

Exploring Complex Chemical Environments of Ocean Worlds
Through Machine Learning

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy
in the Graduate School of The Ohio State University

By

Nicole Marie North

Graduate Program in Chemistry

The Ohio State University

2024

Thesis Committee

Dr. Heather Allen, Advisor

Dr. Abraham Badu-Tawiah

Dr. Zachary Schultz

Dr. Morgan Cable, NASA JPL

Copyrighted by
Nicole Marie North
2024

Abstract

Ocean worlds are defined as bodies throughout the solar system that are known or theorized to contain a large amount of liquid water. Most notably, Earth is an ocean world. Additionally, there are many moons and exoplanets that are also considered to be ocean worlds including Enceladus, a moon of Saturn, among others. Understanding the complex chemical environment of these oceans is important to elucidating the potential chemical reactions occurring on these different interstellar bodies. This includes our ability to find life or lifelike processes on other planets and moons. Performing measurements during planetary missions is incredibly challenging and a great deal of work has been done to improve instrumentation and analysis to increase the impact of mission returned data. Implementing machine learning techniques to the planetary instrumentation data pipeline is another promising way to help further improve our understanding of these systems.

In this dissertation machine learning algorithms are used in a variety of methods to answer both categorical and numerical questions to this end. Classification type machine learning questions have been used to classify the presence of functional groups in analyte molecules using only an electron ionization mass spectrum through logistic regression. Regression based machine learning has been used to develop methods for identifying concentrations of organic classes of compounds (saccharides, fatty acids, and amino acids) in marine samples. This has been accomplished both with a single analyte and with multiple

analytes. Support vector regression proved to be the most effective and accurate at identifying the concentrations of these compounds in a complex chemical matrix.

Beyond only training accurate models, this research also utilizes and develops methodology to further analyze the embedded reasoning behind the model's assignments. These methods included feature analysis, which involved evaluating the associated weights for each of the features (the x axis of spectroscopic and spectrometric data) to determine which features were the most important for the model's assignments. Through this method it was possible to identify trends in how the model was analyzing the data that were consistent with how a chemist would look at the data. Sample dropout as a final validation was also utilized to increase confidence in some models that utilize field samples ensuring that no data leakage was leading to inflated accuracy values for the models. These methods led to more powerful and applicable models by ensuring that the models are based in reality and our understanding of chemistry. As machine learning is being implemented in more and more areas of science it is critical to scrutinize these models to ensure that they are giving accurate answers for the right reasons.

Much work is yet to be done in this area to fully identify the total diversity of organics in these complex aqueous environments, but machine learning has proved to be a powerful tool in this pursuit of understanding. The described research herein utilizes powerful machine learning methods to understand the chemical composition of complex chemical samples quickly and accurately, providing a novel perspective for the analysis of ocean worlds.

Dedication

To my wonderful friends who made this possible: those with whom I can laugh, cry,
drink, fight, learn, love, and grow.

Acknowledgments

For many of life's greatest and most worthwhile pursuits it takes a village. I have had an incredible scientific support system throughout my time completing this work. First off, I would like to thank my advisor, Dr Heather Allen. Thank you for giving me the opportunity to chase my big ideas and for your guidance and support when things were difficult. I am very proud of the work that we accomplished. In addition, I had amazing support and encouragement from Dr Morgan Cable, from NASA JPL. Thank you for believing in me after only knowing me for a few days and teaching me how to pass on the kindness and support that I have been given in my academic journey. My committee members have been powerful assets to help push me and my science while also providing support to help these big ideas come to fruition. Dr Nicole Karn has also been an amazing sounding board for me, providing cheers and laughter when things are going well and support and advice when they aren't. My lab members have also been critical in helping me through day-to-day troubles providing a listening ear, a laugh, and advice. An extra special thanks to Dr Abigail Enders and Jessica Clark for keeping me sane through their friendship, and support, and their kind reminders to know when to say yes and more importantly, when to say no.

Outside of my immediate scientific community, I have many others to thank. First, I would like to thank my husband, Mason North. His support and willingness to carry us through where I couldn't, made all this possible. Here is to the next chapter of our lives being even better than this one! I would also like to thank my parents, Mary and Larry Bishop and my brother, Ken Bishop. Thank you for listening to my ramblings and plights

especially when you didn't understand what I was talking about. The caring and responsive ear was what I needed the most. I also want to thank my cats, Binx and Zanzibar North. Try as they might their personal edits did not make it into the final version of this document, but I do deeply appreciate their efforts. I also want to thank Joshua Prybil and Meredith Varnecky. Thank you for all your advice and support and the late nights trying desperately to not die in one of the many online videogames and D&D campaigns that we have sampled over the years.

I would also like to give a shoutout to three communities that have provided Mason and I so much support since moving out to Columbus in 2019. First, the Forge Tavern, an amazing boardgame bar in Columbus. We have made lifelong friends here and had some amazing whiskies and played so many rounds of trivia! Next, Avery Lodge in Hillard. You welcomed us into your community and have always been willing to help wherever you could. We are forever grateful for that kindness. Finally, I want to thank my fight side family at Endeavor Defense and Fitness. Thank you for teaching me that fighting is a surefire way to make friends! I have learned so much about myself through training with you all and it has instilled in me a confidence that has greatly impacted all areas of my life.

Thank you all for your support; you have truly enriched my life!

Vita

Nicole (née Bishop) North was born on October 23rd, 1995, in Des Moines, Iowa. After graduating from North Polk High School in Alleman, IA she attended the University of Northern Iowa in Cedar Falls, IA. She received her Bachelor of Science in Chemistry in the spring of 2019. Directly after completing her undergraduate degree Nicole began her PhD in the fall of 2019 at the Ohio State University in Columbus, OH. For the academic years of 2019-2020 and 2020-2021 Nicole was a teaching assistant in OSU's Department of Chemistry and Biochemistry where she assisted with general and analytical chemistry courses. In the summer of 2021 Nicole received a Future Investigator in NASA Earth and Space Science and Technology award which funded the remainder of her dissertation work.

Publications

- S. Andrejkovičová, A. C. McAdam, J. C. Stern, C. A. Knudson, R. Navarro-González, M. Millan, S. T. Wieman, J. A. Sebree, N. M. Bishop, E. B. Rampe, D. W. Ming, P. R. Mahaffy, NH₄-smectites as a potential source of N compounds (NO) in SAM analyses 49th Lunar and Planetary Science Conference 2018
- Moran, S. E.; Hörst, S. M.; Vuitton, V.; He, C.; Lewis, N. K.; Flandinet, L.; Moses, J. I.; North, N.; Orthous-Daunay, F.-R.; Sebree, J.; Wolters, C.; Kempton, E. M.-R. .; Marley, M. S.; Morley, C. V.; Valenti, J. A. Chemistry of Temperate Super-Earth and Mini-Neptune Atmospheric Hazes from Laboratory Experiments. *The Planetary Science Journal* 2020, 1 (1), 17. <https://doi.org/10.3847/psj/ab8eae>.
- Enders, A. A.; North, N. M.; Fensore, C. M.; Velez-Alvarez, J.; Allen, H. C. Functional Group Identification for FTIR Spectra Using Image-Based Machine Learning Models. *Analytical Chemistry* 2021, 93 (28), 9711–9718. <https://doi.org/10.1021/acs.analchem.1c00867>.
- Wang, L.; Morita, A.; North, N. M.; Baumler, S. M.; Springfield, E. W.; Allen, H. C. Identification of Ion Pairs in Aqueous NaCl and KCl Solutions in Combination with Raman Spectroscopy, Molecular Dynamics, and Quantum Chemical Calculations. *The Journal of Physical Chemistry B* 2023, 127 (7), 1618–1627. <https://doi.org/10.1021/acs.jpcc.2c07923>.

- North, N. M.; Enders, A. A.; Cable, M. L.; Allen, H. C. Array-Based Machine Learning for Functional Group Detection in Electron Ionization Mass Spectrometry. ACS Omega 2023, 8 (27), 24341–24350. <https://doi.org/10.1021/acsomega.3c01684>.
- Nicole M. North, Abigail A. Enders, Jessica B Clark, Heather C. Allen, Saccharide concentration prediction from proxy sea surface microlayer samples analyzed via infrared spectroscopy and quantitative machine learning, ACS Earth and Space **2024**
- Nicole North, Abigail Enders, Jessica Clark, Kezia, Effie, Morgan Cable, Heather Allen, Multi Analyte Concentration Analysis of Marine Samples Through Regression Based Machine Learning, ACS Earth and Space **2024 – Preprint Available**
- K. G. Hanley , Q. McKown, E. M. Cangi, C. Sands, N. M. North, P. M. Miklavcic, M. Bramble, J. M. Bretzfelder, B. D. Byron, J. Caggiano, J. T. Haber, S. J. Laham, D. Morrison-Fogel, K. A. Napier, R. F. Phillips , S. Ray, M. Sandford, P. Sinha, T. Hudson, J. E. C. Sully ,L. Lowes, The Vulcan Mission to Io: Lessons learned during the 2022 JPL Planetary Science Summer School, AAS Planetary Science **2024 – In Review**

Field of Study

Major Field: Chemistry

Table of Contents

Abstract	ii
Dedication	iv
Acknowledgments	v
Vita	vii
List of Tables	xii
List of Figures	xiii
List of Supplemental Tables	xvii
List of Abbreviations	xix
Chapter 1. Introduction	1
1.1 Motivation	1
1.2 Approach	2
1.3 Dissertation Highlights	3
Chapter 2. Pertinent Background Information	6
2.1 Chemical Systems	6
2.1.1 Terrestrial Ocean Chemistry	6
2.1.2 Enceladus as an Ocean World	7
2.2 Analytical Instrumentation	11
2.2.1 Raman Spectroscopy	11
2.2.2 Fourier Transform Infrared Spectroscopy	21
2.3 Machine Learning Approaches	24
2.3.1 K-Nearest-Neighbors	26
2.3.2 Logistic Regression and Support Vector Machines	28
2.3.3 Decision Trees and Associated Ensemble Algorithms	30
2.3.4 Neural Networks	34
Chapter 3. Array Based Machine Learning for Functional Group Detection in Electron Ionization Mass Spectrometry	36
3.1 Introduction	37

3.2 Methods.....	40
3.2.1 Spectral Preprocessing and Machine Learning Parameter Selection.....	40
3.2.2 Supplemental Experimental Data Collection.....	44
3.2.3 Model Training and Testing.....	44
3.3 Results and Discussion	46
3.3.1 Acquiring the Dataset	46
3.3.2 Comparing Convolutional Neural Networks and Logistic Regression Feasibility.....	46
3.3.3 Logistic Regression’s Ability to Manage Specific Functional Group Classifications	50
3.3.4 Identifying Mass Peaks that Guide Model Assignments	52
3.3.5 Effects of Mass Range on Model Accuracy	56
3.3.6 Specific Examples of the Applications of this Approach	57
3.4 Conclusion	60
Chapter 4. Saccharide concentration prediction from proxy sea surface microlayer samples analyzed via infrared spectroscopy and quantitative machine learning.....	62
4.1 Introduction.....	62
4.2 Methods.....	67
4.3 Results and Discussion	73
4.4 Conclusions.....	81
Chapter 5. Multi Analyte Concentration Analysis of Marine Samples Through Regression Based Machine Learning	82
5.1 Introduction.....	83
5.2 Methods.....	85
5.3 Results and Discussion	92
5.4 Conclusion	106
Bibliography	108
Appendix A. Supplemental Information for Chapter 3	126
Appendix B. Supplemental Information for Chapter 4.....	160
Appendix C. Supplemental Information for Chapter 5.....	167
Appendix D. Supplemental Information for 532 nm Polarized Raman.....	173

List of Tables

Table 1. Reported Mass Fragments from in situ measurements of the Enceladus plume.	10
Table 2. Mass values of the top 5 most impactful positively correlated peaks for each of the functional group generalized models. These were determined by comparing the testing accuracies of the model when it had access to all 300 mass units to when that mass unit of interest was removed. The % effect shown in the right-most column is negative because when those masses were removed the model experienced a reduction in the final testing accuracy. The mass values for the nitrogen containing model are all odd mass values and the mass values for the oxygen containing and the aromatic containing spectra are all even suggesting the utilization of the odd nitrogen rule without explicit training on that detail.	55
Table 3. Results of Selected Models on the Ability to Correctly Identify the NIST Spectra of Tryptophan.	58
Table 4. Results of the models on the ability to correctly identify experimental spectra of Limonene. The experimental spectra were preprocessed in the same way as the NIST data used for training. The results for all the models on limonene, pyridine, and 2 furan methanol are presented in the SI (Table S10).	60
Table 6. Concentrations of all species in the lab-made ocean proxy samples for evaluation of model accuracy on more chemically diverse conditions.	69
Table 7. Marine samples associated with the UM and SM datasets. Concentrations of glycine, butyric acid, and glucose were calculated through mass spectrometry and will be used as the “true” values of concentration for these samples. Histidine concentrations were all beneath the LOQ for the mass spectral method.	97
Table 8. Highest performing models for each analyte compound.	102
Table 9. Effects of dropout sample test on highest performing models for each analyte compound.	106

