Contract No:

This document was prepared in conjunction with work accomplished under Contract No. DE-AC09-08SR22470 with the U.S. Department of Energy (DOE) Office of Environmental Management (EM).

Disclaimer:

Comprehensive Chemical Fingerprinting by Multidimensional GC and Supervised Machine Learning

Project highlight. This poject leverages and combines recent advances in machine learning and analytical chemistry to progress nuclear nonproliferation technologies beyond current capabilities. The developed approaches can be used to identify and detect complex chemical fingerprints of facilities of interest; these tecniques can be applied to other fields including climate sciences, environmental chemistry, and atmospheric physics.

Awards and Recognition

Intellectual Property Review

This report has been reviewed by SRNL Legal Counsel for intellectual property considerations and is approved to be publicly published in its current form.

SRNL Legal Signature

Signature	**Date**

Comprehensive Chemical Fingerprinting by Multidimensional GC and Supervised Machine Learning

Project Team: Joseph Mannion (SRNL PI), Heather Brant (SRNL), Sapna Sarupria (Clemson PI), Jiexin Shi (Graduate Student)

Subcontractor: Clemson University

Project Type: Standard

Project Start Date: October 1, 2019
Project End Date: September 30, 2021

This poject leverages advances in machine learning based data analysis techniques and untargeted omic analytical methods to progress nuclear nonproliferation technologies beyond current capabilities. The developed approaches can be used to identify and detect complex chemical fingerprints of facilities of interest. These techniques have been developed for fields such as metabolomics and genomics but have not been applied to nuclear nonproliferation applications. Adaptation of these techniques for volatile organic compound analysis has far reaching application within the scientific community including environmental chemistry, atmospheric physics, and climate sciences.

FY2020 Objectives

- Analytical method development for multidimesional gas chromatography analysis of volatile organic compounds
- Training data set collection utilizing multidimensional gas chromatography
- Machine learning based data anlysis development utilzing open source data

Introduction

Volatile organic compound (VOC) collection and analysis techniques have been under development at SRNL for more than two decades for national security applications. Traditionally these approaches have attempted to identify one to two "signature" species that are indicative of a given activity. The shortcoming of these efforts has been the fundamental limitations of a silver bullet approach with regards to organic signatures. VOC production and emissions are complex, highly dynamic, and subject to complicating matters such as holdup, chemical transformations, and complex backgrounds (up to 10,000 unique chemical species have been identfied in a single air sample). Despite these challenges, VOC signatures are attractive as their inherent complexity is in part due to their sensitivity to process conditions within a facility; therefore, VOC signatures carry complex process information that can be used to deeply characterize operations.

The major goal of this project is to utilize machine learning based data analysis approaches to develop multi-species chemical signature "fingerprints" of processes relevant to nonproliferation interests. This untargeted approach will assess organic emissions as a comprehensive collection, rather than a "silver bullet" approach, to create more robust and informative chemical fingerprints of activities. The objective is to collect, analyze, and identify patterns present in measured volatile organic emissions from facilities of interest. The product of this work is data collection modalities, machine learning based data analysis alogrithms, and a fingerprint database allowing for identification and assessment of activities (ex. process upsets, efficiencies, etc.).

A multipronged approach was taken for project efforts in FY20. The focus at SRNL was analytical method development for comprehensive VOC analysis utilizing the multidimensional gas chromatograph procured in FY19. This system is one of the most powerfull commercially available instruments for VOC analysis and was found to afford ~4 orders of magnitude improved sentivity (and 1 – 2 orders of magnitude peak capacity) over traditional GC/MS systems at SRNL. The focus at Clemson University was the development of machine learning based data analysis approaches utilizing the open source EPA Speciate database. This database contains more than 3,000 pollution profiles from industrial, commercial, and residential emmission sources. Traditional chemometric techniques were compared to the machine learning based approaches with the intent of publishing the results in FY21. Additionally, a variety of machine learning approaches for plume detection were developed at SRNL utilizing historic single-dimensional GC data sets acquired in previous efforts. Efforts in FY21 will focus on merging the newly developed analytical capabilities at SRNL with data analysis methods developed at Clemson University with the goal of demonstrating these techniques in future SRNL sampling campaigns.

