Contract No:

This document was prepared in conjunction with work accomplished under Contract No. 89303321CEM000080 with the U.S. Department of Energy (DOE) Office of Environmental Management (EM).

Disclaimer:

This work was prepared under an agreement with and funded by the U.S. Government. Neither the U.S. Government or its employees, nor any of its contractors, subcontractors or their employees, makes any express or implied:

1) warranty or assumes any legal liability for the accuracy, completeness, or for the use or results of such use of any information, product, or process disclosed; or
2) representation that such use or results of such use would not infringe privately owned rights; or
3) endorsement or recommendation of any specifically identified commercial product, process, or service.

Any views and opinions of authors expressed in this work do not necessarily state or reflect those of the United States Government, or its contractors, or subcontractors.

LDRD-2020-00020
LDRD External Report Summary

Title of Project
Comprehensive Chemical Fingerprinting by Multidimensional GC and Supervised Machine Learning

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

Project Highlight
No more than one or two sentences to highlight the impact for the nation and overview the technology in “layman’s terms.” Please do not use acronyms. Should include one image that reflects the key achievement or importance of the project.

Project Team
Principal Investigator: Joe Mannion
Team Members: Heather Brant, Stephanie Gamble, Eric Hoar
External Collaborators: Sapna Sarupria (Clemson University PI), Jiexin Shi (Clemson University Graduate Student)

Abstract
This project leveraged advances in machine learning based data analysis techniques and untargeted analytical methods for organic analysis to progress nuclear nonproliferation technologies beyond current capabilities. The developed approaches can be used to detect and identify 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.

Objectives

- Develop multidimensional gas chromatography high resolution mass spectrometry analytical methods for the analysis of volatile organic compounds.
- Collect a training data set utilizing multidimensional gas chromatography for algorithm development.
- Develop machine learning based data analysis algorithms for complex VOC profiles using an open-source data set.

REVIEWS AND APPROVALS

1. Authors:

____

____

____

Name and Signature Date

2. Technical Review:

____

Name and Signature Date

3. PI's Manager Signature:

____

Name and Signature Date

4. 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

____

Name and Signature

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 identified in a single air sample). Despite these challenges, VOC signatures are attractive as they can provide unique information.

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 algorithms, and a fingerprint database allowing for identification and assessment of activities.

A multipronged approach was taken for project efforts. 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 powerful commercially available instruments for VOC analysis and was found to afford ~4 orders of magnitude improved sensitivity (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 emission sources. Traditional chemometric techniques were compared to the machine learning based approaches.

Approach

Multidimensional gas chromatography (MDGC) is an established technique for the analysis of highly complex samples (Figure 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. 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.

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 emissions is an example of the curse of dimensionality. In essence, when the dimensionality 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 implementation of approaches such as dimensionality reduction and feature engineering to overcome this curse of dimensionality. These techniques and various clustering approaches are explored in this work utilizing an adequately complex data set that represents real world data.

Accomplishments

- Secured direct follow-on funding for FY22 at $500K; follow-on project is titled "GC x GC Analysis of Organic Signatures" and is projected to be funded through FY24 at a total project budget of $1,500K.
- Analytical methods developed under this LDRD project have been incorporated into the scope 3 existing projects at a total FY22 funding level of $2,700K.
- Analytical methods developed on new instrumentation improve sensitivity by 4 orders of magnitude over traditional GC/MS systems for species of interest. An example multidimensional chromatogram of air collected at SRNL is shown in Figure 1; thousands of species are detectable in a single air sample.
- Measurement accuracy was increased by a factor of 4 over previous analytical techniques.
- Developed machine learning based approaches for organic fingerprint detection utilizing the open-source EPA SPECIATE database. A comparison of machine learning based clustering vs. classic PCA is shown in Figure 2. An examples of unsupervised cluster identification is show in Figure 3.
- 60-page literature review on machine learning applications in GC/MS data analysis provided by Clemson University

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 2021 Peer-reviewed/Non-peer reviewed Publications

N/A

Intellectual Property

List all invention disclosures, copyright disclosures, patent applications, and patents granted.

