2022

Development of Scent Detection and Categorization Algorithm Using Gas Chromatography and Machine Learning

Alex Driehaus

DEVELOPMENT OF SCENT DETECTION AND CATEGORIZATION ALGORITHM USING GAS CHROMATOGRAPHY AND MACHINE LEARNING

A Capstone Experience/Thesis Project Presented in Partial Fulfillment

of the Requirements for the Degree Bachelor of Science

with Mahurin Honors College Graduate Distinction

at Western Kentucky University

By

Alexandra L. Driehaus

May 2022

*****

CE/T Committee:

Dr. Ivan Novikov, Chair

Dr. Vladimir Dobrokhotov

Dr. Eric Conte

ABSTRACT

There are many looking to connect human senses to quantifiable data. Scents are categorized by their descriptions into scent families. These include citrus, floral, and woody. Similar descriptors designate similar families, while different descriptors correlate with different families. Dravnieks compiled an Atlas of chemical descriptors **[1]**. Such descriptors are cinnamon, fruity, and cadaverous. By analyzing the applicability of these descriptors, the chemicals will be sorted into their scent families.

Gas chromatography generates sample-specific signals of voltage over time. Chromatograms of known scents will serve as a basis for a convolutional neural network. This algorithm will be trained on these signals and tested with unknown scents to categorize scents with no human participation. We seek to generate verbal descriptions of scent through machine learning analysis of GC signals.

I dedicate this thesis to my cat, Henry Fluffypants Driehaus, who is responsible for my emotional stability, even from more than 750 miles away, and his puppy pal Cassidy. I also dedicate this to my parents, Mr. Paul Driehaus and Dr. Tracy Driehaus, who have supported me throughout this experience.

ACKNOWLEDGEMENTS

I would like to thank Dr. Dobrokhotov for his continued support and allowing the use of his GC and API's facilities. Also, I would like to thank Mr. Matthew Pimienta for his contributions. This project was funded by KY EPSCoR and a WKU FUSE grant. I would like to profusely thank Dr. Ivan Novikov for his support throughout the last two and a half years.

VITA

EDUCATION

Western Kentucky University, Bowling Green, KY — May 2022
B.S. in Physics – Mahurin Honors College Graduate
Honors CE/T: *Development of Scent Detection and Categorization Algorithm using Gas Chromatography and Machine Learning*

Pottsgrove High School, Pottstown, PA — May 2018

PROFESSIONAL EXPERIENCE

Department of Physics and Astronomy, WKU — May 2020-Present
Research Assistant, Learning Assistant, Tutor

Alabama Plasma Internship Program, UAH and Alabama A&M — May 2021-Aug. 2021
Research Intern

AWARDS & HONORS

Cum Laude, WKU, May 2022
Dr. Randall Harper Award for Outstanding Research in Physics and Astronomy, WKU, May 2021
Dr. Douglas Humphrey Award for Outstanding Service in Physics and Astronomy, WKU, May 2021
Sarah and Mark Rogers Physics Enhancement Fund, WKU, 2019-2022

PROFESSIONAL MEMBERSHIPS

Sigma Pi Sigma Physics Honor Society (**ΣΠΣ**)
American Physical Society (APS)
Society of Physics Students (SPS)
American Institute of Physics (AIP)

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

There are many working on the connection between the human experience and quantitative data. Tools that gather and analyze this data include electronic noses ("e-Nose") and electronic tongues ("e-Tongue"). Examples of usage for e-Tongues include tasting human sweat to determine stress levels **[2]** and categorizing components of cheddar cheese **[3]**. These e-Tongues utilize high performance liquid chromatography (HPLC). Chromatography practices generate signals of voltage over time specific to the chemical make-up of the sample. From these signals, the type of molecule, and the concentration of each type, can be found.

Like e-Tongues, e-Noses have been studied for many different applications. e-Noses have been studied for a while. In 1995, Rastogi applied GC techniques to analyze cosmetics **[5]**. More recently, Keller's group set to predict how the shape of a molecule could predict the human perception of the molecule's scent **[4]**. Their findings suggest that there is still no way to determine, from appearance alone, how a human olfactory system will perceive a chemical's scent. The goal of this project is to utilize gas chromatography (GC) to create an algorithm that can categorize the scent of chemicals without the use of human participation.

GCs work by taking a sample of gas, heating it up, and then allowing it to diffuse through a column. GC signals have two aspects, both sensor and substance-specific, as part of their signals. Retention time is how long it takes for a given substance to migrate through stationary phase, and the detection limit is the minimum amount of a specific substance

needed to confirm its presence. The width of the peak correlates to the resolution of the column. The thinner the peak, the more efficient the column.

Scents are categorized into families by their descriptions. For example, a scent with notes of lemon or orange would be in the citrus family, and a scent with notes of oak and cedar would be woody. Much like the color wheel, those versed in scent categorization have a scent wheel, as shown in **Fig. 1**. The way that scents are categorized leads to those from the same family having similar descriptions, and those from different families having different descriptions.

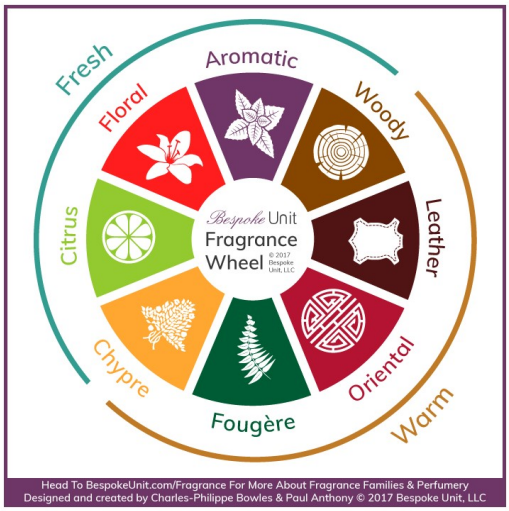

Figure 1. Scent wheel showing the different scent families.

In 1985, Andrew Dravnieks published his Atlas of Odor Character Profiles. The Altas is a collection of 160 tables. The chemicals chosen are mostly scent or flavor additives. Each table is for a different chemical and measures the applicability of 146 different descriptors on a scale of 0-5, 5 being entirely applicable and 0 being entirely non-applicable. To obtain the applicabilities, Dravnieks had groups of about 130 people sniff

each chemical and fill out a survey, where they were asked to rate each applicability. The Atlas serves as the basis of the scent categorization utilized to create the algorithm, and choose chemicals to use as the known scents in the experimental portion of the project. **[1]**

DESCRIPTION OF THE PROCESS

To start, the chemicals from *The Atlas* were sorted into their respective scent families. Then, a bank of 160 data points existed, each 146 dimensional. The dimensionality made the data difficult to visualize. So, visualizations of the data, in two and three dimensions, were generated using Python to better understand the data.

During this time, chemicals were also purchased for the experimental portion of the project. Nine chemicals were chosen, three chemicals each from three families. Those chosen were chemicals deemed to have the greatest alignment for their given family. For each of these chemicals, the retention time and detection limit were found. The detection limit for the GC used in this project was determined to be the concentration of a sample that corresponds to a peak of 100 raw ADC units on the program created to gather the chromatographic data. These raw signal units can easily be converted to Volts for data analysis. As retention times differ, some samples needed to be run multiple times to find the actual retention time, as some occurred after the pre-set timing of acquisition for the GC device provided by API.