Approach

Multidimensional gas chromatography is an established technique for the analysis of highly complex samples.[1] It is uniquely suited to applications involving complex matrices and hundreds to thousands of analyte species. Thousands of volatile organic compounds have been identified in the atmosphere that arise from both biogenic and anthropogenic sources.[2, 3] When MDGC is applied to complex samples and coupled with multichannel detectors, such as mass spectrometers, enormous amounts of data are generated (on the order of gigabytes for a single run). Traditional data analysis methods are not practical with such large data sets; therefore, modern data analysis techniques must be applied that take advantage of the higher order dimensionality of the data sets. These methods convert chemical data into information using algorithms. Machine learning based clustering and pattern recognition is utilized for this work.

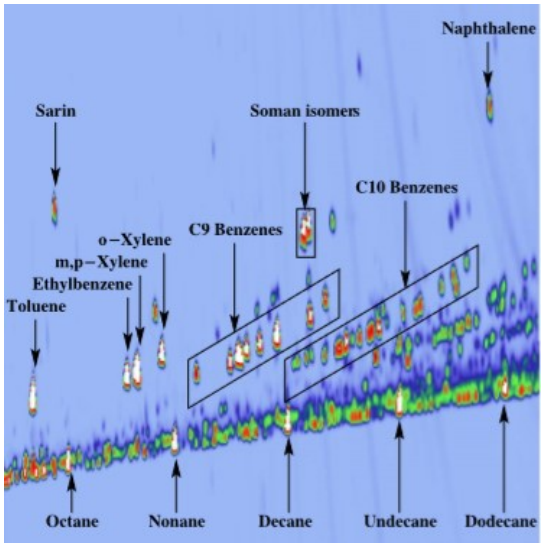

Figure 1: Separation of chemical warfare agents from a complex matrix via multidimensional GC

The number of applications utilizing machine learning has vastly expanded in recent years; however, limitations, pitfalls, and hurdles exist in the implementation of machine learning techniques to some applications. Analysis of complex VOC emmisions is an example of the curse of dimensionality.[4] In essence, when the dimesionality of a problem increases (i.e. the number of chemicals present in a sample) the volume (i.e. data space) grows so quickly that the data becomes sparse. As the number of features (i.e. chemicals) increases, the data (i.e. number of samples) must grow exponentially to maintain accurate representation; for example, a system with 15 features may require millions of samples to accurately classify the system. Application of machine learning techniques for complex systems such as atmospheric VOC analysis containing thousands of features therefore requires implentation of approaches such as dimensionality reduction and feature engineering to overcome this curse of dimensionality.[5] These techniques and various clustering approaches are explored in this work utilizing an adequately complex data set that represents real world data.

Results/Discussion

The EPA Speciate database was used for data analysis algorithm development. This database contains pollution profiles from more than 3,000 industrial, commercial, and residential emmision sources (i.e. samples) and over 20,000 unique chemicals (i.e. features). This database represents a high dimensionality data set with more features than observations. This situation will likely be encountered in any atmospheric VOC analysis application due to the abundance and variety of VOCs in the atmosphere and the limited analytical throughput of instrumentation (typically 25 – 60 samples/day for a GC system). The database was first curated to remove spurious and incomplete entries then feature engineered to reduce dimesionality based on chemical features. This was achieved by grouping chemicals with similar chemical functionality using Python and the CAS or SMILES index of a given chemical. This resulted in the reduction of >20,000 features to 54 features representing chemical classifications such as alkanes, alcohols, carboxylic acids, etc. and greatly reduced the sparcity of the data set. The reduced dataset was then subjected to various linear and non-linear data reduction methods including principle component analysis

(PCA), t-distributed stochastic neighbor embedding (t-SNE), locally linear embedding (LLE), uniform manifold approximation and projection (UMAP), and autoencoders (AE). Simple features scaling such as the standard scalar method were investigated. Various clustering algorithms were then used to assessthe quality of data reduction and scaling based on the silhouette coefficient including density-based spatial clustering of applications with noise (DBSCAN) and k-nearest neighbor (KNN) clustering. Non-linear machine learning dimensionality reduction methods such as t-SNE were found to greatly outperform classic chemometric approaches such as PCA (Fig. 2).