List of Figures

Figure 1. Simplified diagram of proposed physical and chemical processes occurring on Enceladus.	8
Figure 2. Overview schematic for Raman spectroscopy.	11
Figure 3. Simplified vibrational diagram of Raman processes.	12
Figure 4. Diagrams depicting the theoretical (A) and practical (B) setup and implementation of a laser.	14
Figure 5. Diagram of collection of polarized Raman signal from excitation source to spectrograph.	17
Figure 6. Theoretical Czerny Turner Spectrograph (II) and Schmidt-Cserny-Turner Spectrograph (I) layouts of the described spectrograph. Similar components have been labeled A-E.	18
Figure 7. Diffraction of light with a blazed grating.	19
Figure 8. Image of Raman CCD (I). Summed regions of interest (ROI) converted into Raman spectra (II).	20
Figure 9. Energy diagram of the absorbance of IR light generating an excited vibrational state.	21
Figure 10. Path diagram of an attenuated total reflection (ATR) Fourier transform infrared (FTIR) spectrometer.	22
Figure 11. Evanescent wave moving through an attenuated total reflection (ATR) crystal and sampling a sample droplet through its contact with the ATR crystal.	24
Figure 12. Summary of K-Nearest-Neighbor classification (I) and regression (II) approaches.	27
Figure 13. Summary of logistic regression classification approach.	29
Figure 14. Summary of support vector machine kernel transformation for the application of hyperplanes.	30
Figure 15. Generalized decision tree schematic.	32

Figure 16. Generalized scheme for random forest models.	33
Figure 17. Generalized scheme for gradient boosting.	34
Figure 18. Generalized scheme for artificial neural networks.	35
Figure 19. Distributions of available mass spectra from NIST included in this study are presented here. (A) Shows the generalized functional group classifications. (B) Shows the specific functional groups. Aromatic is listed as a specific functional group to help correlate the relative distribution between the generalized models and the functional group specific ones.	41
Figure 20. Histogram depicting the number of unique functional groups (duplicate functional groups within a molecule are not counted) present in each molecule from the NIST database. The largest distribution is molecules that contain 3 unique functional groups. Because most molecules contain multiple functional groups, they can be used to represent the positive case for multiple functional group models.	43
Figure 21. The results of the training and testing for four specific functional groups and the three functional group classifications are shown above. (A and B) Show the final training accuracy, accuracy of identifying the training portion of the data after the final training step has passed for both the functional group specific and functional group generalized models, respectively. For example, these plots would suggest that the CNN based approach should be better at correctly identifying the Aldehydes and the Ketones and that the LR based approach should have an edge on the Nitro group and the Alkyl Aldehydes. This, however, does not tell the full story. (C and D) The final test accuracies for the functional group specific and functional group generalized models respectively. The testing accuracy of the models is the accuracy of the models when presented with new previously unseen data shows that a high training accuracy does not correlate necessarily with a high final testing accuracy. The final training and testing scores for each of the functional group's models are presented in the SI (Table S3/S4 for specific functional groups and S5/S6 for the generalized functional groups).....	48
Figure 22. Scatter plot depicting all the final training and testing accuracies of each of the 20 different models. These final accuracies are highly variable with respect to the functional group that they are to be classifying.	51
Figure 23. Model coefficients for each of the different trained models as a function of mass fragment. (A) Depicts the coefficients for the generalized functional group models and (B) does the same with the specific functional group models. All the coefficient plots for the individual models are presented in the SI.	53
Figure 24. Scatter plots depict the effect of decreasing the utilized mass range (A) from 300 mass units to 100 mass units and increasing the utilized mass range (B) from 300	

mass units to 500 mass units on the final testing accuracy of the models. The presence of points that are positive on the x axis (shaded in green, right-most box) show a net benefit in accuracy whereas a negative x value (shaded in red, left-most box) indicates a worsening accuracy..... 57

Figure 25. Schematic flow chart of data collection process to the ML pipeline. 68

Figure 26. Heat map of the ATR-FTIR dataset as sorted by the concentration of glucose (0 – 1 M). The band of intensity growing in between 1100 and 1000 cm^{-1} corresponds to the increasing C-O stretching within the IR fingerprint region from the increased concentration of glucose. We do not see a strong spectral signature for the ESA relative to that of glucose also in solution (0 – 5 mg/mL) where we would expect the amide bands to exist between 1700 and 1500 cm^{-1} 74

Figure 27. Molecular structures of both glucose (left) and sucrose (right). Both saccharides contain similar vibrational bonds and vibrational environments in regions of the structure. The simplified proxy (SP) dataset contains only glucose and egg serum albumin whereas the ocean proxy (OP) dataset contains both glucose and sucrose in solution with egg serum albumin, bovine serum albumin, and 1-butanol..... 75

Figure 28. (a-f). Scatter plots depicting the accuracy of each of the utilized machine learning models on the simplified proxy (SP) dataset. The y-axis represents the difference between the model assigned and the actual concentrations of the testing dataset divided by the actual concentrations multiplied by 100% (circles) and the withheld validation dataset (triangles). The gradient boosted regression, multilayer perceptron, and support vector regression models do experience an increased error at low concentrations. 77

Figure 29. Bar graphs depicting the associated root mean squared error (RMSE) in each part of the training process for the simplified proxy (SP) dataset. All models have a final testing error of less than 0.07 M, but the MLR performed the best in this evaluation. The asterisk indicates that for the decision trees the training error was 0.00 M..... 78

Figure 30. Predicted concentration divided true concentration of combined saccharide for ocean proxy (OP) saccharide concentrations. Solid line at 100 represents 100% meaning that the predicted concentration equals the predicted concentration. The dotted lines represent +/- 20%. The darkest markers in each column represent the highest concentration of saccharide in OP (0.20 M) and the lightest represent the least concentrated (0.10 M). The models have varied levels of success at identifying samples that are far removed from the original training set. The highest performing models were GBR and SVR..... 80

Figure 31. Sample organization for model training datasets. The SL sample dataset (I) contains two sample arrays one in which there are anti-correlated concentrations (the species on opposite sides of the array have inverse calibration curves), and in the second

the calibration curves move in the same direction. The SM sample dataset (II) contains first a dilution series of the field samples to ensure that the calibration curves were done lower than the concentration of the UM samples and then an anti-correlated array of spikes. The row numbers show the solution array being used 1-5 is dilutions, 7-11 is anti-correlated calibration curves, and 12-16 is the correlated calibration curves. Not pictured: 6 represents the UM samples that are withheld as the final validation set for the trainings. 95

Figure 32. Diagram depicting the process of mapping the calibration curves to make unique combinations of concentrations for each spike-containing sample array. 98

Figure 33. Test stage root mean squared error (RMSE) values for each combination of ML approach and chemical species. 99

Figure 34. UM sample estimates from each ML approach on SM models (left - circles) and on SL models (right - triangles). Solid grey line denotes a difference between actual and predicted concentrations of 0. The dotted lines represent +/- 20% of the most concentrated marine sample for the given chemical species (glucose, glycine, and butyric acid). The SM models show more clustering within these boundaries than the SL models suggesting that the SM models were more accurate at identifying the concentrations within the UM samples. 100

Figure 35. Counted values out of 10 for the correctly quantified UM samples within 20% of the max true values in a single UM sample. These counts are separated by ML approach and chemical species. Importantly, the SM models perform higher than the SL models in nearly every case. SVR achieved the highest accuracies for all three analyte concentrations. 102

Figure 36. Marine sample analysis using dropout sample method. For each model training one column of samples was dropped (e.g., SM samples column A) then the model was evaluated using the UM sample associated with that marine sample (for column A: sample A6). The dropped sample results are in black or grey and the original analysis is left in the color associated with that ML approach in Figure 4. The accuracy of models is well maintained for glycine (I) and butyric acid (II). The largest loss in accuracy was in the measurement of glucose. This variance, due to it mostly being overestimates, may be associated with the presence of other saccharides in these samples that cannot be determined using the stated mass spectral method. 104

List of Supplemental Tables

Table S1. Structure of the 16 different functional groups explored during the specific functional group portion of our modeling experiments. In each of these structures the R groups stand for an undefined organic structure attaching the functional group to the rest of the molecule.....	127
Table S2. Functional groups that make up each of the different functional group classifications.....	128
Table S3. Final training accuracies for all models. These tests seek to show how these different architectures succeed at building specific models for a given functional group. The final training accuracy is defined as the accuracy of the identification of the training data set after the final step in training for the case of the image-based models, and after model convergence for the array-based models. A higher training accuracy suggests a closer fit to the data that makes up the training dataset. The training set looks at how the adjustment of the weights of the model during the training phase is working for new data that wasn't used to adjust the weights. The highest final training accuracy for a given functional group is highlighted. The image-based models were trained under multiple different training parameters, because of this the highest accuracy is reported and the average accuracy is within the parenthetical.	132
Table S4. Final testing accuracies for each of the functional group models are shown here. These tests seek to show how these different architectures succeed at building specific models for a given functional group. Final testing accuracies are defined as the final accuracy of identification of the test data, data that was withheld from the entire training and testing process. A higher accuracy at identifying the test data suggests an increased ability to generalize from the training and testing data and thus suggests an increased ability to identify novel data in the future. The highest final testing accuracy for a given functional group is highlighted. The image-based models were trained under multiple different training parameters, because of this the highest accuracy is reported and the average accuracy is within the parenthetical.....	135
Table S5. Final training accuracy for the models for the functional group classifications is presented here. The goal of these tests is to show how these architectures can create models to fit a more generalized classification of functional groups. The final training accuracy is defined as the accuracy of identification of the training dataset after the last training step in the case of the image- based approach, and after model convergence in the array-based approach. The highest accuracy is highlighted and in the case of the image-based approach multiple models were trained with different parameters and the average of all the different models for each functional group is shown in the parenthetical.....	136

Table S6. Final testing accuracies for each of the functional group models. These models seek to show how these different architectures succeed at building generalized models for a given functional group classification. Final testing accuracies are defined as the final accuracy of identification of the test data, data that was withheld from the entire training and testing process. A higher accuracy at identifying the test data suggests an increased ability to generalize from the training and testing data and thus suggests an increased ability to identify novel data in the future. The highest final testing accuracy for a given functional group is highlighted. The image-based models were trained under multiple different training parameters, because of this the highest accuracy is reported and the average accuracy is within the parenthetical. 138

Table S7. Top five most beneficial peaks for analyzing each of the different functional groups and functional group classifications. The beneficial peaks were defined as the peaks that caused the largest decrease in final training accuracy when removed from the feature set. 150

Table S8. Top five most hindbersome peaks for analyzing each of the different functional groups and functional group classifications. The hindbersome peaks were defined as the peaks that caused the largest increase in final training accuracy when removed from the feature set. 154

Table S9. Model results for looking at two different amino acids, tryptophan and histidine. The overall description of the molecules was approximately 75% accurate. . 158

Table S10. Model results for experimental data outside of the NIST dataset. The description of the functional groups for all three was approximately 80%. 159

Table S11. Results of the linear fits from Figure S3 on identifying the concentration of saccharide (glucose and sucrose) from the ocean proxy samples. These results show that an individual linear fit is insufficient to identifying the generalized concentration of saccharide in aqueous solution..... 164

Table S12. Numerical results from the error and fit analysis for each ML method. 165

Table S13. Numerical results from the estimates of the models for the lab proxy samples. 166

List of Abbreviations

ANN	Artificial Neural Network
ATR	Attenuated Total Reflectance
BSA	Bovine Serum Albumin
BW	Bulk Water
CART	Classification and Regression Tree
CDA	Cosmic Dust Analyzer
CNN	Convolutional Neural Network
DOC	Dissolved Organic Carbon
EI-MS	Electron Ionization Mass Spectrometry
ESA	Egg Serum Albumin
FTIR	Fourier Transform Infrared Spectroscopy
GBR	Gradient Boosted Regression
HGBR	Histogram Gradient Boosted Regression
INMS	Ion Neutral Mass Spectrometer
KNN	K Nearest Neighbors
LOQ	Limit of Quantification
LR	Logistic Regression
MCT	Mercury Cadmium Telluride
ML	Machine Learning
MLP	Multi-Layer Perceptron
MLR	Multivariate Linear Regression
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
NASA	National Aeronautics and Space Administration
NIST	National Institute of Standards and Technology
PSI-MS	Paper Spray Ionization Mass Spectrometry
RBF	Radial Bias Function
RF	Random Forest
RMSE	Root Mean Squared Error
SL	Spiked Lab Dataset
SM	Spiked Marine Dataset
SSML	Sea Surface Microlayer
SVM	Support Vector Machine
SVR	Support Vector Regressor
TLU	Threshold Logic Unit
TOC	Total Organic Carbon
UM	Unspiked Marine Dataset

Chapter 1. Introduction

1.1 Motivation

The research presented here within this dissertation seeks to advance the science in two unique areas. Firstly, it seeks to advance our understanding of ocean worlds, bodies throughout the solar system that are proposed to contain liquid water, through the lens of utilizing our understanding of terrestrial marine chemistry. Secondly, this research aims to advance our utilization of machine learning and advanced data analysis techniques in the chemical space. The motivation behind these themes and this research arises from the rapid expansion of each of these disciplines and the benefits that arise from this interdisciplinary knowledge transfer.