Once these values were gathered for each of the nine experimental chemicals, the chromatographic data was used as the basis for the Convolutional Neural Network (CNN).

EXPERIMENTAL SETUP

Our data was collected on a GC apparatus supplied by API at WKU. The GC's work by heating up the sample and allowing it to diffuse over a detector. Diagrams of the apparatus are shown in **Fig. 2** and **3**.

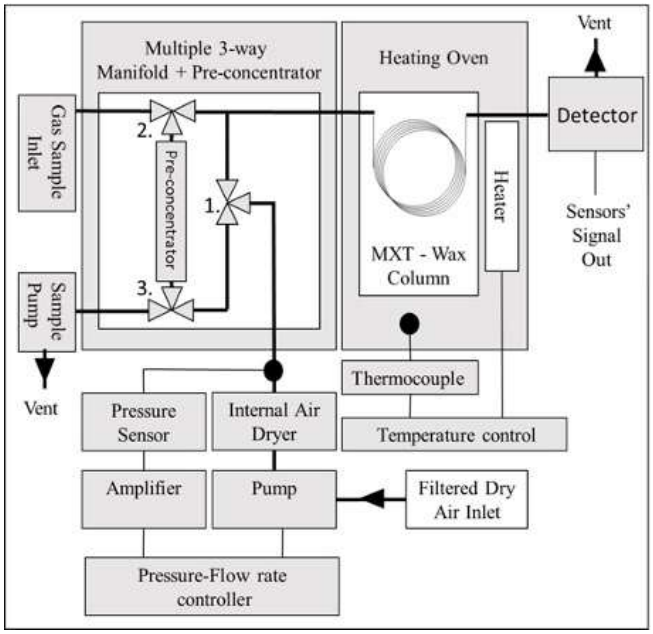

Figure 2. Diagram of GC

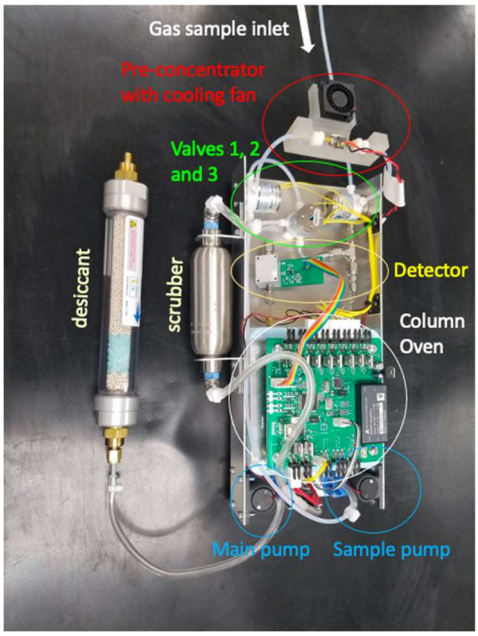

Figure 3. Picture of the GC with labels

Samples are prepared by first filling a bag with 1L of filtered air. The volume of a substance needed for a specific concentration is calculated. Then, a gas tight syringe, as shown in **Fig. 4**, is used to inject the necessary volume of the substance into the bag. The bag is then connected to the sample inlet on the GC.

Figure 2. Picture of gas-tight syringe

When the apparatus is run, a subsample of the prepared sample is injected into the apparatus. This sample is then heated by the oven and allowed to diffuse over the detector. The detector, then, will create a signal of raw ADC data over time, which can be converted to a Voltage over time signal, based on the specific sample.

EXPERIMENTAL DATA

The chemicals from *The Atlas* were sorted into their scent families. It was found that the number of samples from each family was different. For example, only three chemicals would fit into the citrus family, but forty align with a woody family description. To better identify these scents, spider plots were generated using Python to test the similarity between families. The citrus spider plot is shown in **Fig. 5**, but the others are all shown in Appendix A. In the plot, each color corresponds to a different chemical. As previously stated, three are citrus, so there are three different plots in **Fig. 5**. The distance from the center on each line corresponds to one quarter of the percent applicability for each descriptor along the edge of the shape. As shown, the terms with high applicability for one chemical tended to also have high applicability for the other

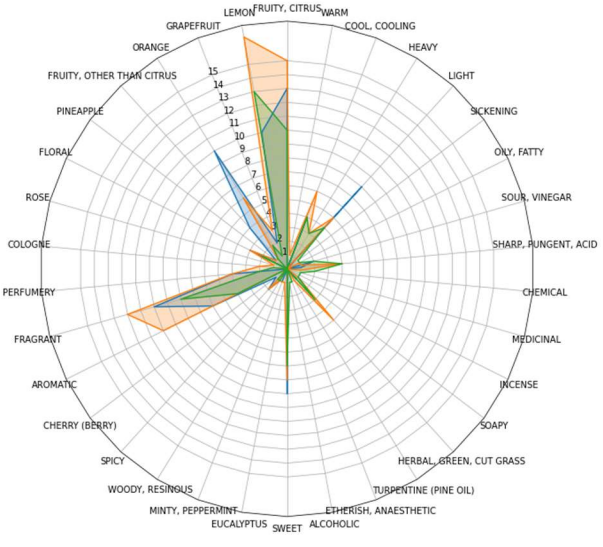

Figure 5. Spider graph for the citrus family

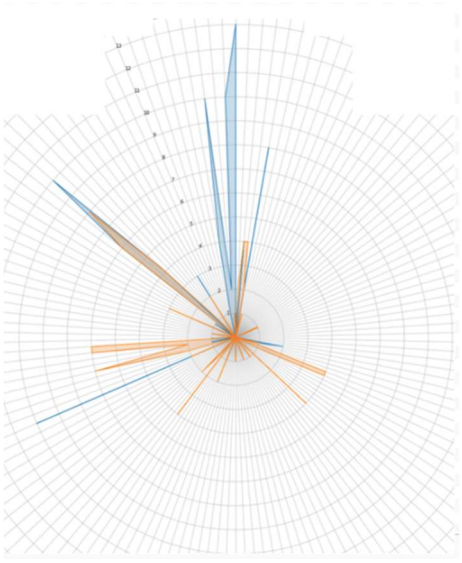

Figure 6. Spider graph of Limonene and Patchouli Oil

chemicals within this family. This trend carried through each family.

In contrast, the plots for chemicals from different families tended to have spikes for one chemical with none for the other, as shown in **Fig. 6**. In this plot, the blue corresponds to Limonene, a citrus scent, and the orange is patchouli oil, a woody scent. There is very little overlap between two scents of different families. The only true crossovers are "Fragrant" and "Aromatic," which are not family specific for these chemicals. More data visualizations are shown in Appendix A, including an LDA (Linear Discriminate Analysis) plot.

The nine chemicals chosen for this project are shown in Table 1. Several trials were performed to determine the retention time and detection limit. Once a detection limit was found, the sample bag was cleaned using lab air, and the GC was run with filtered air samples. To make sure they were accurate for each chemical, the detection limits were then tested again, using the clean bags.