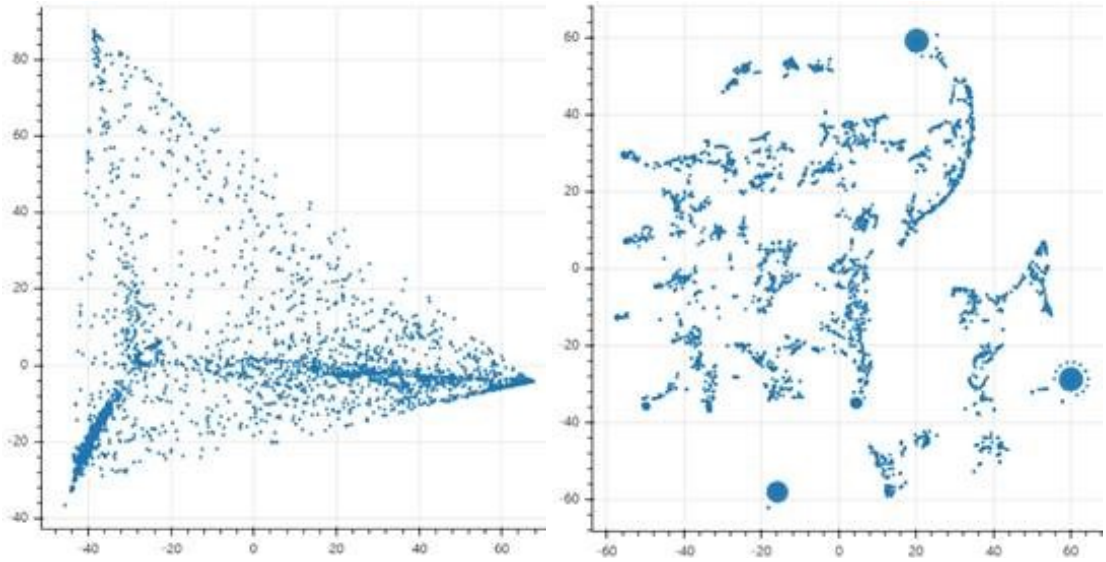

Figure 2: Data visualization of the EPA Speciate database after processing; each point represents a polution profile contained in the database. (Left) 2-component 11 feature PCA data reduction representing classic chemometric approaches and (right) t-SNE non-linear machine learning based dimensionality reduction represented (visually?) in 2 components.

Analytical methods for comprehensive organic analysis were developed utilizing an Agilent 7250 GC quadropole time-of-flight high resolution mass spectrometry (Q-TOF) equiped with a Zoex ZX2 thermal modulator for multidemsional analysis. Methods were developed for a variety of potential sample types of interest for future training data set development. Paramter optimization included standard GC method development such as inlet type/temperature and oven ramp programs as well as multidimensional parameters such as modulation period and primary/secondary column selection. The viability of various training data set collections was assessed including on-site environmental sampling at SRS. The multidimensional GC/Q-TOF was found to be several orders of magnitude more sensitive than traditional GC/MS systems, with sub-picogram detection limits. This is in part due to the high resolution mass spectrometer, which allows for the seperation of isobars and the isolation of target ions. This is illustrated in Figure 3; 6 ions are identifiable at a nominal mass-to-charge (m/z) ratio of 96. Traditional unit mass resolution mass spectrometers would display a single summed peark of all ions near m/z 96. The high resolution Q-TOF allows for the extraction of trace target ions from interfering species, greatly improving detection limits.