This research was funded by the National Aeronautics and Space Administration (NASA) Future Investigators in NASA Earth and Space Science and Technology (FINESST) program. The research was selected under the planetary science research program, aiming to fund projects that can improve our understanding of processes that occur throughout the solar system or for individual objects or systems. It also funds work that seeks to improve scientific return of past planetary missions. These aims and directives helped to guide the direction and motivation of this work.

A capstone to this work was attending the first NASA Networking for Ocean Worlds (NOW) retreat. This was a weeklong workshop on Catalina Island, CA. The goal of the

retreat was to bring together people from different disciplines and perspectives that have research related to our understanding of ocean worlds. This research provided a unique perspective to the discussions arising from the interpretation and utilization of prior studies on terrestrial marine chemistry.

1.2 Approach

The approaches to achieve these goals were two-fold. From the perspective of applying data science techniques and methodologies to marine chemistry, it involved exploring both classification (sorting data into discrete classes) and regression (providing continuous numerical answers) type problems. This has resulted in 10 different machine learning techniques being tested in their ability to answer chemical questions. Beyond simply training accurate models advancements were also made in the realm of feature analysis and model scrutinization, approaches for evaluating how the models learned the data. It is not sufficient to simply provide accurate answers, models providing the right answers for the wrong reasons can be more hazardous and problematic than models simply being inaccurate.

From a marine chemistry perspective, it was critical to simplify the system to a level that the questions being asked of the data were concise and the results were interpretable. The aim of the work was to have a method to describe the organic contents of the ocean in a generalized method. Having the ability to describe a “recipe” of the ocean provides incredible access to understanding ocean health on a variety of time and spatial scales through rapid field analysis. This began with single analytes in a simple chemical

matrix and ramped into analyzing multiple analytes simultaneously on true marine samples with their intrinsic large chemical complexity and diversity.

The themes of data science and marine chemistry collided in defining the dataset. These projects ranged from web scraping to the hand curation of datasets. In web scraping data volume is a nonissue, for certain datasets thousands of potential datapoints exist online. The limitations arise from unexpected irregularities in the data and the associated struggles with identifying a dataset of unknown representatives. At the other end of the spectrum, is making chemical matrices in the lab. In this context practically total control over chemical representatives and concentration ranges is possible. However, in this context the limitations become time, personnel, and cost of chemical reagents.

1.3 Dissertation Highlights

The roadmap of this dissertation is as follows. Chapter 2. Pertinent Background Information” provides context and background information to key aspects of the dissertation. 2.1 Chemical Systems” Provides background into the chemical systems of terrestrial (Earth) marine chemistry, and Enceladus, a moon of Saturn. 2.2 Analytical Instrumentation Explains the analytical instrumentation used to generate datasets for the work, these include the vibrational spectroscopic techniques of Raman and IR. These sections go into both the theory and explanation of the techniques as well as information regarding the specific analytical instrumentation used. 2.3 Machine Learning Approaches Explores the different machine learning approaches utilized throughout the dissertation.

Chapter 3. Array Based Machine Learning for Functional Group Detection in Electron Ionization Mass Spectrometry” explores the utilization of machine learning

methods to classify the presence of functional groups in compounds using electron ionization mass spectrometry. Convolutional neural networks and logistic regression were tested for their ability to identify 18 specific functional groups and generalized functional group classifications. Logistic regression was found to be much more effective than the convolutional neural networks. By analyzing the associated features of the more successful logistic regression models a suggested mass range and resolution for future planetary methods was suggested of unit resolution with a mass range that at minimum covers 1 – 100 m/z units.

Chapter 4. Saccharide concentration prediction from proxy sea surface microlayer samples analyzed via infrared spectroscopy and quantitative machine learning.” creates a framework for identifying organic compounds in marine samples. Saccharide concentrations were quantified using regression-based machine learning. A dataset was generated in lab using concentrations of glucose and egg serum albumin (ESA). Concentrations of glucose in these samples were used to train models utilizing six different machine learning methodologies. The models were further tested using a small supplemental dataset that beyond just containing glucose and ESA, also obtained bovine serum albumin (BSA), 1-butanol, and sucrose. The target saccharide concentration for these samples arose from the sum of glucose and sucrose. It was determined that support vector machines and gradient boosted regressors were the best models for accomplishing this task.

Chapter 5. Multi Analyte Concentration Analysis of Marine Samples Through Regression Based Machine Learning” expands upon the results from chapter 4. In this work

the same six methods, along with two additional methods were tested on two different datasets with the goal of identifying three different chemical analytes, saccharides, fatty acids, and proteins through the lens of amino acids. Two unique datasets were developed with this goal in mind. The first was a series of concentration gradients on ultrapure water, the goal of this dataset was to provide well resolved and isolated spectral signatures for the analyte compounds. The second applies the same concentration gradients on top of 10 unique marine samples from a variety of sources. After analyzing all of the combinations of dataset and machine learning technique it was found that support vector machines combined with the marine sample dataset was the most successful. A model dropout test was used to further scrutinize the model results to find model limitations and increase confidence in model results.

Chapter 2. Pertinent Background Information

2.1 Chemical Systems

Ocean worlds are bodies throughout the solar system which are theorized to contain liquid water. These bodies include many moons and exoplanets but also include Earth. All of the work presented here focused on the chemical systems of Enceladus, one of the moons of Saturn, and Earth as an analogue for ocean worlds while in tandem creating methods to understand our own marine environment.

2.1.1 Terrestrial Ocean Chemistry

When sampling ocean worlds, our own ocean provides the easiest target to explore due to its proximity. Advancements in this area directly affect our ability to understand aspects of ocean chemistry throughout the other ocean worlds. Even with the differences in the interactions between ocean and the atmosphere (or lack thereof) or the energy sources within the system may change, there is still a bulk liquid ocean that can utilize this transferred knowledge. A major factor in this understanding is the concept of the interfacial sea surface microlayer (SSML)

The SSML is chemically complex and different than the bulk ocean.¹⁻⁷ Acting as the “ocean’s skin”, the interactions of the SSML affect climate^{5,8-10} and ice nucleation.^{4,11-13} Because the SSML has different properties (organic and ion composition among others) than the bulk ocean,¹⁴⁻¹⁷ this layer is of particular interest for understanding the marine ecosystem. Due to the higher concentration of organics partitioning to the SSML, it is

enriched with lipids, proteins, and saccharides which all contribute to the total organic carbon (TOC).¹⁸⁻²²

2.1.2 Enceladus as an Ocean World

Enceladus has been a target of chemical intrigue since the discovery of its plume, an eruption of ice grains and volatiles from the surface of the moon that have the potential to contain biosignatures, if indeed the subsurface ocean contains life.²³⁻²⁸ The plume is proposed to be sourced from the subsurface ocean due to the presence of molecular H₂, methane, and nanograins of silica.^{29,30} These observations are consistent with water rock interactions and temperatures of over 90 C.³⁰ The observation of organic macromolecules in the plume required an explanation, leading to planetary scientists calling on prior literature and understanding of terrestrial marine chemistry (2.1.1 Terrestrial Ocean Chemistry). The current hypothesis is the presence of a layer of organic materials at the interface of the subsurface ocean and ice shell that is being sampled by the plume (**Figure 1**).^{23,26,27,29,31,32} Although the proposed SSML has not been directly sampled the presence of this layer would be consistent with observations of organics being sampled by the plume.

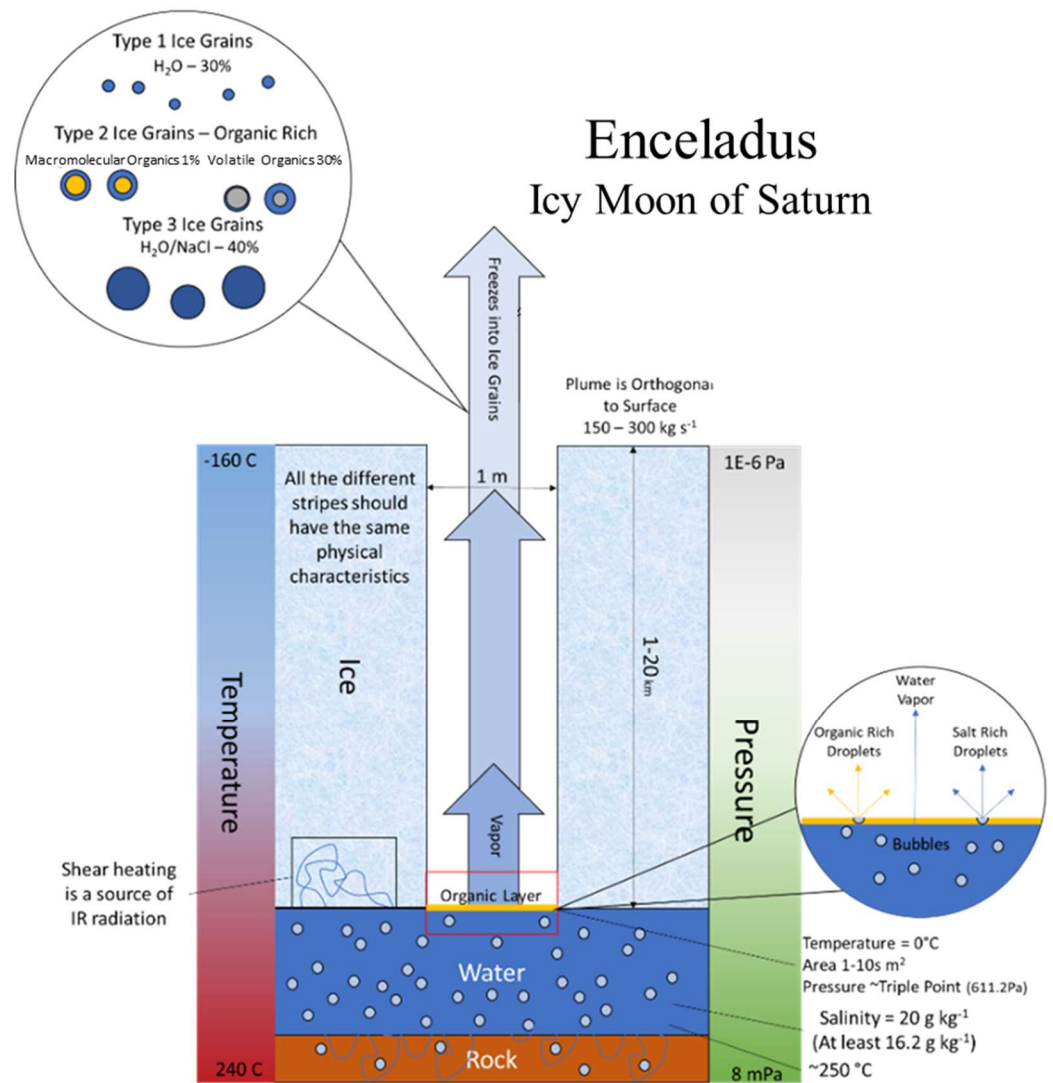


Figure 1. Simplified diagram of proposed physical and chemical processes occurring on Enceladus.