In **Fig. 7**, multiple chromatograms are shown for different concentrations of acetone. The detection limits for this column were chosen to be the concentatrations of substances at which the peak was 100 units (in the raw ADC units) above the baseline to account for noise. Each color corresponds to a different concentration. As shown, the correlation between concentration and peak height is non-linear, as the jump in concentration is the same for each. The peaks all occur at the same time, at the retention time for acetone. The smallest curve is at acetone's detection limit, which is 48 parts per million (ppm). So, for every million parts of that sample, there were 48 acetone molecules. For this visual, the raw ADC has been converted to Volts.

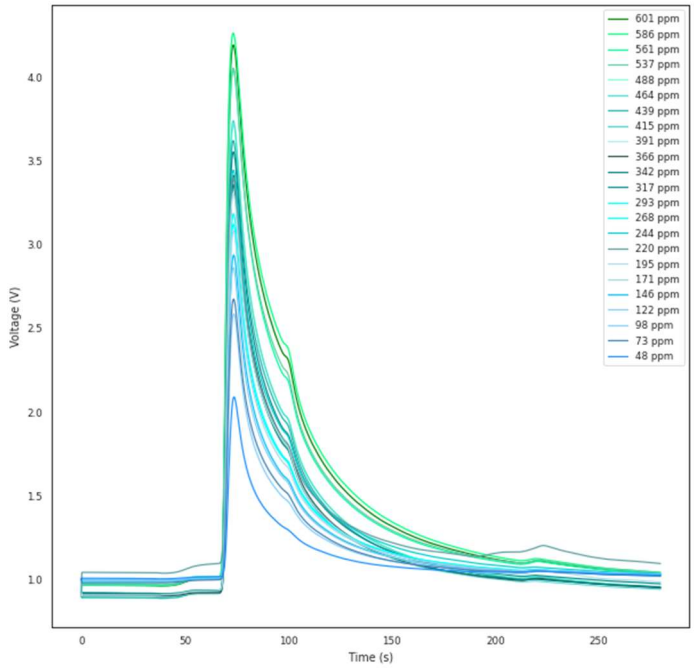

Figure 3. Peaks corresponding to different concentrations of acetone

Table 1. Chemicals purchased from the Atlas and their families

Chemical	Retention Time (s)	Detection Limit	Family
1-Hexanal	1486	61 ppt	Green
Diphenyl Oxide	1490	.28 ppt	Green
Haxanol	270	13 ppt	Green
(-) Menthol	550	10 ppm	Aromatic
Eucalyptol	541	.25 ppt	Aromatic
Methyl Salicylate	584	9 ppm	Aromatic
Acetophenone	52	2089 ppb	Floral
Hydroxy Citronellol	546	328 ppt	Floral
Phenyl Ethanol	227	13 ppb	Floral

Due to the limited sample size, the correlation between family and retention time might be unreliable. Also, some chemicals have multiple peaks, so the primary peak is the one used in the table. Graphically, as seen in **Fig. 8** and **9**, there seems to be a correlation between the retention time and the scent families of molecules. To substantiate this claim, there would need to be further samples run, and more chemicals from each scent family for comparison's sake.

In **Fig. 5** and **6**, the horizontal lines show the detection limit of the sensor. So, any peak below that line is considered noise, and not part of the chemical specific portion of the chromatogram. Experimental data for all other chemicals can be found in Appendix B.

The gathered chromatograms were used as the basis for a machine learning algorithm to determine scent.

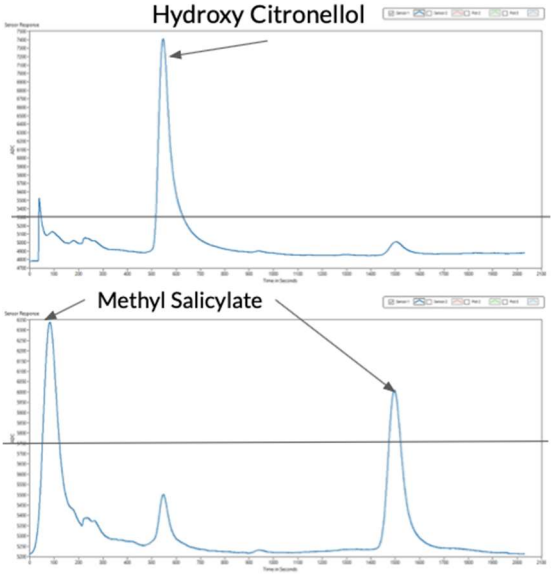

Figure 8. Experimental data for Hydroxy Citronellol and Methyl Salicylate

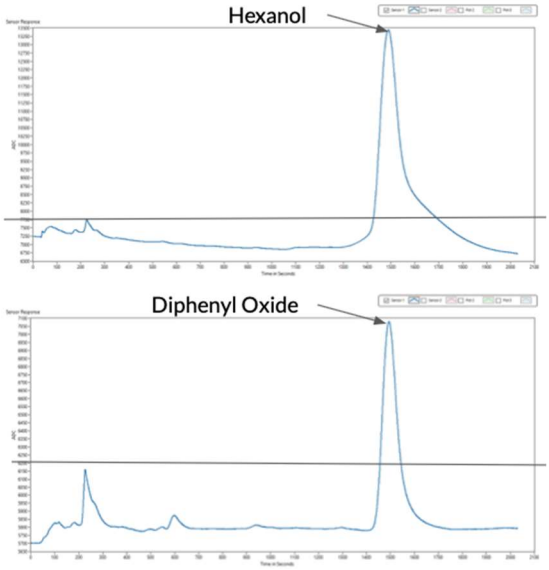

Figure 9. Experimental data for Hexanol and Diphenyl Oxide

The chromatograms were correlated to their scent family, and the algorithm was trained on that correlation. The algorithm was fed plotted chromatograms, and these served as the basis of the algorithm. The algorithm took in the whole signal, including any noise. Due to the difference in run times for samples, some signals had to be lengthened for the algorithm to train on the experimental data. Otherwise, the algorithm would take a chromatogram gathered over 250 seconds, and directly compare it to a signal taken for 2500 seconds, even though the signal has a factor of ten times the number of data points. Iterations of the algorithm have an average accuracy of 85% in correlating the correct scent family to a chromatogram from the three families used for data collection.

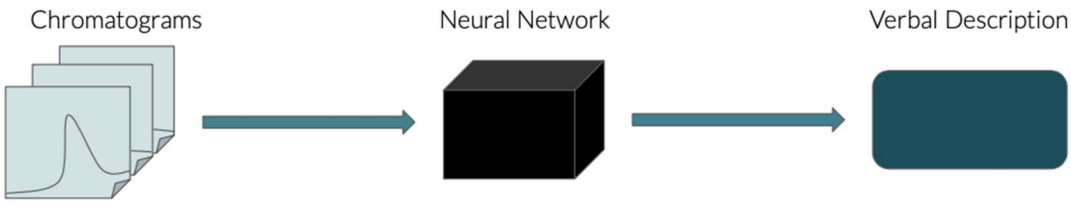

Figure 4. Diagram showning Machine Learning Architecture

CONCLUSION AND FUTURE WORK

The project is still ongoing, as the algorithm is only 85% accurate. If more chromatograms were gathered and more chemicals were used from the Atlas, the algorithm would be more accurrate. The time for running each sample could also have been more streamlined, although the amount of data collected was limited by the time constraints of the semester and the GC device being used. Some samples required the machine to collect data for half an hour, but between the cool down time and the actual run time, those samples took almost an hour each. Despite this, there does seem to be some correlation between retention time and human scent perception, although this would require further testing to state.