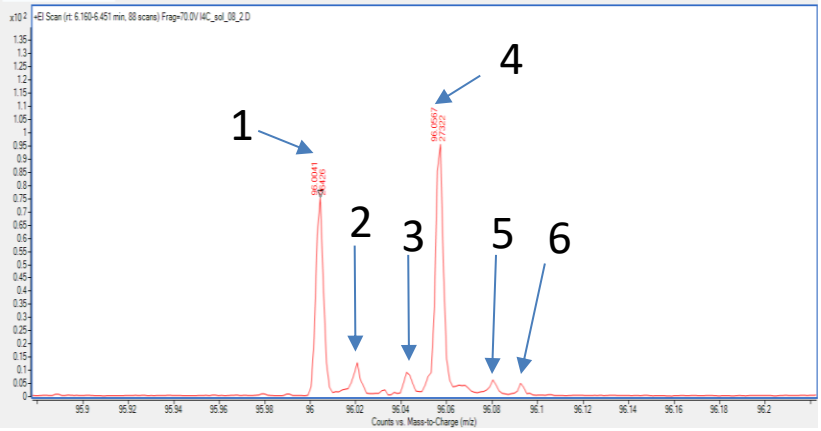

Figure 3: High resolution Q-TOF allows for the seperation of isobaric interferences greatly improving the sensitivity for target trace ions.

In addition to sensitivity improvements derived from the use of a high resolution mass spectrometer, the use of a thermal modulator to create multidimensional GC chromatograms further improves the sensitivity of analyses by approximately 1.5 orders of magnitude (Figure 4). This is due to a chromatographic focusing effect which improve the signal to noise ratio of the chromatographic peak. These combined improvements result in a MS based system that is more sensitive for known target compounds than the traditional electron-capture-detectors (ECD). Although ECD detectors are highly sensitive, species identification is ambiguous due to solely relying upon a retention time. Traditionally, GC/MS systems have not been sensitive enough for end-use applications which necessitated the use of more sensitive but ambiguous ECD systems. Developing MS based analytical methods capable of out-performing ECD systems is a major milestone in operationalization of SRNL developed technologies. Additionally, the use of MS based systems allows for identification of a broad range of chemical species, enabling development of multi-species chemical fingerprints of activities of interest.

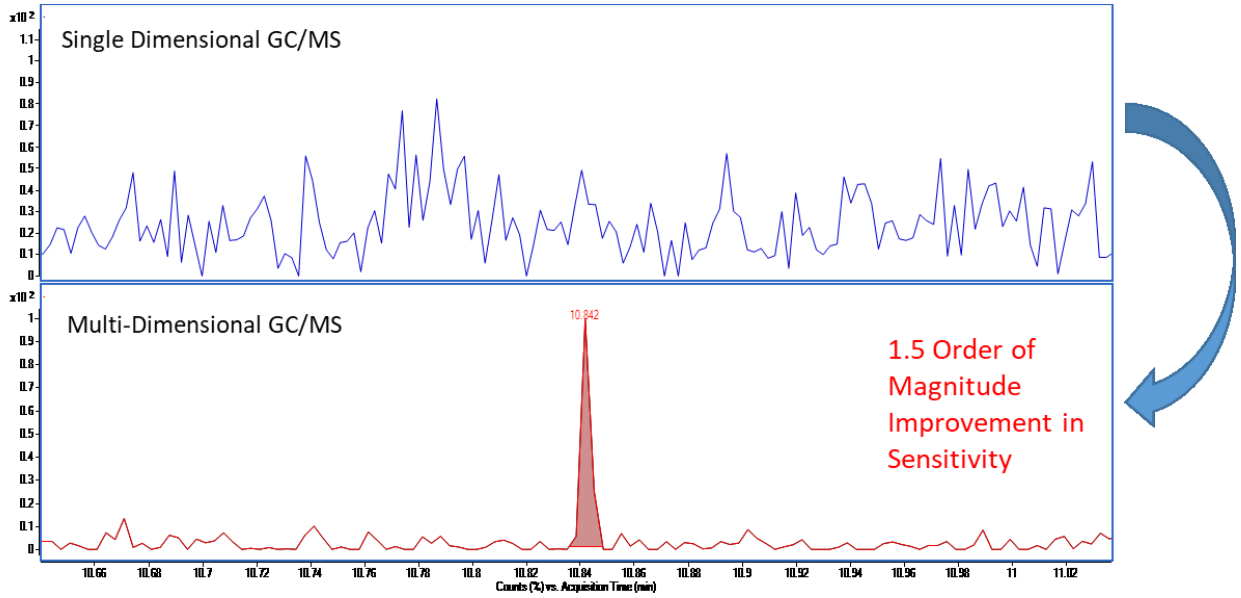

Figure 4: Comparison of sensitivity of the GC/Q-TOF in single-dimensional operation vs. multi-dimensional thermal modulation. The improvement in sensitivity is due to chromatographic focusing of the sample.