The putative organic layer has been sampled indirectly during multiple flybys of Enceladus by the Cassini Mission.^{23,25,27} The two onboard mass spectrometers – the Ion Neutral Mass Spectrometer (INMS) and the Cosmic Dust Analyzer (CDA) – performed measurements of the gas and ice grains that were ejected from the interior of Enceladus via

the plume. The difficulty in the full utilization of the INMS and CDA databases largely arises from the low mass cutoff and the low mass resolution ($m/\Delta m$ of 100 for the INMS where a traditional quadrupole has a 10x greater resolution, the CDA only has a $m/\Delta m$ of 20-50) of these instruments.³³

The complicated chemical environment of Enceladus' plume is of particular interest because it serves as a window into the (bio)geochemistry occurring within the subsurface ocean.^{23,34} This is convolved with processes happening at the ice-ocean interface, at fissures where ocean material is expressed into space via the plume.^{35,36} Low and high mass surface-active organic molecules are important to categorize for obtaining a more thorough understanding of the chemical and physical interactions occurring at this interface. Developing this type of analysis could therefore enhance the return of NASA Planetary Science Division missions such as the Cassini Mission as well as future missions sampling the plume of Enceladus or materials from other Ocean Worlds.

The mass spectrometry data obtained from Enceladus during the Cassini Mission has already had important impacts on planetary science. Postberg and colleagues identified a subset of the organic-molecule-containing (subset of Type II) ice grains as being high mass organic cations (HMOC). This was done by finding repeated patterns of +12-13 mass units for molecules with mass-to-charge ratios greater than 80 m/z . This pattern was important as it indicated the presence of compounds with a carbon atom count of 7 to 15. They also identified fragments that suggested the presence of benzene rings.²⁷ On the other end of the mass spectrum (pun intended), Khawaja and colleagues identified low mass organic molecules that contained N- and O- bearing and aromatic components. They were

also able to predict the presence of low-mass amines and carbonyls.²⁵ The quest to understand the ice grain data from Enceladus has also been supplemented by analogue experiments on Earth. Klenner and colleagues in multiple papers explored what possible biosignatures might look like if they were detected coming from Enceladus.^{37,38} This has been done by experimentally simulating the impacts of ice grains as we try to replicate the Cassini dataset conditions on Earth. They worked to determine the limit of detection for amino acids, fatty acids, nucleobases, and metabolic intermediates setting the benchmark of what biosignature detection might look like for Enceladus.^{23,26}

Table 1. Reported Mass Fragments from in situ measurements of the Enceladus plume.

Relevant Chemical Fragments of Type II Ice Grains from CDA and INMS Data					
	Category	HMOC, LMCS or Both	Mass of Ion (u)	Possible Identity of Ion	Possible Functional Groups the Parent Ion Would Contain
Type II Ice Grains ²⁹	Carbon Chains	HMOC ²⁷	N ± 12*	Unsaturated Carbon Chains	Unsaturated C ₇ to C ₁₅ chains
	O Containing Functional Groups	Both ^{25,27}	29-31	CH _{1,2,3} O ⁺	Carbonyl ^{25,27}
					Ethoxy ^{25,27}
			43/45	C ₂ H _{3,5} O ⁺	Hydroxyl ^{25,27}
	N Containing Functional Groups	Both ^{25,27}	18	NH ₄ ⁺	Amine ²⁵
					Nitrile ²⁵
					Amide ²⁵
	Aromatic Functional Groups	Both ^{25,27}	77/79	C ₆ H _{5,7} ⁺	Benzene (Non-Fused) ^{25,27}
			89/91	C ₇ H _{5,7} ⁺	Phenyl ^{25,27}
					Benzoyl ²⁵
*N = nominal mass > 80 m/z.					

Table 1 exhibits a brief review of the identification of fragments within the ice grain dataset thus far. There have been studies assigning ions to possible organic fragments; the difficulty is that each of these fragments may have come from multiple possible functional groups or fragments of larger molecules. These types of assignments are not as difficult to do with traditional mass spectrometer instruments and they commonly use tandem mass spectrometry techniques, where ions in a single mass window are trapped and isolated prior to analysis, simplifying the mass spectrum significantly.³⁹⁻⁴¹ but this was not possible for the two MS instruments aboard Cassini.

2.2 Analytical Instrumentation

2.2.1 Raman Spectroscopy

Raman spectroscopy measures inelastic scattering, in which the vibrational energy changes after interaction with light. This means that upon excitation the bonds vibrate.⁴² These vibrations cause there to be a difference in the excitation wavelength before and after the collision with the molecule (

Figure 2).

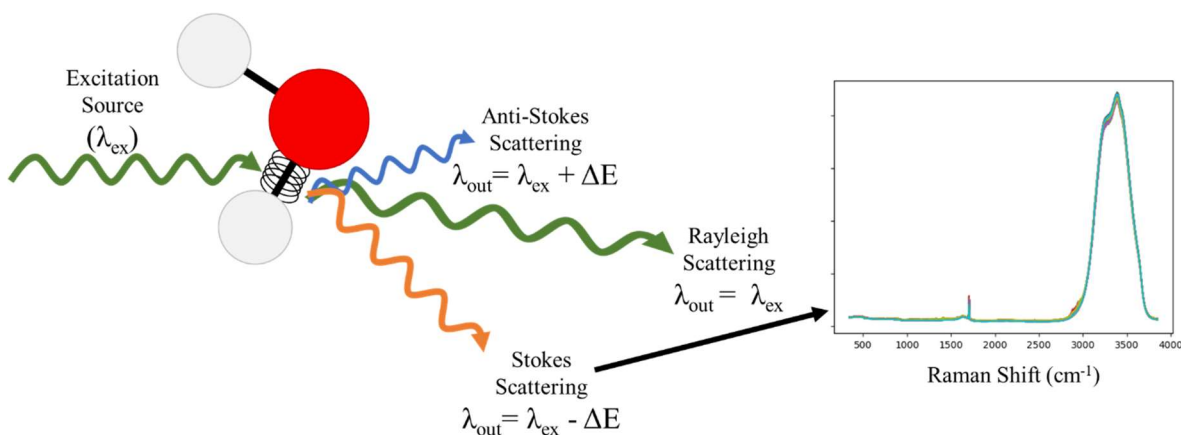


Figure 2. Overview schematic for Raman spectroscopy.

When light interacts with a molecule there are three possible outcomes. In order of drastically decreasing likelihood, the light can be elastically scattered and only its direction is changed, some of the energy can be absorbed through bond vibrations leading the light to lose energy and red shift upon scattering, and finally the already excited bond vibration can add energy to the energy of the collided light leading the scattered photon to be of higher energy and blue shifted. In Raman spectroscopy these are referred to as Rayleigh, Stokes, and anti-Stokes respectively (**Figure 3. A-C**).

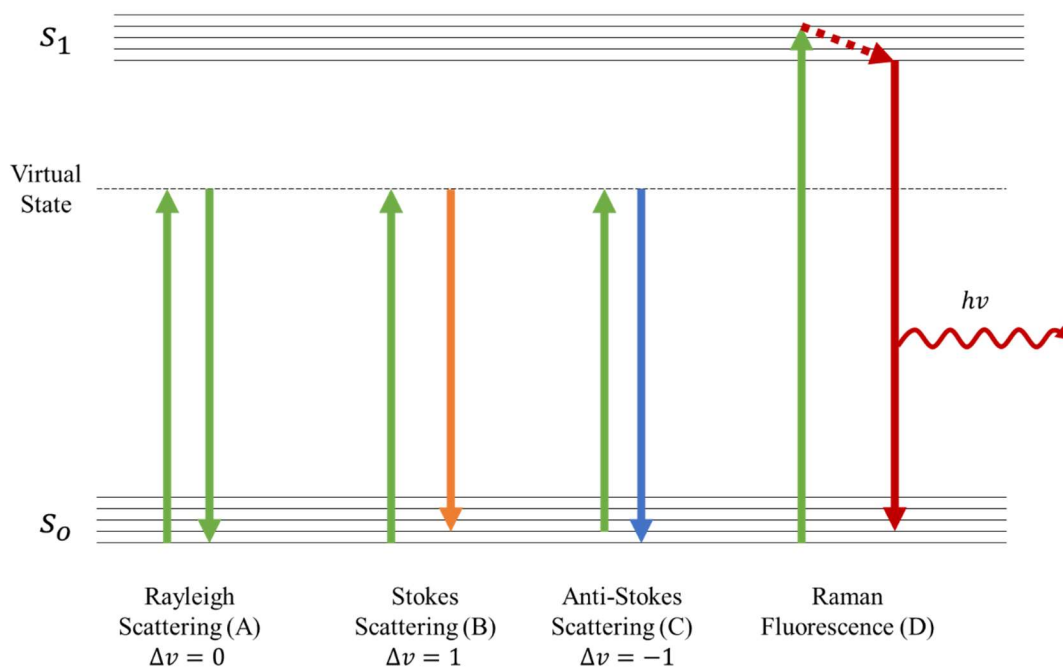


Figure 3. Simplified vibrational diagram of Raman processes.

Raman is typically reported as a function of frequency shift which has units of wavenumbers (cm^{-1}). This Raman shift can be calculated using the following equation.

Equation 1 Raman Shift:

$$Raman\ Shift\ (cm^{-1}) = \frac{10^7}{\lambda_{ex}\ (nm)} - \frac{10^7}{\lambda_{out}\ (nm)}$$

Excitation efficiency increases as the wavelength used for excitation decreases (proportional to $1/\lambda^4$). However, using these shorter and shorter wavelengths has downsides. For aqueous samples, the largest downside to using shorter wavelengths is an increased prevalence of Raman fluorescence. This occurs when the energy used to achieve the virtual state overlaps with a molecule's vibrational states (**Figure 3. D**). After relaxing through the excited vibrational states, a photon can be released as fluorescence. In the spectra this typically presents itself as broad peaks and elevated baselines.

The initial excitation source for Raman spectroscopy is a laser. This is due to the minimal wavelength variance within the source allowing for high wavenumber resolution on the back end. Laser stands for light amplification by stimulated emission of radiation. Stimulated emission refers to when an incoming photon interacts with an atom in an excited electronic state. This interaction causes the electron to drop in energy level releasing another photon in the process. That released photon can then go on to cause a stimulated emission event for another excited atom. The group of atoms that will be excited and used to emit photons is referred to as the lasing medium. These media can be solids, liquids, or gases.

From an electronic structure standpoint, lasers require many atoms to be in an excited electronic state. Under standard conditions most atoms do not exist in an excited state. To utilize stimulated emission to generate a laser beam a population inversion of

these excited states is required to generate the chain-reaction of stimulated emission events. Practically, to cause this to occur a minimum of a three-level laser is necessary (**Figure 4. A**). These three levels require three different electronic energy state changes. First atoms in their ground state are excited through a pump phase. This excitation typically comes from a light source or through running a current through the lasing medium (**Figure 4. B**). Next an internal relaxation occurs, and this reaction needs to happen quickly. After that the lasing transition occurs and this must occur orders of magnitude slower than the internal relaxation. This creates the necessary population inversion to make a laser work.

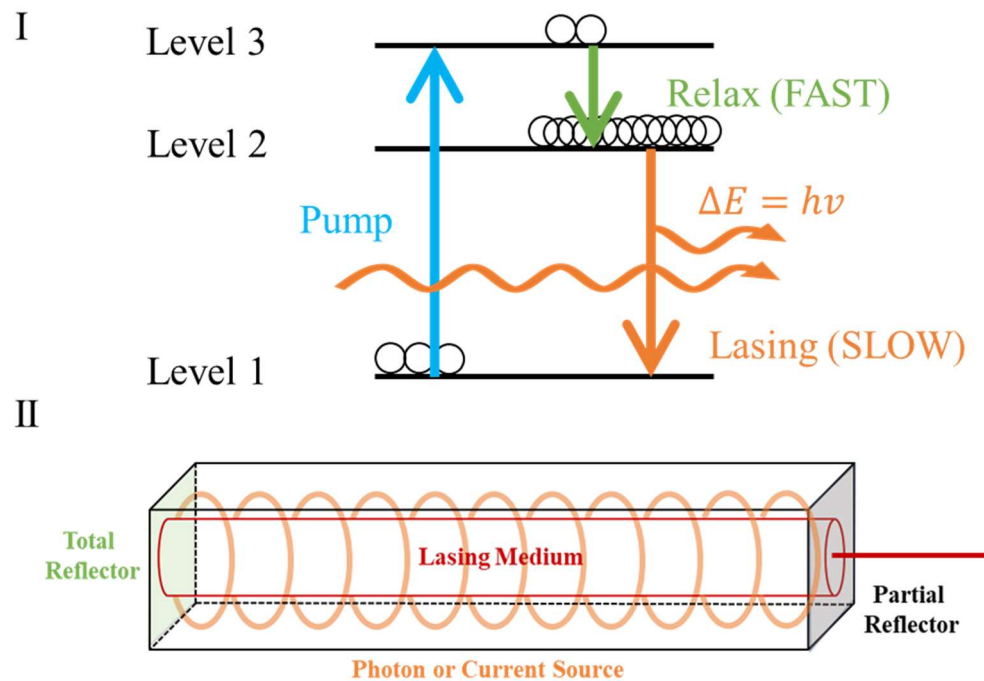


Figure 4. Diagrams depicting the theoretical (A) and practical (B) setup and implementation of a laser.