This project can also be expanded in the future in several ways. For example, this couls have connections in agriculture. For ten weeks in the summer of 2021, I worked at Alabama A&M's agricultural research station where we treated industrial hemp with low temperature plasma (LTP) to test the affects on growth. Industrial hemp is *Cannabis sativa* within a certain threshold of THC (Tetrahydrocannabinol), measured in relation to CBD (Cannabidiol). When farmers grow industrial hemp, they must make sure that the percentage of THC in their crops is not over the threshold. In Alabama, this threshold was 0.3%. Once a crop surpasses this limit, the entire batch is considered a controlled substance and is destroyed by the state as such. Currently, the only way to test THC content in the plants is through a destructive process. If an e-Nose could be fashioned to determine THC

content through non-destructive means, many harvests can be planned when the crops approach the limit so an entire season's work is not lost.

Throughout this process, I have learned a lot about both experimental physics and myself. I have discovered that I like working in a lab, and doing things with my hands, rather than only theory. I discovered the importance of coding in an experimental project, and that I was not made to directly interpret 146 dimensions. Also, I found that I prefer a lab with windows, despite the beauty of API. Moreover, I found that I love research in the field, and this process has led me to pursue a Ph.D. in physics.

The goal of the project was to connect GC to scent, and to generate verbal description of scents without the use of a human nose. That has been accomplished in a rudimentary sense. I hope that, at some point, someone takes this to the point where we can generate more descriptive verbal descriptors for scents using this process.

REFERENCES

[1] Dravnieks, A., & ASTM Committee E-18 on Sensory Evaluation of Materials and Products. (1985). Atlas of odor character profiles. Philadelphia, PA: ASTM.

[2] Falk, M., Nilsson, E. J., Cirovic, S., Tudosoiu, B., & Shleev, S. (2021). Wearable Electronic Tongue for Non-Invasive Assessment of Human Sweat. *Sensors (Basel, Switzerland)*, *21*(21), 7311. https://doi.org/10.3390/s21217311

[3] Karametsi, K., Kokkinidou, S., Ronningen, I., & Peterson, D. G. (2014). Identification of bitter peptides in aged cheddar cheese. *Journal of Agricultural and Food Chemistry*, *62*(32), 8034–8041. https://doi.org/10.1021/jf5020654

[4] Keller, Andreas, et. al. (2017). Predicting human olfactory perception from chemical features of odor molecules. Science. 355. eaal2014. 10.1126/science.aal2014.

[5] Rastogi, Suresh. (1995). Analysis of fragrances in cosmetics by gas chromatography-mass spectrometry. Journal of High Resolution Chromatography. 18. 653 - 658. 10.1002/jhrc.1240181008.

[6] Roa, Mary & Fernandez, Proceso. (2018). Development of an Electronic Nose for Olfactory System Modelling using Artificial Neural Network. Transactions on Machine Learning and Artificial Intelligence. 6. 10.14738/tmlai.64.4985.

APPENDIX A: ATLAS VISUALIZATIONS

Plots were generated in Python to analyze the raw data from the Atlas. Two-Dimensional Slices, Parallel Coordinates, Principal Component Analysis (PCA), Linear Discriminate Analysis (LDA), t-distributed Stochastic Neighbor Embedding (t-SNE), and spider (radar) plots were used to visualize the raw data.

Two-Dimensional slices take only two of the 146 dimensions for each point. For example, we plotted Vanilla v. Lemon, but this would work for any of the descriptors in the atlas. There is a point for each chemical, color coded to its scent family.

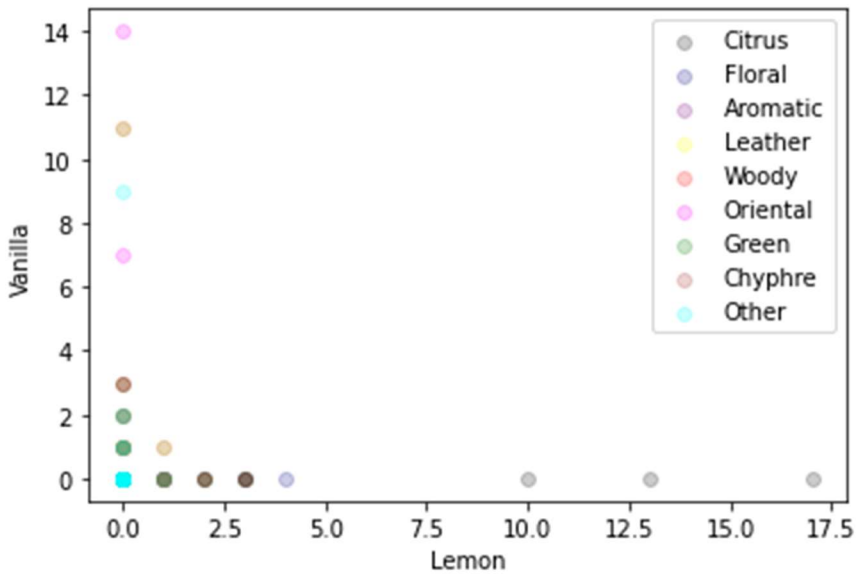

Figure 5. 2-D slice with Lemon and Vanilla descriptors

Parallel Coordinates has a point on the x-axis for each descriptor, and the applicabilities on the y-axis. Then, there is a shape plotted for each chemical. These are scatterplots where the points are connected linearly.

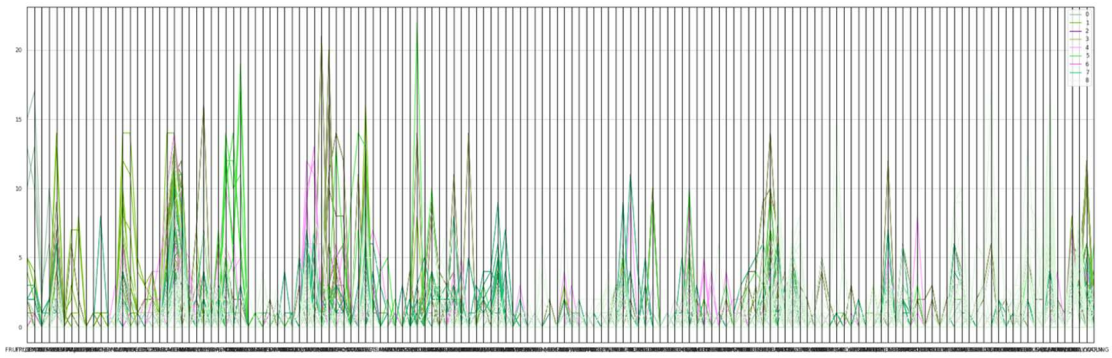

Figure 6. LDA plot

PCA is used to reduce data dimensionality. PCA can be plotted in two or three dimensions. PCA plots clusters of like points, color coded for their group. The downside to PCA is that the distance between these groups is arbitrary. The clusters, however, do show correlation.