FY2020 Accomplishments

- 60 page literature review on machine learning applications in GC/MS data analysis provided by Clemson University
- Developed machine learning based approaches for organic fingerprint detection utilizing the open source EPA SPECIATE database
- Analytical methods developed on new instrumentation improve sensitivity by 4 orders of magnitude over traditional GC/MS systems for species of interest and 1 order of magnitude over GC/ECD systems representing an order of magnitude increased detection range
- Developed multiple machine learning algorithms capable of detecting plume “hits” using historic single dimensional data sets (SRNL)
- SRNL PI taught himself Python and R for data processing and algorithm development
- LDRD reserach presented to 5 NNSA program managers
- New skill-set development for SRNL researcher
- Completion of instrument installation including facility modifications
- Establishment of a University subcontract and collaborative relationship with new external researchers and graduate students

Future Directions

- Continued development of machine learning based data analysis approaches utilizing the open source data including the further development of auto encoder and uniform manifold approximation and projection dimensionality reduction methods
- Merge SRNL analytical methods and data collections with data analysis algorithms developed at Clemson University
- Continued collection of training data sets utilizing SRNL analytical methods
- Development of adversarial controls to test the accuracy of machine learning based classifications

FY 2020 Peer-reviewed/Non-peer reviewed Publications

Publication of algorithm development activities is expected in early-mid 2021.

Presentations

2 Presentations to NNSA headquarters program managers

Planned: 19th Conference on Artificial Intelligence for Environmental Science
Planned: ISCC & GCxGC 2021

References

[1] S.E. Prebihalo, K.L. Berrier, C.E. Freye, H.D. Bahaghighat, N.R. Moore, D.K. Pinkerton, R.E. Synovec, Multidimensional gas chromatography: advances in instrumentation, chemometrics, and applications, Anal. Chem. 90(1) (2017) 505-532.
[2] C. Arsene, D. Vione, N. Grinberg, R.I. Olariu, GC× GC-MS hyphenated techniques for the analysis of volatile organic compounds in air, Journal of Liquid Chromatography & Related Technologies 34(13) (2011) 1077-1111.
[3] J. Hamilton, P. Webb, A. Lewis, J. Hopkins, S. Smith, P. Davy, Partially oxidised organic components in urban aerosol using GCXGC-TOF/MS, Atmospheric Chemistry and Physics 4(5) (2004) 1279-1290.

[4] M. Verleysen, D. François, The curse of dimensionality in data mining and time series prediction, International work-conference on artificial neural networks, Springer, 2005, pp. 758-770.
[5] E.W. Newell, Y. Cheng, Mass cytometry: blessed with the curse of dimensionality, Nature immunology 17(8) (2016) 890-895.

Acronyms

AE- autoencoders
DBSCAN- density-based spatial clustering of applications with noise
ECD- electron capture detector
EPA- Environmental protection agency
GC- Gas chromatograph(y)
KNN- k-nearest neighbor
LLE- locally linear embedding
m/z- Mass-to-charge ratio
MS- Mass spectrometer
NNSA- National nuclear security administration
PCA- principle component analysis
Q-TOF- Quadropole time-of-flight
SRNL- Savannah river national laboratory
t-SNE- t-distributed stochastic neighbor embedding
UMAP- uniform manifold approximation and projection
VOC- Volatile organic compound

Intellectual Property

None to report.

Total Number of Post-Doctoral Researchers

0

Total Number of Student Researchers

1, Clemson University PhD Student, Jiexin Shi

External Collaborators (Universities, etc.)

Clemson University Department of Biomolecular and Chemical Engineering