Laser light has three important characteristics. The light is monochromatic, coherent, and directional. Monochromatic means that the laser light is only one

wavelength. This is critical to Raman spectroscopy as the wavenumbers that make up the x-axis are relative with respect to the incident wavelength. If this monochromaticity is disrupted it can lead to duplication of Raman bands separated by the difference in wavenumbers of the true laser wavelength and any other wavelengths being scattered. Laser light is also coherent, this means that the waves of the light line up trough to trough and peak to peak and the photons are moving in the same direction. This coherence arises from the stimulated emission of photons and is critical to increasing the power of emitted laser light through the interference of the waves being constructive. Finally, the laser light is directional meaning that all of the light is being focused in the same direction. This directionality is in part governed by the series of mirrors in the laser housing. Having fully reflective mirrors throughout the laser apart from where the laser is leaving from where it is still largely but not entirely reflective allows the laser to have a low degree of beam divergence which is important to achieving high levels of signal and also minimizing risk of injury from stray photons.

The described instrument also can perform polarized Raman. This provides additional information that traditional Raman spectroscopy does not. This additional information includes the aspect of symmetry. Depolarization ratios (ρ) are calculated for resolved vibrational bands corresponding to individual vibrational signatures using **Equation 2**.⁴²

Equation 2 Depolarization Ratio:

$$\text{Depolarization Ratio } (\rho) = \frac{I_{\perp}}{I_{\parallel}}$$

Depolarization ratios of less than 0.75 are defined to be polarized bands and are expected to be fully symmetrical. Ratios of more than 0.75 are defined as depolarized bands and are assumed to be not completely symmetric.

The laser system for the described system contains a diode pumped 532 nm green laser continuous wave laser from CrystaLaser. The laser itself has two built in optical components including a laser line filter that ensures that the laser output is within 531.5 and 532.5 nm and a polarizing filter that ensures that all the light is vertically polarized with respect to the laser casing. This laser is then emitted directly into a sample holder that holds a quartz cuvette. The subsequent scattering of the laser light after interacting with the sample is collected by a custom-built Raman probe from InPhotonics. It can collect both light that is parallel (\parallel) and perpendicular (\perp) to the polarized laser light into two independent fiberoptic channels. Which are then vertically stacked when introducing them to the spectrograph. This makes the orientation of the fiberoptic critical to maintain when detaching and reattaching the combined fiberoptic to and from the spectrograph as the misorientation of this junction can cause issues with the signal intensity of one or both polarized channels.

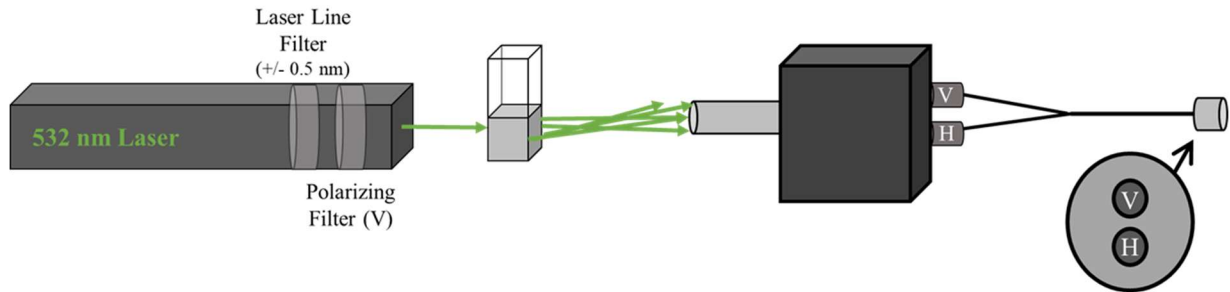


Figure 5. Diagram of collection of polarized Raman signal from excitation source to spectrograph.

The end of the fiberoptic is connected to an entrance slit connected to a micrometer (for reading micrometer see Appendix D. Supplemental Information for 532 nm Polarized Raman) with a 50 μm slit width. Adjusting this entrance slit width to be more narrow increases resolution but reduces overall signal whereas opening the slit has the opposite effect. The slit has an operating range of 0.010 – 3 mm (10 – 3,000 μm).

The spectrograph in the described system is a IsoPlane® from Princeton Instruments. The physical layout of the spectrograph is similar to a Czerny-Turner spectrograph (**Figure 6. B**) however Princeton has made propriety changes to avoid astigmatism so they refer to the design as a Schmidt-Czerny-Turner spectrograph (**Figure 6. A**). The IsoPlane® also contains three different gratings attached to a rotating turret.

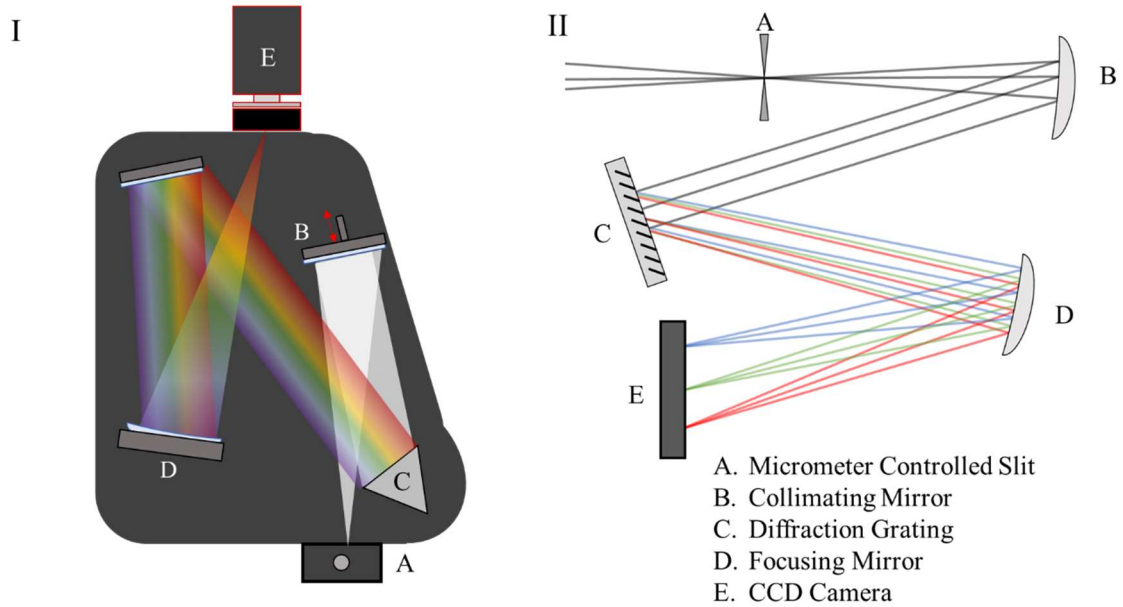


Figure 6. Theoretical Czerny Turner Spectrograph (II) and Schmidt-Czerny-Turner Spectrograph (I) layouts of the described spectrograph. Similar components have been labeled A-E.

These grating characteristics describe the efficiency of the wavelength separation by diffraction. Having a higher grating density, typically reported in the units of grooves per millimeter (g mm^{-1}) results in a greater spectral resolution by reducing the distance between grooves.

It is important to note that, for the described spectrograph due to the fixed size CCD, increasing the grating density from 600 g mm^{-1} to 1200 g mm^{-1} increases the resolution but it also reduces the range of the spectrum that can be collected by the CCD camera as that is fixed by length and pixel density.

Utilizing a blazed grating, like what is in the described spectrograph allows for the optimization of a certain order diffraction of light. Doing so maximizes for the grating

efficiency (amount of light being diffracted into the desired diffraction order relative to the total flux of incident light) without having to change other structural characteristics of the spectrograph. For the described setup the blaze values are referenced in units of nm. This shows which region of the electromagnetic spectrum that that blaze angle is optimized for (i.e. 715 nm blaze optimizes for the visible region and 300 nm blaze would optimize for the UV region). The described setup has the following grating and blaze options, 600 g mm⁻¹ with a 750 nm blaze, 1200 g mm⁻¹ with a 750 nm blaze, and 1200 g mm⁻¹ with a 500 nm blaze.

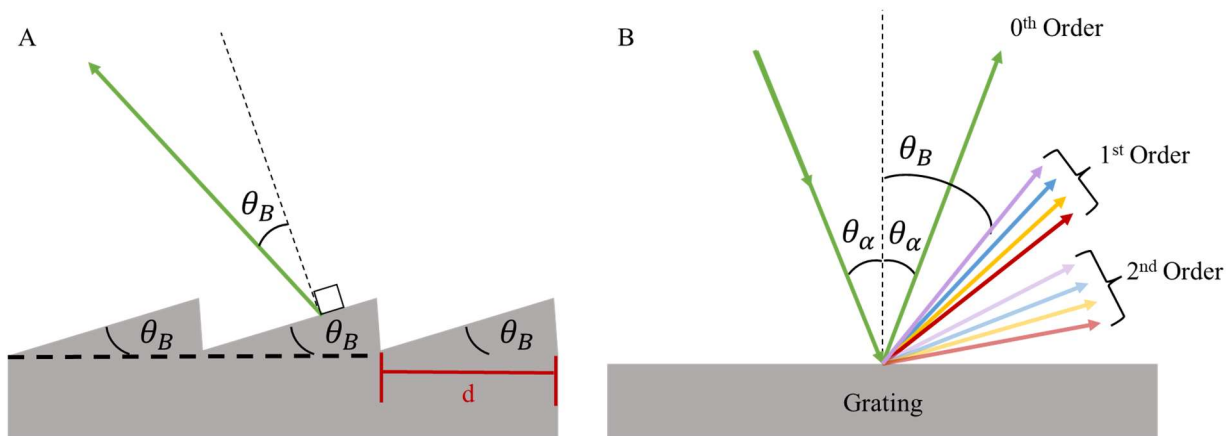


Figure 7. Diffraction of light with a blazed grating.

After diffraction the light is focused through a mirror and is detected by a liquid nitrogen cooled CCD camera. The CCD determines intensity at a given wavenumber through counting photons at a given pixel. The photons are converted to electrons through the internal metal-oxide-semiconductor. These electrons are collected for a given time (the exposure time) and that voltage is then measured. The voltage is then proportional to the intensity of the light at that pixel. This is why Raman signals are typically reported as

arbitrary detector units on the y axis these values vary from instrument to instrument and vary with exposure time. The most important thing for detection is to ensure that none of the spectra will oversaturate the pixels which will decrease the usability of the spectra collected but can also cause permanent damage to the CCD detector. The vertically polarized and horizontally polarized spectra are collected by vertically summing bands of the CCD camera (**Figure 8**).

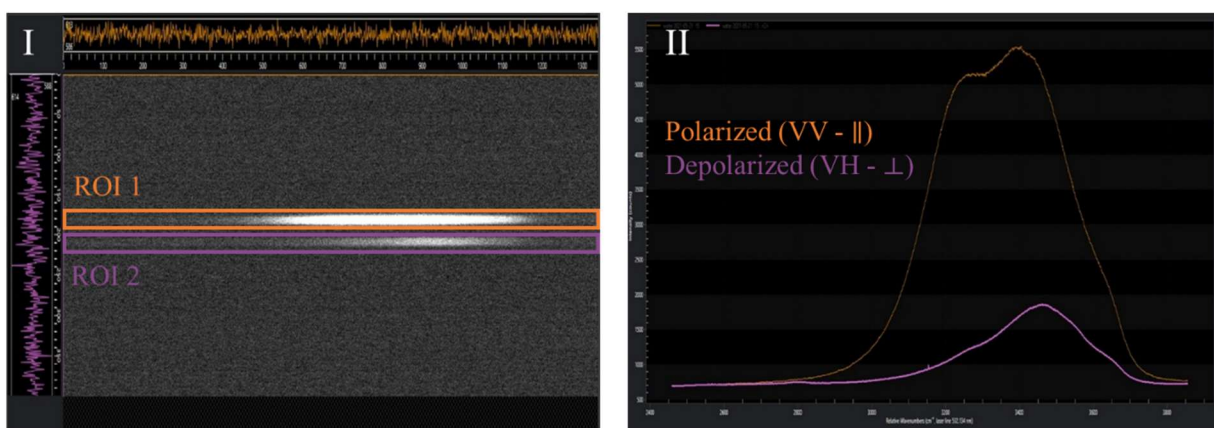


Figure 8. Image of Raman CCD (I). Summed regions of interest (ROI) converted into Raman spectra (II).