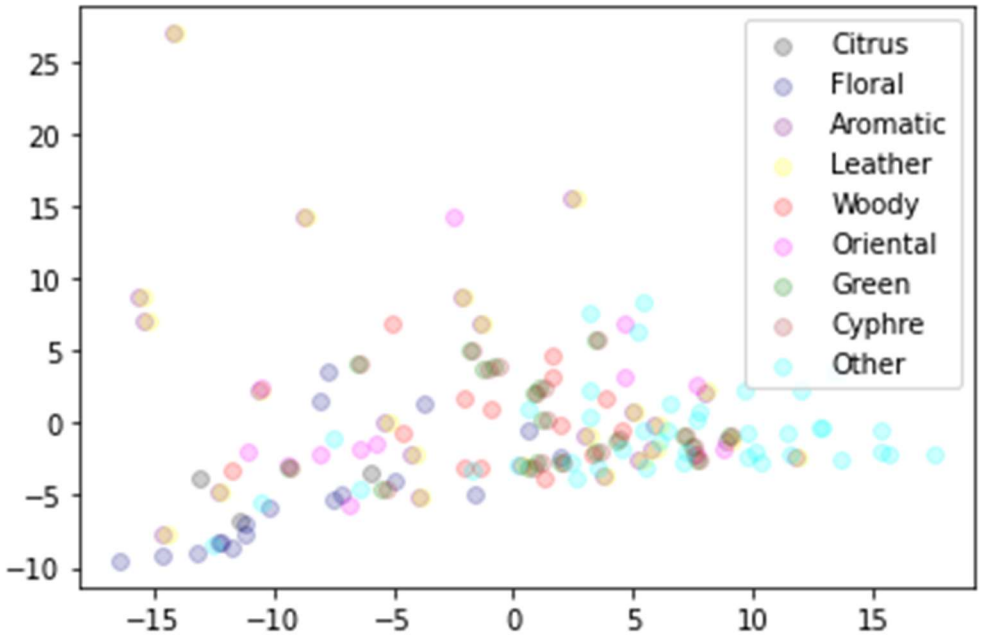

Figure 7. PCA plot

LDA is like PCA, but the tool looks to separate the classes, not just the data points. The plots still look like clusters of points color coded by group identifier.

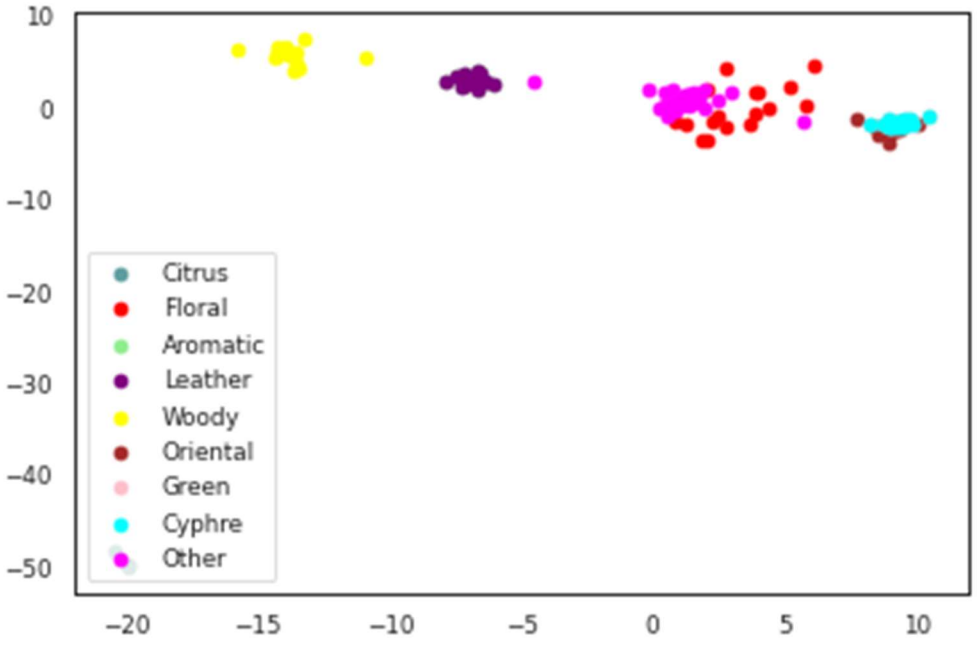

Figure 8. t-SNE plot

Also, like PCA, t-SNE is a data reduction tool that can be used in two or three dimensions. It also appears to be clusters of data points color coded by a group identifier. Unlike PCA, the distance between clusters is not meaningless.

Spider plots are the most dissimilar, as explained in the introduction.