Total Number of Post-Doctoral Researchers

Stephanie Gamble – On-site post-doctoral researcher
Eric Hoar - On-site post-doctoral researcher

Total Number of Student Researchers

Jiexin Shi – PhD Student at Clemson University Department of Chemical and Biomolecular Engineering

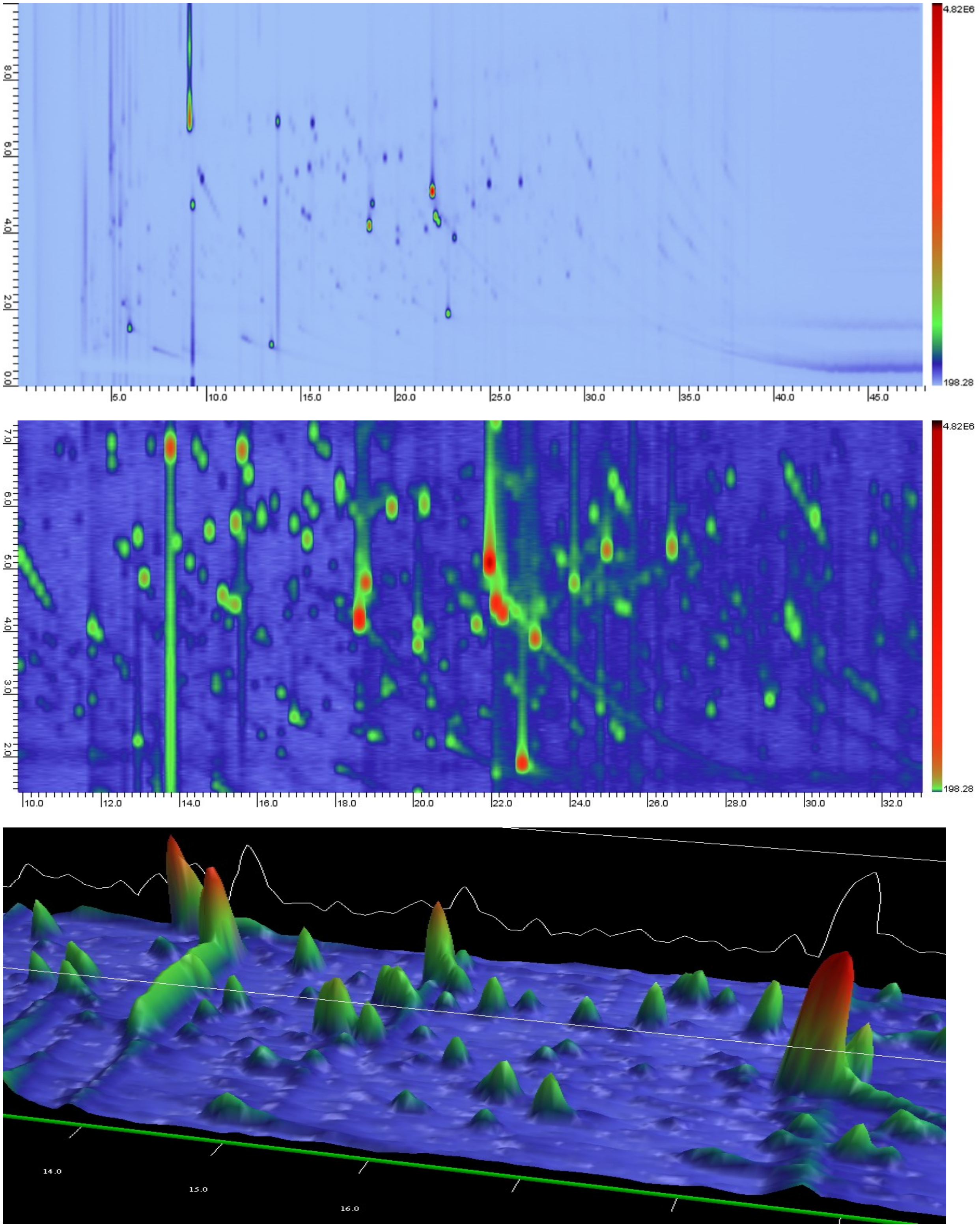

Figure 1: (Top) An example multidimensional chromatogram of air collected at SRNL. (Middle) Increasing magnification of the air sample reveals thousands of trace level chemical species present in a single air sample represented by green ovals. (Bottom) 3D rendering of a trace levels compound detection.