Note that in the polarized spectra presented there is a small artifact of intensity variability at approximately $3400\text{--}3450\text{ cm}^{-1}$; normalization did not prove useful to remove this small variability.⁴³ The consistent observation of this artifact did not have repercussions for the machine learning models as it is read just as a small systematic error in the data and is within every training spectrum and thus contains no predictive power.

2.2.2 Fourier Transform Infrared Spectroscopy

Infrared spectroscopy (**IR**) generally utilizes wavenumbers between 4,000 and 400 cm^{-1} for the mid IR range. When a sample is exposed to IR light some of that light is absorbed by the molecules generating an excited vibrational state. These energies are much smaller than the energies observed in Raman spectroscopy (section 2.2.1). **Figure 9** demonstrates these vibrational transitions observed.

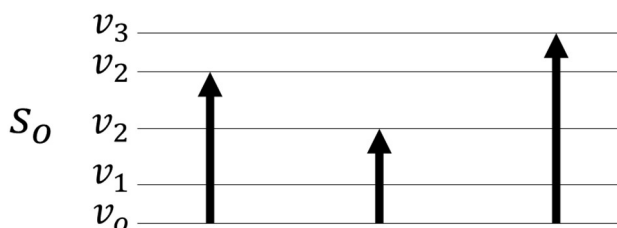


Figure 9. Energy diagram of the absorbance of IR light generating an excited vibrational state.

IR sources are typically black body radiators meaning that they take in voltage and emit broadband light in a variety of wavelengths depending on the material. The described work utilizes a glow bar which is typically made of silicon carbide.

The data described in this dissertation was taken at a resolution of 1 wavenumber from 4,000 to 440 cm^{-1} . This large range of wavenumbers coupled with the sampling of each individual wavenumber would lead to 3660 monochromatic values to be collected per sample. This slow data collection process is unnecessary when the IR is coupled with a Michelson interferometer. This interferometer allows for the collection of all of the wavelengths of light simultaneously within the time domain rather than the frequency domain.

A Michaelson interferometer works using a series of optics and mirrors to create reproducible patterns of constructive and destructive interference. This pattern is made using a fixed mirror and a rapidly moving mirror (**Figure 10**). Both of these mirrors are illuminated by the IR beam through a beam splitter. The difference in distances from the beam splitter between the fixed and moving mirror is what generates the interference pattern based on the phase of the light when it is reflected. After reflection the recombined beam is directed to the sample. This can be done through a variety of methods but for the described work the beam is directed through a crystal to perform attenuated total reflection (ATR).

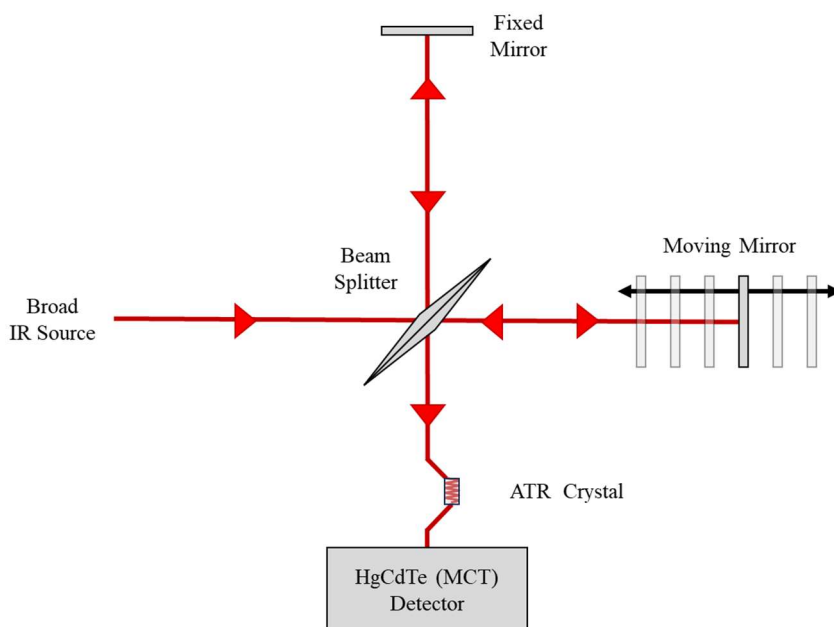


Figure 10. Path diagram of an attenuated total reflection (ATR) Fourier transform infrared (FTIR) spectrometer.

In ATR sampling, the sample is exposed to the IR source beam many times through the form of an evanescent wave. This occurs because the IR beam undergoes total internal

reflection many times throughout the ATR crystal.⁴² When a sample is in contact with this ATR crystal there is penetration of that IR beam into the sample where the vibrational signature of the sample changes the IR beam as certain wavenumbers are absorbed into vibrational modes (**Figure 11**). For this to be optically possible it is necessary for the crystalline material to have a higher refractive index than the sample that is to be analyzed, and for there to be sufficient contact between the crystal and sample. In the described work the ATR crystal is made of diamond which has a refractive index of ~2.4 and the samples are largely made of water which has a refractive index of 1.3~ (all refractive indices are reported for a given wavelength and temperature and there is small variation in the refractive indices if different temperatures or wavelengths are used).

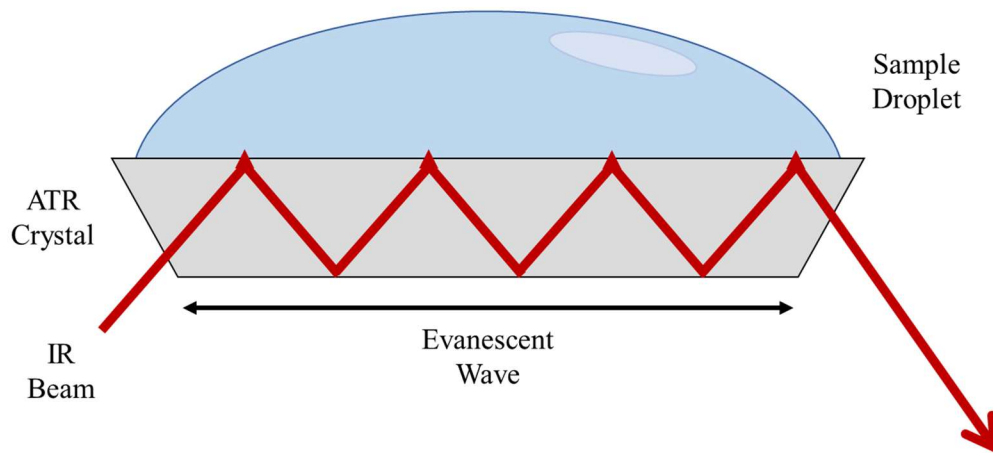


Figure 11. Evanescent wave moving through an attenuated total reflection (ATR) crystal and sampling a sample droplet through its contact with the ATR crystal.

After the IR beam interacts with the sample it is then sent to the detector. In the described work this is an HgCdTe or Mercury Cadmium Telluride (MCT) detector. This detector is particularly sensitive to mid IR light and is an example of a photovoltaic detector. This simply means that the detector is able to convert photons into an electrical signal. The collected signal is in the time domain, a Fourier transform is then used to convert the data back into the frequency domain.

2.3 Machine Learning Approaches

The learning in machine learning (ML) can be described as supervised, unsupervised, or reinforcement learning. Supervised learning involves the models being trained on labeled data, meaning that during the training the model is provided the data as well as the “correct” answer that the model is expected to reproduce if it were to see this

piece of data again in the future. Unsupervised learning is not provided with labels, instead the models must utilize similarities within the data to identify potential clusters of similar data through a variety of mathematical approaches.⁴⁴ Finally, reinforcement learning is what people typically think of when they consider machine learning for the first time. Reinforcement learning tasks consist of tasks that have their own internal scoring metric, like playing a game, or driving a car and following road rules. The described work are all examples of supervised ML meaning that from this point forward when referring to ML, it is implied that the ML is supervised.

Machine learning is typically broken down into two kinds of questions. First classification type problems work to answer the question, “What is this?” In this type of ML the data is broken up into labeled classes. And the data coupled with these quantized labels is used to train a model that typically will generate a numerical score to determine which class a new piece of data should be sorted into. Classification ML is the basis for Chapter 3. Array Based Machine Learning for Functional Group Detection in Electron Ionization Mass Spectrometry Regression type problems are the other kind of questions that ML can answer. Regression type problems are those that answer the question “How much?”. In this context the models’ answers come in the form of a float and can be any numerical value. Regression type questions make up the basis for Chapter 4. Saccharide concentration prediction from proxy sea surface microlayer samples analyzed via infrared spectroscopy and quantitative machine learning. and Chapter 5. Multi Analyte Concentration Analysis of Marine Samples Through Regression Based Machine Learning.

In all kinds of ML, the data is described as having a consistent number of features. In the general ML context these features could be anything from numerical values to labels. For the purposes of this work features are typically chemical data in the form of mass spectral data or spectroscopic data. For these examples a single feature would be a single wavenumber in a Raman spectrum or a single mass in a mass spectrum.

The data for ML is also typically split into 3 parts. These are referred to as the training, validation, and test data. The training data is used to adjust the model parameters so that it can “learn” the data and generalize the patterns within. The validation data is typically utilized during the training between each iteration of changing the model parameters. This provides a metric of how the training is going and predicts how well that model will perform on new data. Test data is typically withheld so that final model parameters can be calculated by observing how the model performs on truly new data.

2.3.1 K-Nearest-Neighbors

K-Nearest Neighbors (KNN) can be utilized for classification or regression analysis. In either case the algorithm utilizes the placement of previously learned data to identify either the class or value of new data. The number of reference points is defined as the K-Neighbors used to evaluate new data.⁴⁴ The scoring function that this algorithm uses is calculating the distance from the new point to a defined (K) known data points. For classification the model then assigns the new data to the class that has the smallest distance from the new data point (**Figure 12. I**). For regression the output value is calculated by comparing the values that are the spatially closest to the new data point (**Figure 12. II**).

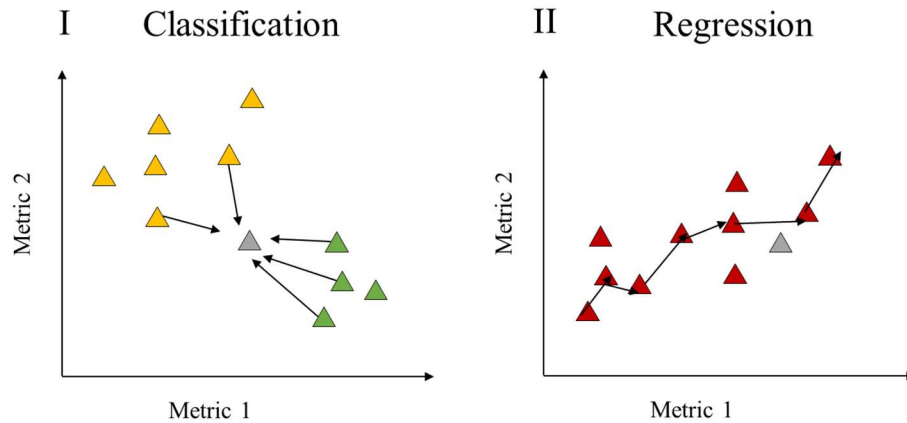


Figure 12. Summary of K-Nearest-Neighbor classification (I) and regression (II) approaches.

This distance can be calculated in many ways but the default for the scikit-learn function is the Minkowski distance⁴⁵ (**Equation 3**) When $p = 2$ this is a generalized form of Euclidian distance and is calculating the hypotenuse between two points to determine the distance.