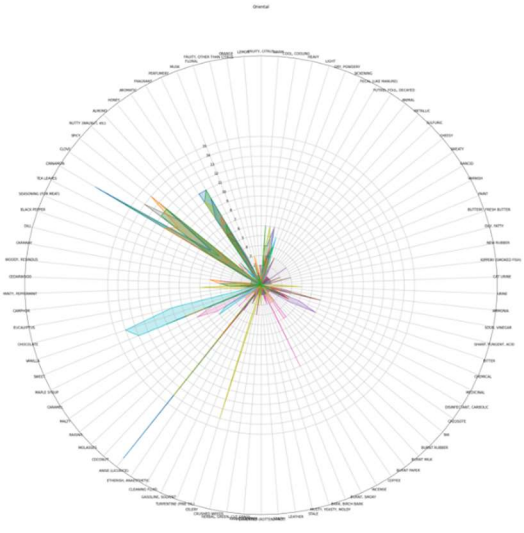

Figure 12. Oriental spider plot

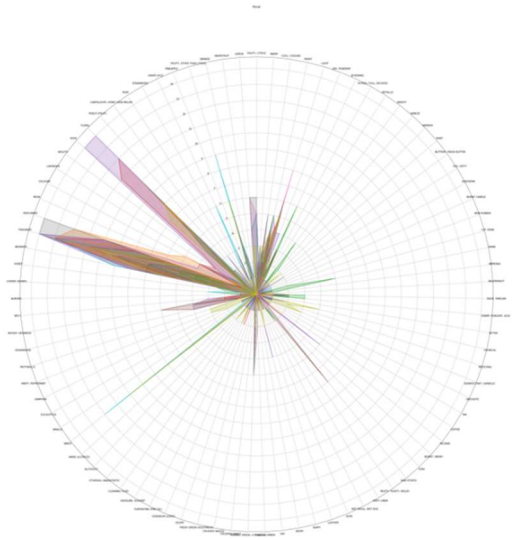

Figure 11. Floral spider plot

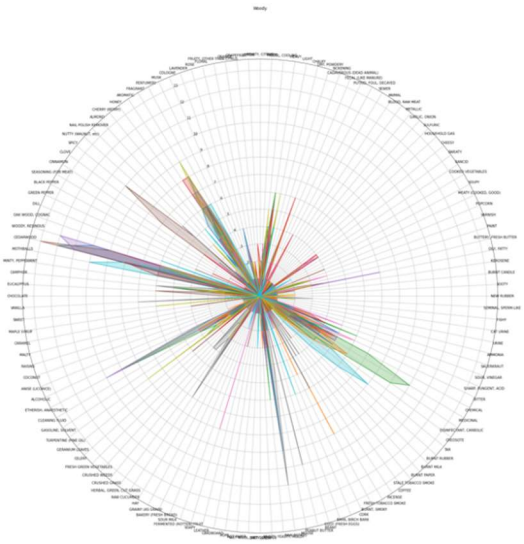

Figure 10. Woody spider plot

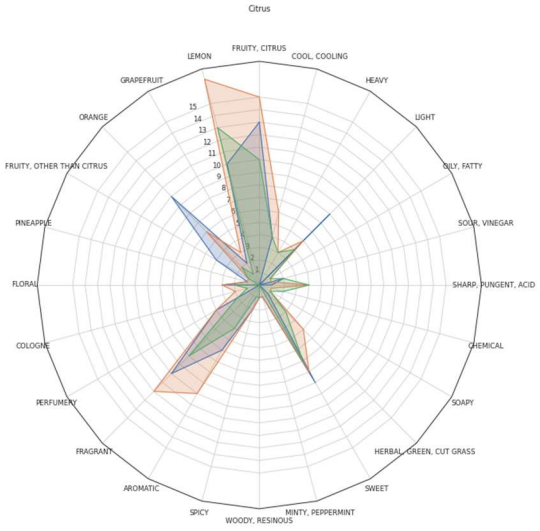

Figure 9. Citrus spider plot

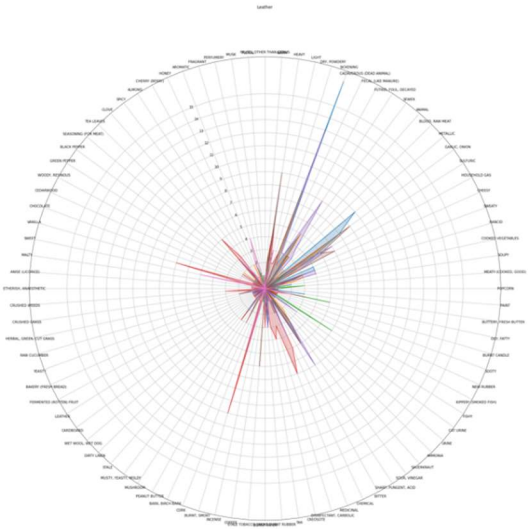

Figure 19. Leather spider plot

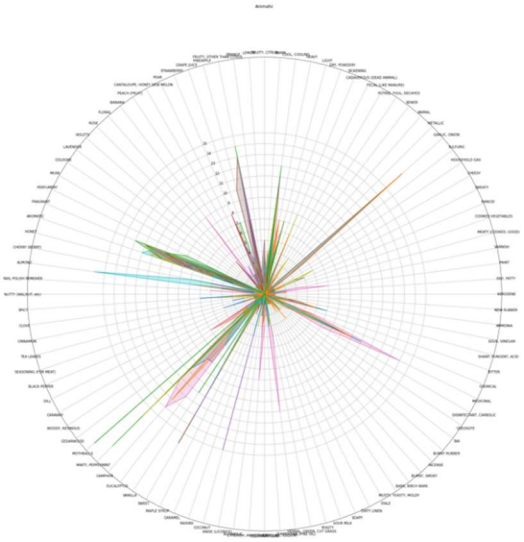

Figure 13 Aromatic spider plot

APPENDIX B: EXPERIMENTAL CHROMATOGRAMS

Experimental GC Data was gathered for all purchased chemicals. Included in this appendix are nine such chromatograms, one for each chemical.

