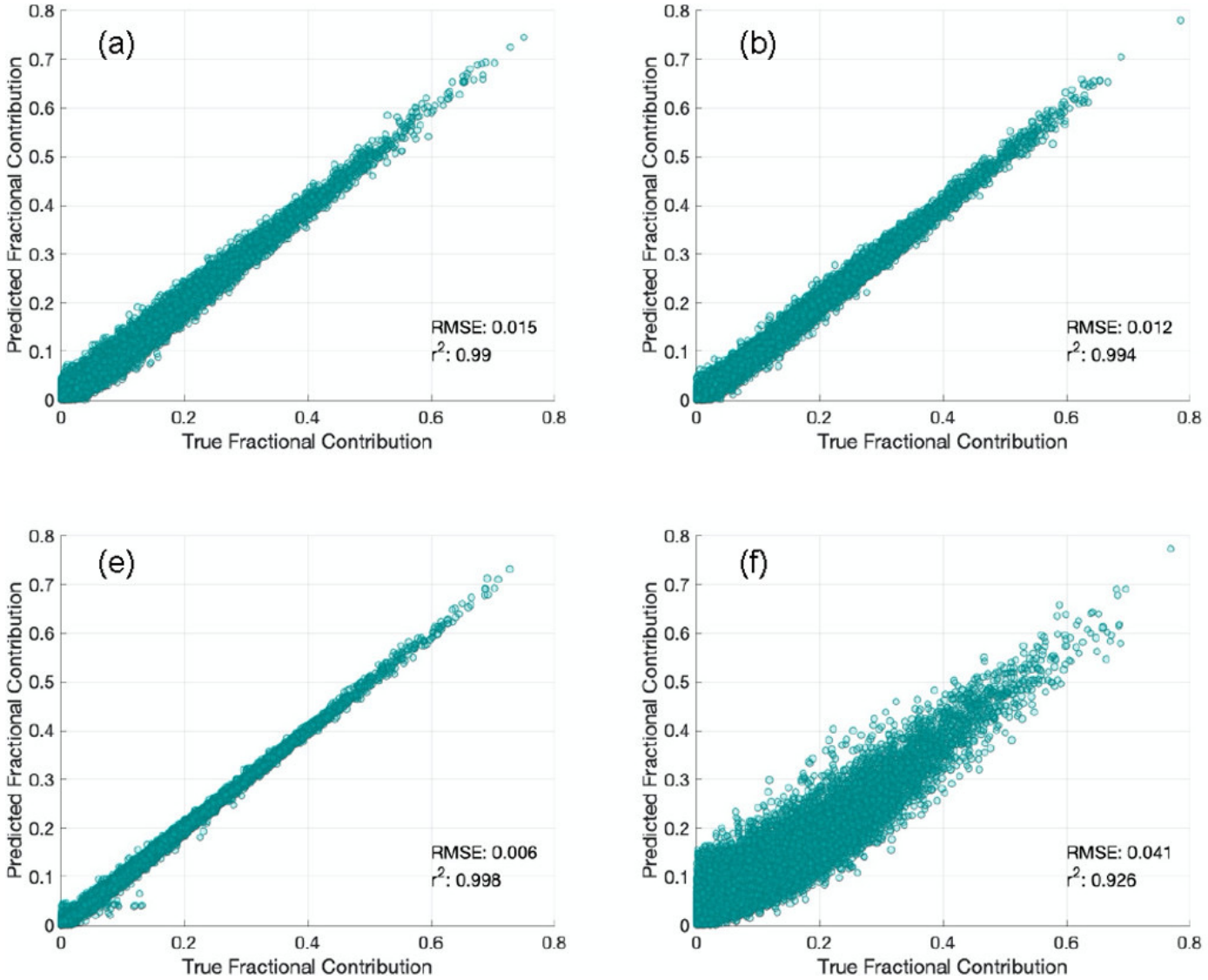
$$f(\boldsymbol{X}) = f_1(\boldsymbol{X}) + f_2(\boldsymbol{X}) + \cdots + f_T(\boldsymbol{X})$$

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

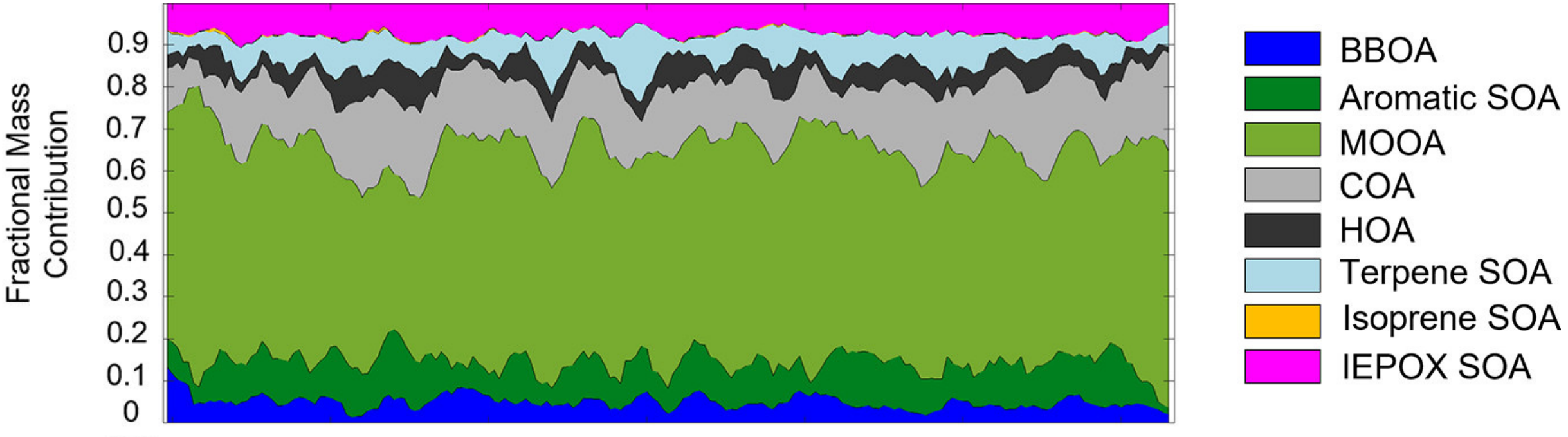


- The weights determined by maximizing the log-odds probability of identifying OA classes confirm surprisingly well with our domain knowledge of marker peaks (m/z 60, 73 for biomass burning, 44 for oxidized OA, 82 for IEPOX-SOA etc.)

Ensemble non-linear regression trained to determine source contributions from their probabilities



Applied for source apportionment of AMS data onboard aircraft during HI-SCALE 2016, May 6 2016



Pande, Shrivastava et al. 2022

Conclusions

- Developed a novel 2-step machine learning technique that leverages laboratory and field measurements to perform rapid online source apportionment of aerosol mass spectrometer data
- ML is trained a-priori to identify 8 different well characterized laboratory and field SOA types
- It can be used to rapidly classify SOA mass spectra as they are being collected
- It can be useful when source profiles are changing rapidly e.g., over aircraft and long-term observations to analyze ACSM data from ARM over multiple years
- Once trained, the ML predictions are rapid, since they are just running matrix multiplications (simple algebraic calculations)
- This technique does not require any user judgement during field applications