Equation 3 Minkowski equation for distance:

$$Minkowski\ Distance = \left(\sum_{i=1}^n |x_i - y_i| \right)^{1/p}$$

Limitations of this model include being very data heavy in the final product. Unlike other ML methods that once they have a function fitting the data no longer need constant access to the training dataset, KNN utilizes the dataset with every decision thus making the final model no smaller than the data size of the entire training dataset. This model also

tends to struggle when the input data has many more features than it has unique data points. This can lead to overfitting thus over confidence of the model .⁴⁶

2.3.2 Logistic Regression and Support Vector Machines

Logistic regression, despite having regression in the name is typically used for classification type problems. Logistic regression is a binary classifier which uses the training data to maximize the distance between two classes numerically labeled as 1 and 0 arising from the likelihood that an event will occur. A sigmoid function is a common metric of choice for logistic regression (**Equation 4**).

Equation 4 Sigmoid function:

$$\sigma = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x}$$

When applying the sigmoid function to ML questions the x in **Equation 4** is replaced with a linear combination of features of the training dataset and associated weights. These weights are iteratively adjusted to maximize separation after the sigmoid activation is applied (**Figure 13**).

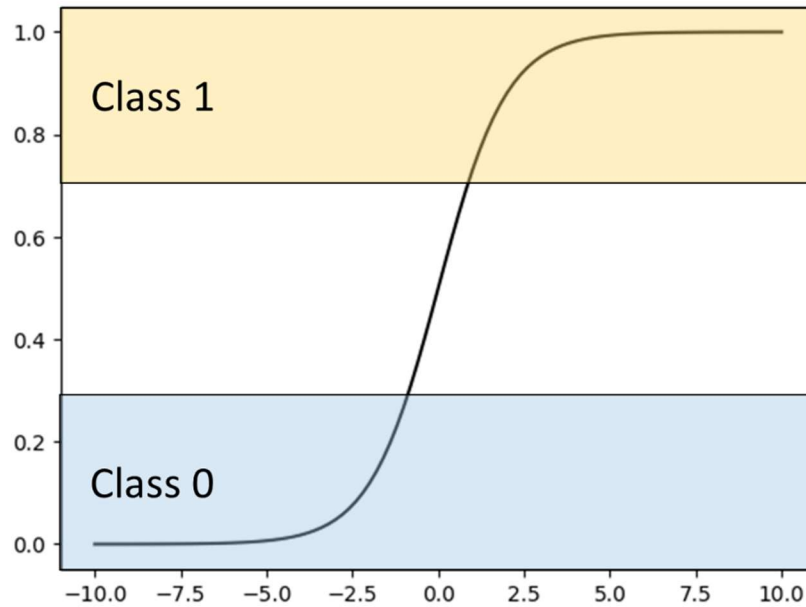


Figure 13. Summary of logistic regression classification approach.

Unlike logistic regression support vector machines (SVM) can be used for both classification or regression type analysis. In either case the SVM would be called a support vector classifier (SVC) or a support vector regressor (SVR) for classification and regression respectively. Similarly, to logistic regression SVMs also work to use a mathematical function to separate data for classification but it can also utilize high dimensionality data and functions to fit training data for regression purposes as well.

This fitting is done by transforming the data into a higher dimensionality space via a mathematical kernel to apply a hyper plane to either fit or separate data points of the training dataset (**Figure 14**). The data points that are closest to that fitting or splitting hyperplane are referred to as the support vectors and directly influence the placement and adjustments of the hyperplane.

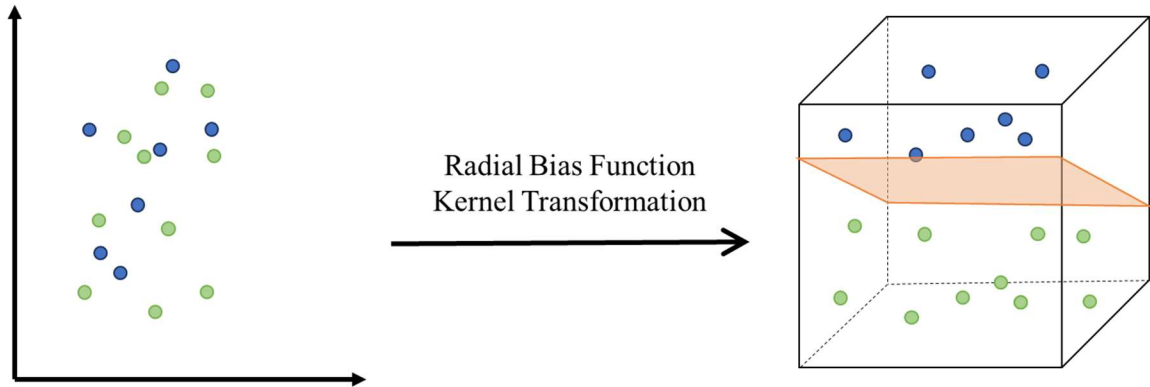


Figure 14. Summary of support vector machine kernel transformation for the application of hyperplanes.

In this work the mathematical kernel used for transformations is a radial bias function (RBF), also known as a squared-exponential function. The subfunction d in the following equation represents the Euclidian distance (**Equation 3** when $p = 2$) and it is scaled by a length scale (l) which has a default value of one.⁴⁷ RBF is a common kernel for both classification and regression type problems and is used in this work (**Equation 5**).

Equation 5 Radial Bias or squared exponential function:

$$k(x_i, x_j) = \exp \left(-\frac{d(x_i, x_j)^2}{2l^2} \right)$$

2.3.3 Decision Trees and Associated Ensemble Algorithms

Decision trees are algorithms that can analyze both classification and regression problems in ML. They work similarly to how dichotomous keys work in phylogenetics in which a series of binary questions are asked to separate data into discrete classes. In this work the specific loss function being minimized to split up the data is the classification and regression tree (CART) used by scikit-learn (**Equation 6**). In the following the n

variables correspond to the number of data points in total or in either class (right or left), and the H variables correspond to the group homogeneity of the data that has been grouped together.

Equation 6 Cost function for classification and regression tree (CART):

$$J(f, f_k) = \frac{n_{left}}{n} H_{left} + \frac{n_{rig}}{n} H_{rig}$$

This cost function looks at a single feature (f) and a threshold metric (t_f). As the model is trained. An example of this split could look like splitting data based on data that has an intensity of 1000 ADU at 3010 cm^{-1} for a Raman spectrum. Data that met or exceeded that intensity would be placed into the left group and the data that did not meet the threshold would be placed into the right group. These decisions are repeated and again until all of the data has been separated based on the function parameters that were originally stated in the Python script. Each internal decision node can use a different feature and they are all evaluated to find the features that best separate the data. Once the model has been trained a new datapoint can be run through the model to determine which leaf node it belongs to and then the tree can decide the assignment of the new datapoint (**Figure 15**). This training method is prone to overfitting based on the limitations of using only one feature at a time. If a dataset has many features the number of possible combinations of utilized features balloons quickly.

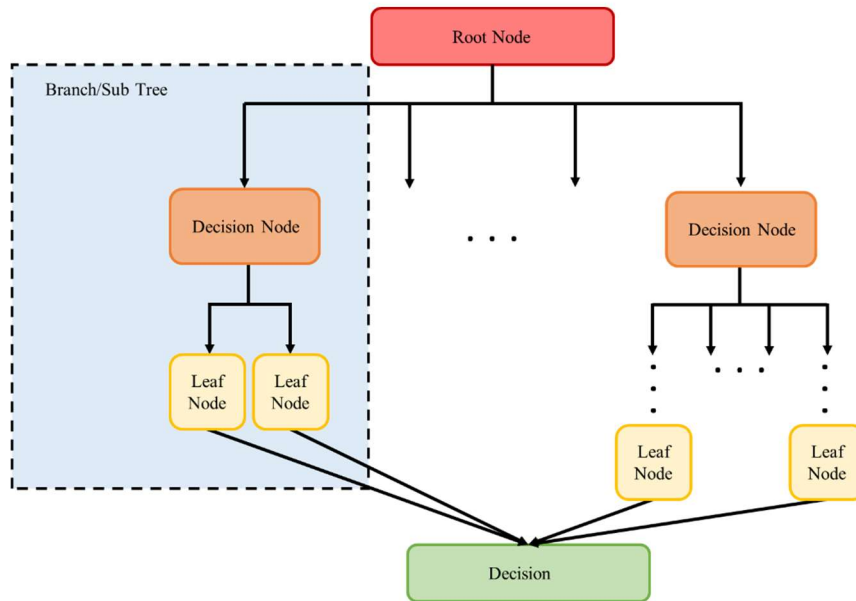


Figure 15. Generalized decision tree schematic.

Because decision trees are computationally inexpensive to train, and to offset the overfitting limitations of decision trees it is possible to use many decision trees simultaneously to generate more complex models. Approaches that utilize the training and implementation of many smaller, less complex models are referred to as ensemble models. In ensemble training all models are trained on the same dataset but each model is able to come to its own conclusion and make a decision on the data. Then larger scale decisions can be made by looking at the decision distributions of all of the models within the ensemble. It is possible to customize ensemble approaches to use multiple kinds of models simultaneously but in this work only decision trees are used within the ensembles. There are two general methods to use many decision trees to make an ensemble model, these are random forests and gradient boosting.

The main difference between random forests and gradient boosting is in what order are the decision trees trained? For random forests all of the trees are trained simultaneously and typically in parallel. Each tree gains no information as to how the other trees are making their decisions. Each tree then gets a “vote” in the final ensemble solution. The solution with the most votes is what the model outputs as its final answer (**Figure 16**).

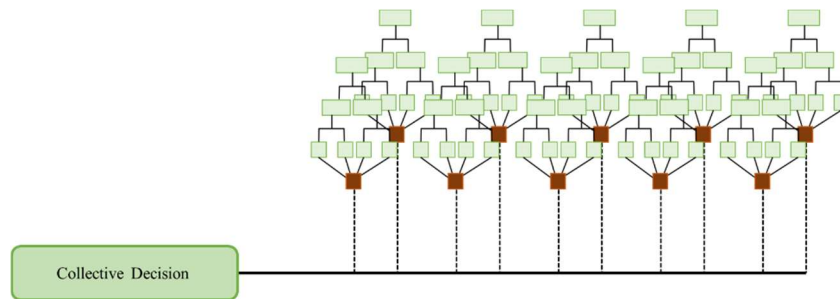


Figure 16. Generalized scheme for random forest models.

Gradient boosting, in contrast, only trains one tree at a time but each tree is trained on the residuals on the previous tree. In short, each tree is directed to where the previous tree went wrong in its assignments and is able to adjust for each subsequent tree (**Figure 17**). Over many iterations the trees ideally get better and better until they converge on a optimized splitting structure. Unlike random forests, gradient boosting requires many more computational resources. This computational expense can be reduced by utilizing different mathematical solvers, functions like this include the histogram gradient boosting.

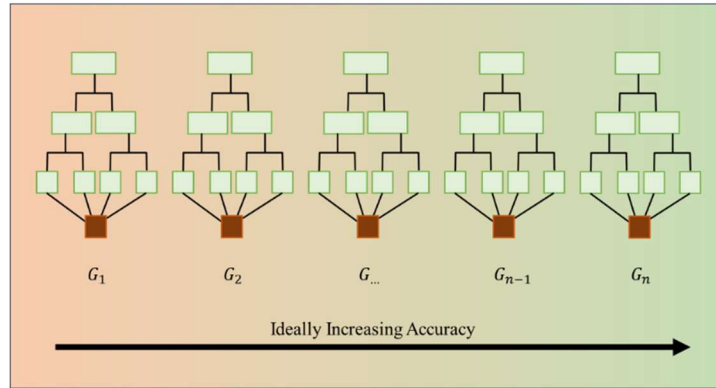


Figure 17. Generalized scheme for gradient boosting.

2.3.4 Neural Networks

Artificial neural networks (ANN) are named such for their similarities in structure to brain and nervous system neurons. Neural networks can be used for classification and regression ML problems. The earliest neural networks came in the form of the single perceptron in the 1950s. The general structure of the single perceptron is a single layer of threshold logic units (TLU) and a bias value. These TLU have input values, and each input value has an associated weight. These inputs and weights are used to calculate a weighted sum which is then provided to a step function which traditionally would provide a 0 or 1 as a kind of classification output based on whether the weighted sum was above or below a threshold value.

The ways that these perceptrons are trained are based in the understanding of biological neurons as well. The idea of neurons that fire together wire together was coined in the 1950s and suggests a method of adjusting weights and biases to improve perceptron training. These models are trained in a step-wise iterative process the adjusted weights (w_N) from one step to the next can be described as the previous step's weight (w_{N-1}) plus the

learning rate (η) multiplied by difference between the output of the previous step's output and the target or true output multiplied by the input value (x) This is shown in **Equation 7**.