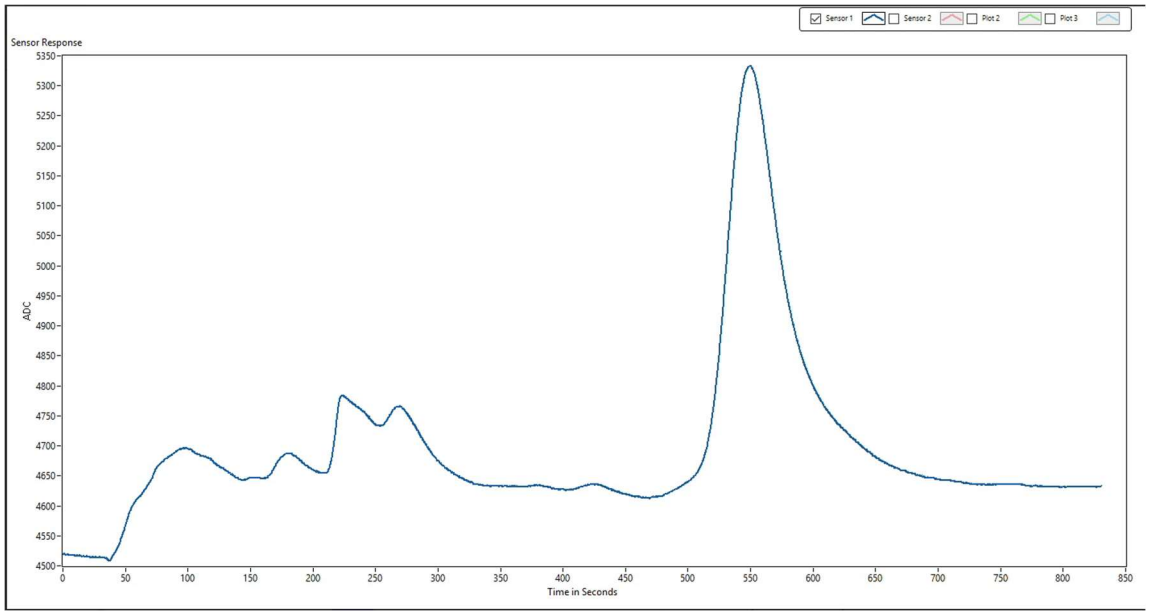

Figure 14. Menthol(-) chromatogram

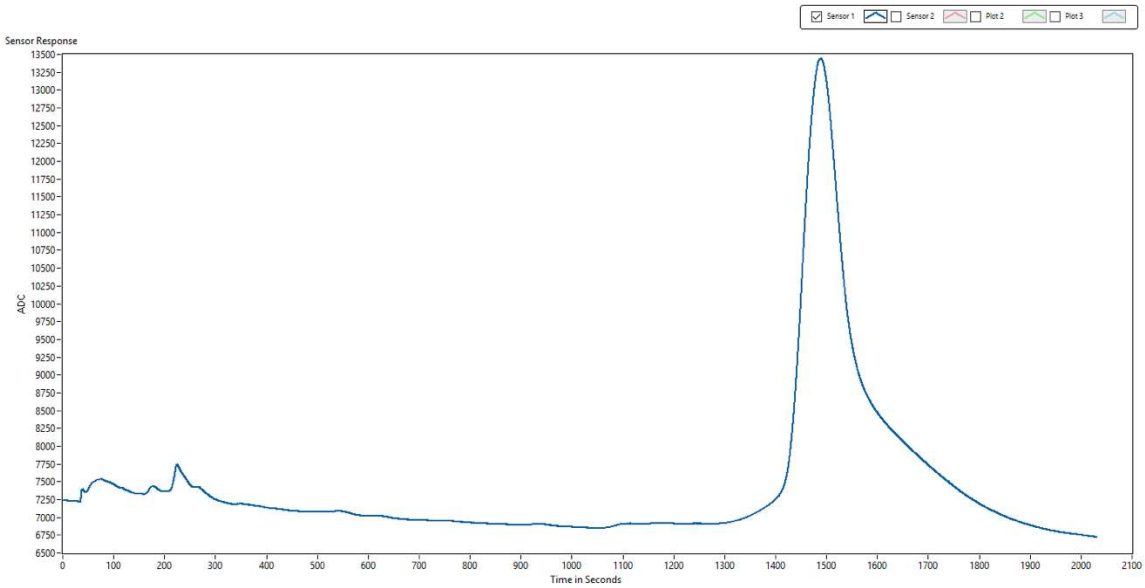

Figure 15. 1-Hexanol chromatogram

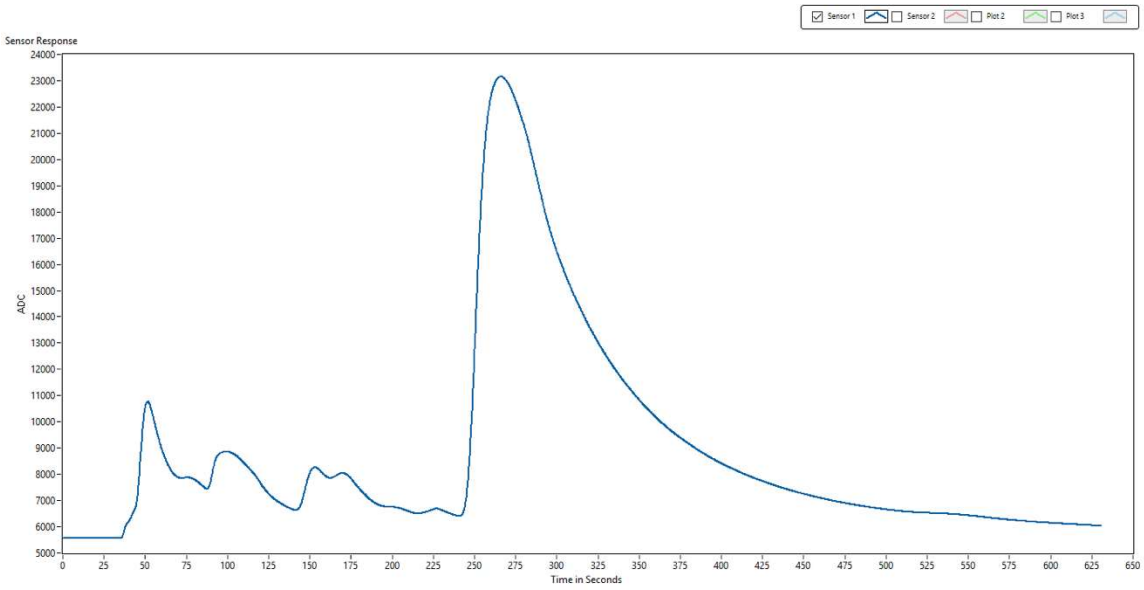

Figure 16. Haxanal chromatogram

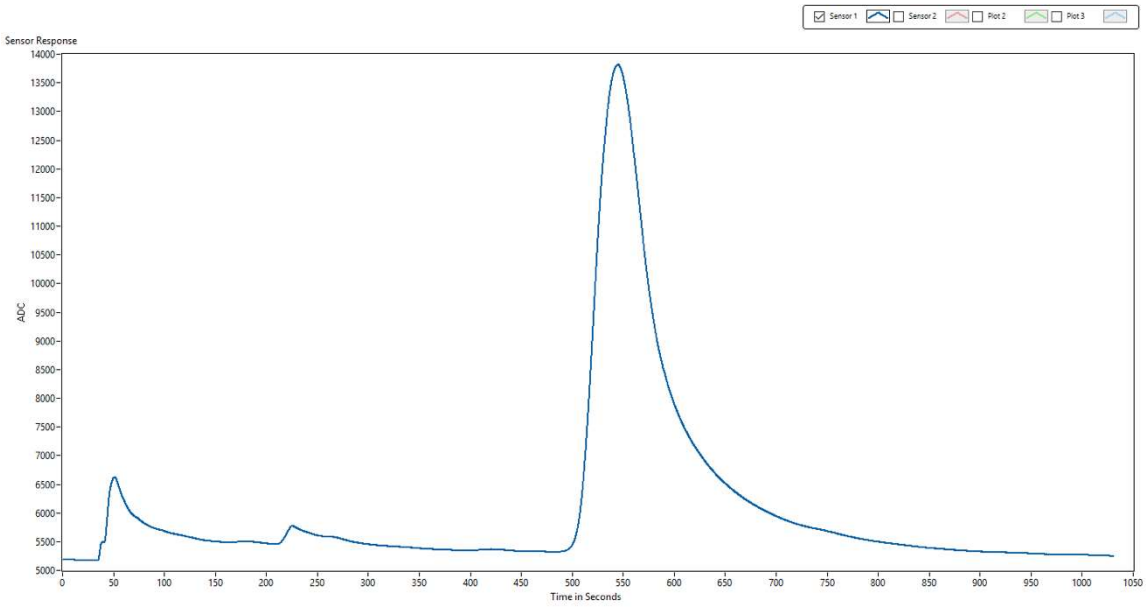

Figure 17. Acetophenone chromatogram