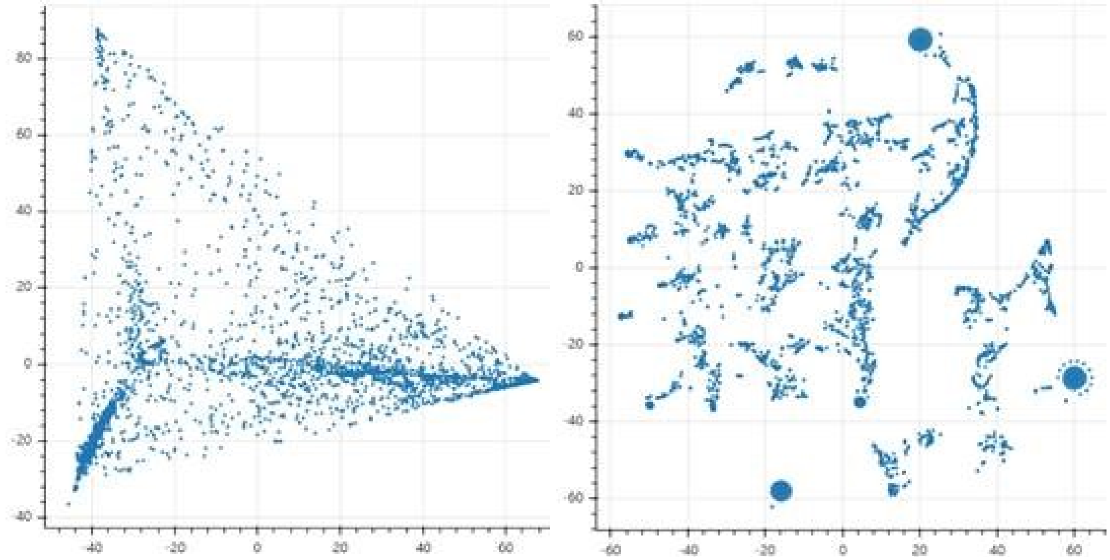

Figure 2: Data visualization of the EPA Speciate database after processing; each point represents a pollution profile contained in the database. (Left) 2-component 11 feature PCA data reduction representing classic chemometric approaches and (right) t-SNE non-linear machine learning based dimensionality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

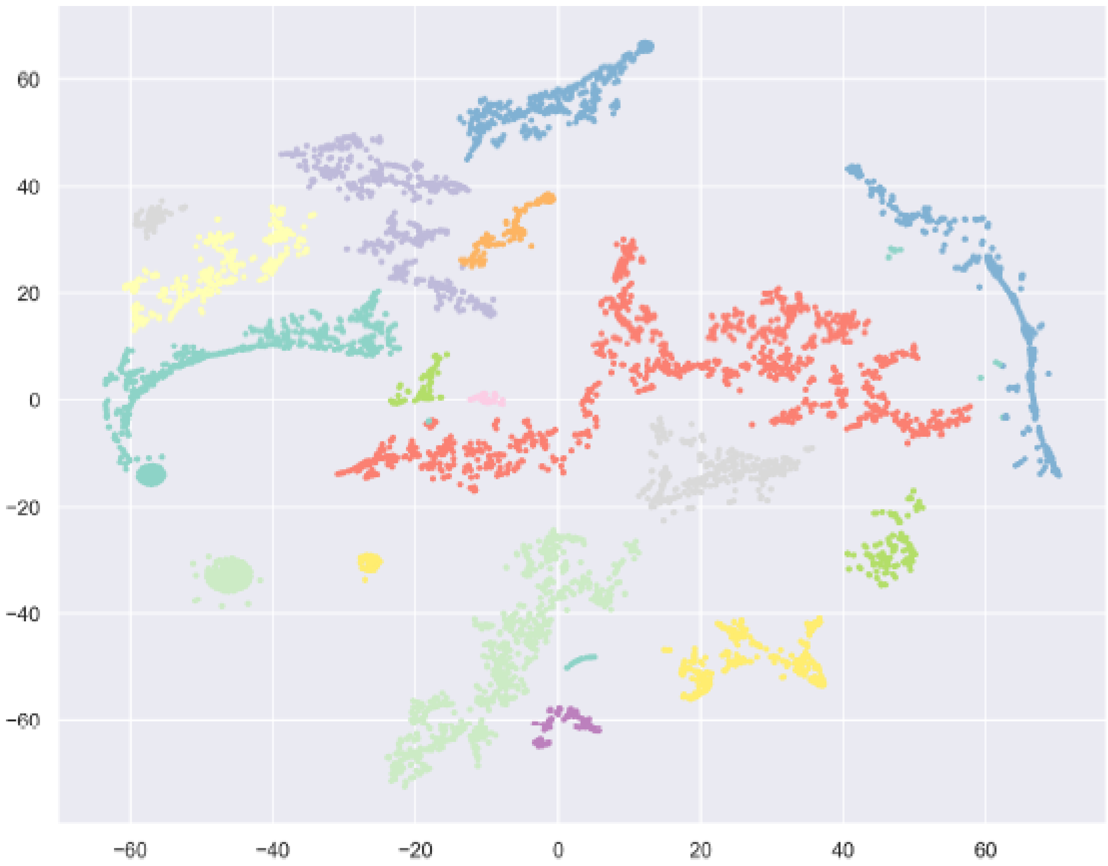

Figure 3: Data visualization of unsupervised cluster identification of the EPA Speciate database.