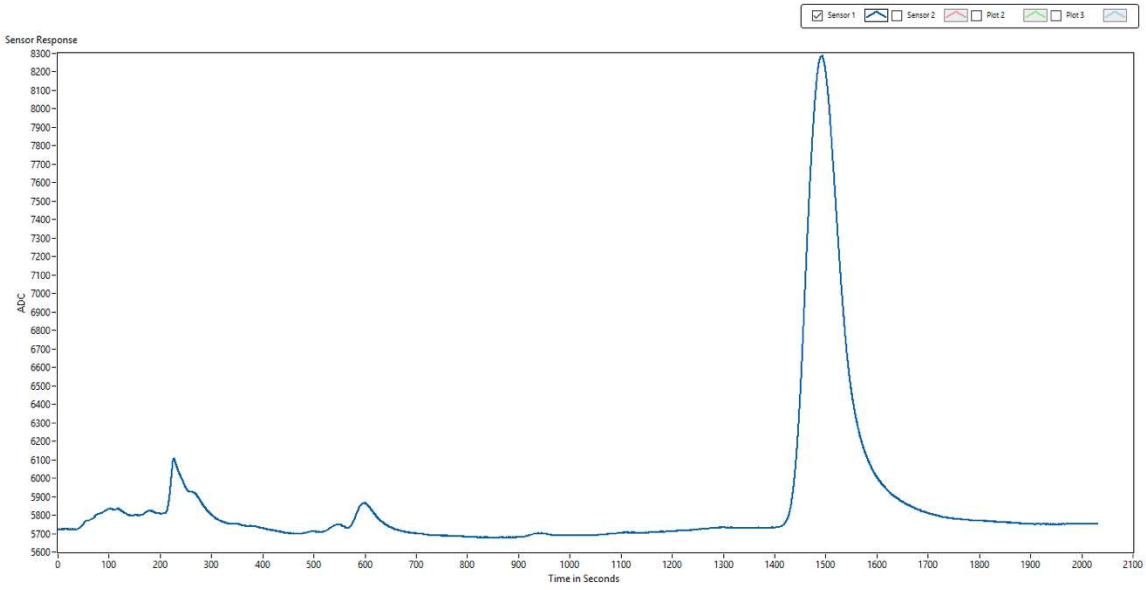

Figure 18. Diphenyl Ether chromatogram

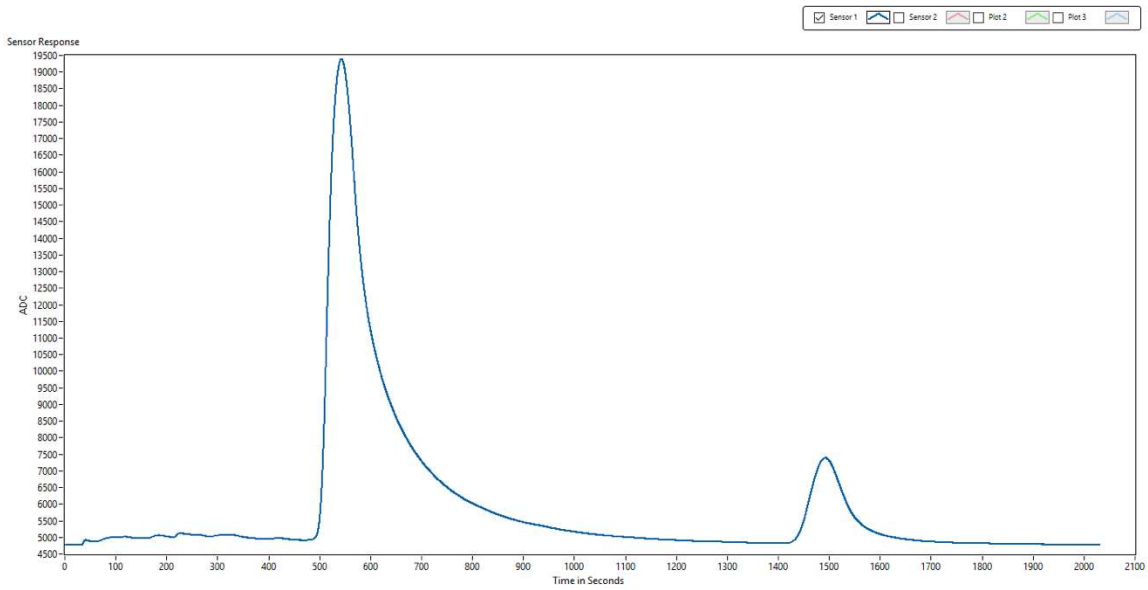

Figure 19. Eucalyptol chromatogram

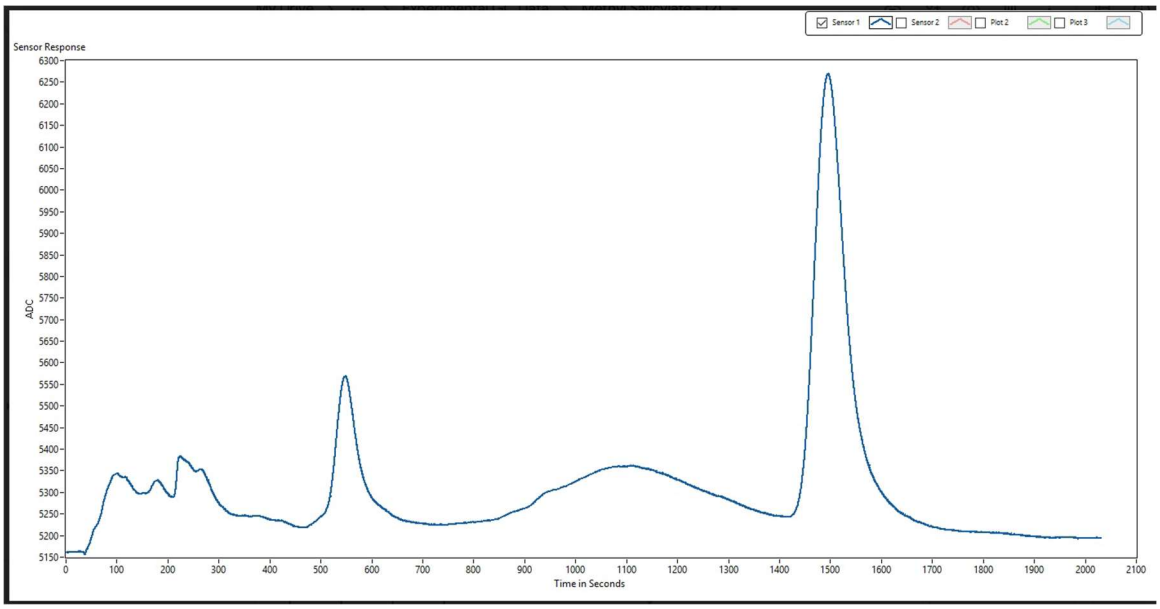

Figure 20. Methyl Salicylate chromatogram

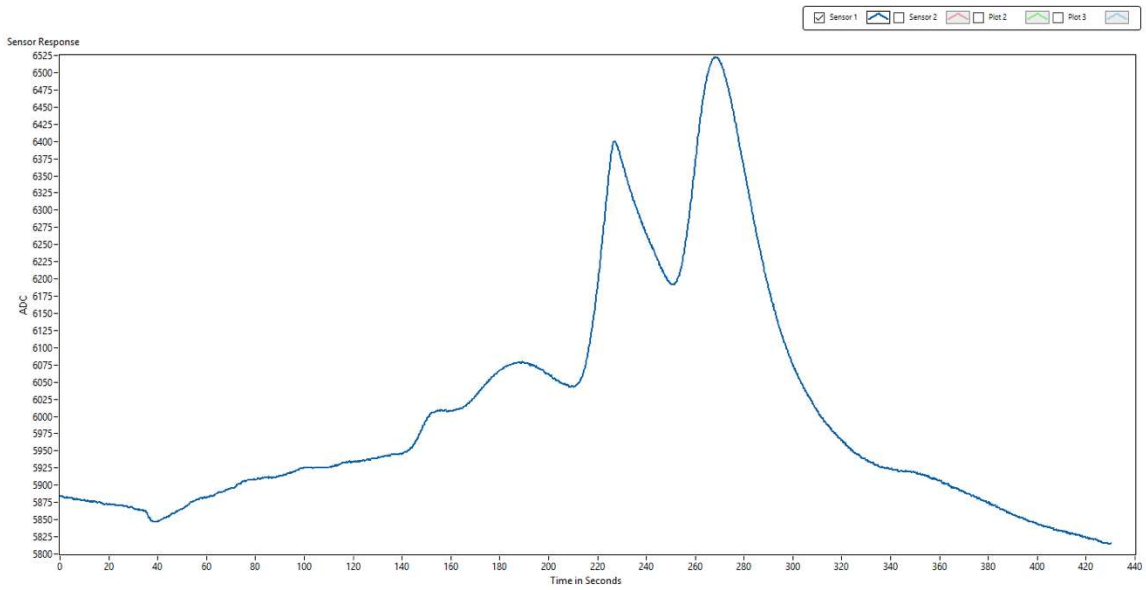

Figure 21. Phenyl Ethanol chromatogram