Equation 7 Perceptron Learning:

$$w_N^{Next Step} = w_{N-1} + \eta(y_{N-1} - \hat{y})x$$

Instead of only using one layer of these simple neurons it is possible to stack multiple of these layers together to create a multi-layer perceptron. After the first layer instead of receiving the inputs of the model, all subsequent layers (N) would receive the result of the $N-1$ layer's computation that the weights and associated biases on the initial input or the output of the $N-2$ layer. The internal layers between the input and the eventual output are referred to as hidden layers. These additional layers add computational complexity to the model, allowing it to analyze more and more complex problems and datasets (**Figure 18**). Eventually when enough layers are used this is referred to as deep learning as there is a deep stack of hidden layers involved in the computation.

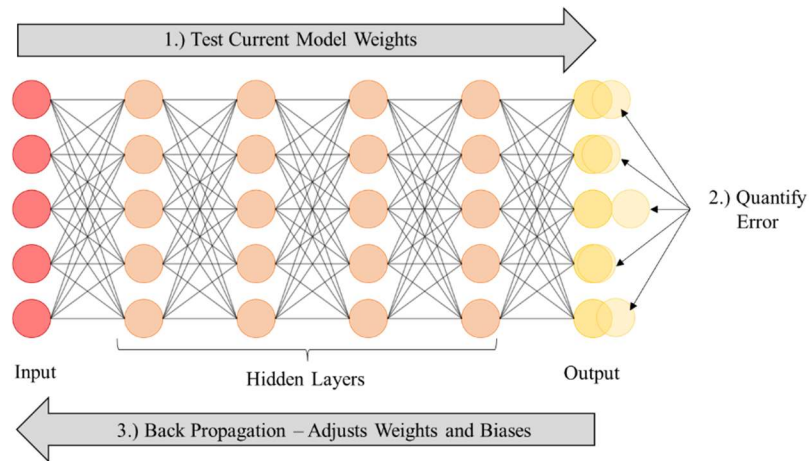


Figure 18. Generalized scheme for artificial neural networks.

Chapter 3. Array Based Machine Learning for Functional Group Detection in Electron Ionization Mass Spectrometry

Mass spectrometry is a ubiquitous technique capable of complex chemical analysis. The fragmentation patterns that appear in mass spectrometry are an excellent target for artificial intelligence methods to automate and expedite analysis of data to identify targets such as functional groups. To develop this approach, we trained models on electron ionization (a reproducible hard fragmentation technique) mass spectra so that not only the final model accuracies, but moreover, the reasoning behind model assignments could be evaluated. Convolutional neural network (CNN) models were trained on 2D images of the spectra using transfer learning of Inception V3 and logistic regression models were trained using array-based data and a Scikit-Learn implementation in Python. Our training dataset consisted of 21,166 mass spectra from the United States' National Institute of Standards and Technology Webbook. The data was used to train models to identify functional groups, both specific (e.g., amines, esters) and generalized classifications (aromatics, oxygen-containing functional groups, nitrogen-containing functional groups). We found that the highest final accuracies on identifying new data were observed using logistic regression rather than transfer learning on CNN models. It was also determined that the mass range most beneficial for functional group analysis is 0 – 100 m/z. We also found success in correctly identifying functional groups of example molecules both selected from the NIST database and experimental data. Beyond functional group analysis we also have developed a methodology to identify impactful fragments for the accurate detection of the models'

targets. The results demonstrate a potential pathway for analyzing and screening substantial amounts of mass spectral data.

3.1 Introduction

Functional group identification is an important strategy for molecular structure analysis in analytical techniques such as mass spectrometry.⁴⁸⁻⁵¹ Mass spectrometry often looks at fragmentation of molecules so that the original (parent) structure may be elucidated.⁵²⁻⁵⁴ Such analyses can be challenging. The presence of functional groups can aid in predicting where fragments will occur, however, identifying specific fragments corresponding to the presence of functional groups proves difficult.⁵⁵ Machine learning (ML) methods aid in pattern recognition when supplied with large data sets. This couples nicely with mass spectrometry's fragmentation patterns, making ML a promising tool to identify functional groups, and thus, fragments of interest.⁵⁶⁻⁵⁹

Generally, mass spectrometry is not as commonly used for bulk functional group analysis without the use of extra sample preparation or tandem mass spectrometry techniques (MS/MS).^{60,61} For example, in previous work the analysis of amino acids has been aided by derivatization via ninhydrin prior to using high performance liquid chromatography and tandem mass spectrometry for analysis.^{62,63} It is also possible to use tandem mass spectrometry approaches including triple quad mass spectrometry to perform precursor ion scanning to screen for functional groups.⁶⁴ These approaches are invaluable to the mass spectrometry community because they allow for in-depth analysis of chemical compounds. In addition, these approaches have created a higher level of understanding of complex analyte mixtures inclusive of those containing high mass molecules, for example

in the field of proteomics. However, there are circumstances in which prior derivatization, separation, and tandem methods are not feasible. Situations in which time, resources, and/or location make such analysis impossible, such as with field-based analyses and planetary probes.

The employment of ML has the potential to overcome many of the challenges faced in analyzing mass spectra under limiting conditions. ML approaches have a strong backing in the literature regarding their ability to classify organic molecules through their fragmentation patterns. For example, CANOPUS⁶⁵ which works to predict thousands of classes of molecules using MS/MS data or MSNovelist⁶⁶ was able to identify the structures of molecules that the model had never seen in the training phase. Similarly, CSI:FingerID⁶⁷ also utilizes MS/MS spectra to assist in searching a molecular structure database. Another application that takes advantage of the intersection of mass spectrometry and machine learning is in the understanding of metabolite chemistry.^{68,69} There are also many papers utilizing machine learning with mass spectrometry to perform rapid screening methodologies for specific analytes of interest.^{70,71} These machine learning methods have had powerful results and have been revolutionary in our implementation of mass spectral methods.

In this study, we aim to achieve meaningful fragment analysis using machine learning methods that do not require the use of tandem mass spectral techniques or controlled sample preprocessing. We generate a simplified method that can be applied in situations in which more sophisticated mass spectrometry techniques are not feasible, opening the door to many applications that have, to this point, been inaccessible with the

current analytical techniques. We achieve this goal by only using single mass analyzer data, meaning that further fragmentation information on parent fragments is unavailable. By doing minimal preprocessing, particularly in not manually selecting peaks of interest, we generate models that need to develop their own understanding of fragmentation patterns, which we can evaluate. In doing so, we explore how a generalized method for analyzing mass spectra informs interpretation of mass spectra for functional group analysis. Our methodology enables us to probe the model assignment mechanism, which further improves how we understand the functional group assignment and ML techniques.

Herein we present a comparison of functional group analysis methods from electron ionization – mass spectrometry (**EI-MS**) spectra. We evaluate the success of two ML approaches, transfer learning on a previously trained convolutional neural network (**CNN**) and logistic regression (**LR**). Transfer learning has previously been successful in identifying functional groups from infrared (**IR**) spectral data,⁷² therefore its application to functional group analysis in mass spectra was evaluated. In contrast to transfer learning, LR provides a simpler architecture to allow for further analysis into the impact of the features themselves on the outcome of the models.

The transfer learning on a CNN and LR algorithms were used with the same set of mass spectral data to identify specific functional groups (e.g., amines, esters) within molecules, as well as place the molecules into generalized classifications based on these functional groups (aromatics, O-containing functional groups, N-containing functional groups). We first explain the process of organizing the spectra obtained from the National Institute of Standards and Technology (**NIST**) webbook through web scraping. We then

show how the classifications of molecules are assigned prior to training followed by adjusting the different training parameters and how they affect both the final training and testing accuracies of the models. We then dive deeper into the LR based models to explore how adjusting mass ranges affects the model accuracies as well as exploring methods to quantify how the model is making its predictions.

3.2 Methods

3.2.1 Spectral Preprocessing and Machine Learning Parameter Selection

Prior to training the CNN and the LR models, the data was sorted and labeled. Jupyter notebooks describing these processes along with the model training will be available on our GitHub (https://github.com/Ohio-State-Allen-Lab/Mass_Spec_Functional_Group_ML). Data sorting and labeling was completed by identifying the functional groups that each molecule contained; this identification was done by looking at the InChiKeys. Segments of the InChiKey can be correlated with specific functional groups allowing for labeling of molecules. This process was tailored for our purposes from another publication.¹⁸ After identifying the presence or absence of individual functional groups, the molecules were then sorted into the more generalized functional group classifications (e.g. alcohol, amine, etc.). After defining each of the functional groups, the number of available spectra for each functional group identification was determined. **Figure 19** shows the distribution of the functional groups present in the NIST mass spectra.

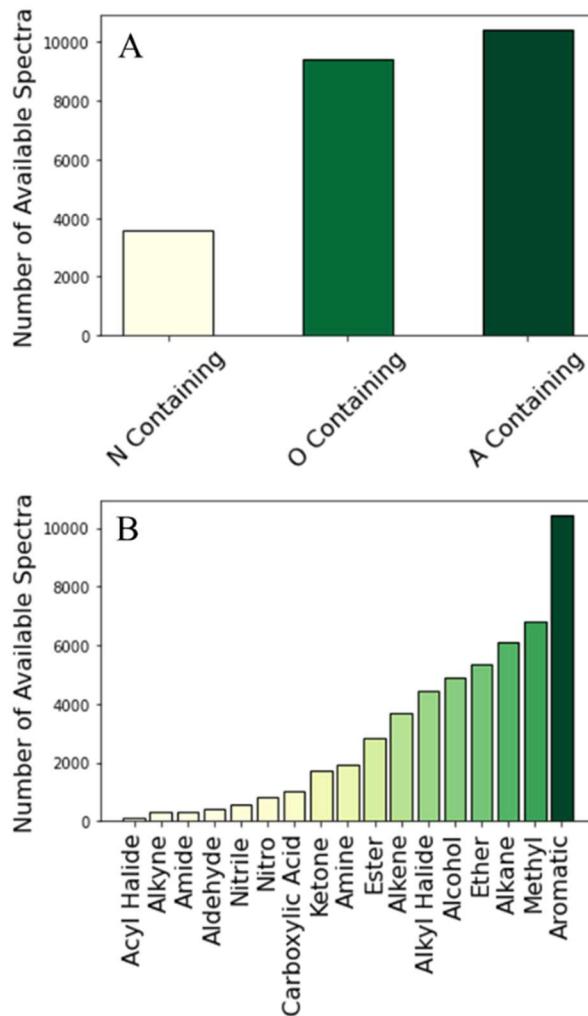


Figure 19. Distributions of available mass spectra from NIST included in this study are presented here. **(A)** Shows the generalized functional group classifications. **(B)** Shows the specific functional groups. Aromatic is listed as a specific functional group to help correlate the relative distribution between the generalized models and the functional group specific ones.

All the mass spectra were normalized to their most intense fragment peak to ensure that all the y axes were scaled the same way. NIST mass spectra only reports intensities for mass fragments over a certain intensity. For these data to be able to be compared to each

other they all needed to have the same dimensionality. To match up the data the unreported peaks were filled with a correlated 0 intensity. This was done based on the fact the non-reported peaks were assumed to be in the noise of the instrument. This is a limitation because the addition of the zeros although necessary for the training of the models does artificially inflate the signal to noise ratio of the data. This preprocessing was sufficient to prepare the data for the LR based models. The CNN based models required further preprocessing.

For the CNN based models the data was plotted. These plots were then used as the input data. All the spectra were saved with the same output parameters, so the resolution of the plots is consistent. However, further analysis of the pixel resolution of the exported plots showed that the plots are fewer pixels wide than there are mass values. This means that each pixel is not defining one mass channel as one would expect, this leads to an artificial reduction in mass resolution which likely is the source of the lack of success for this approach. We do run into an artificial reduction in the resolution of the mass spectral data. The exported plots are 2D representations of the data and should not be confused with hyperspectral imaging which would generate 3D data.

For both methodologies it was necessary to scale the number of spectra that did not contain the model's functional group of interest. The number of spectra that did contain a given functional group or functional group classification was always outnumbered by the number of spectra that did not contain the given functional group or functional group classification. Because of this, spectra were randomly removed from the negative case in order to even out the classes preventing the models from always predicting the not present

class due to a disparity in the data. This also means that if a molecule had multiple unique functional groups, that spectra would be used in some way for each of the represented functional group models. **Figure 20** shows a histogram that shows the average number of unique functional groups to be present in molecules in the NIST database is three.

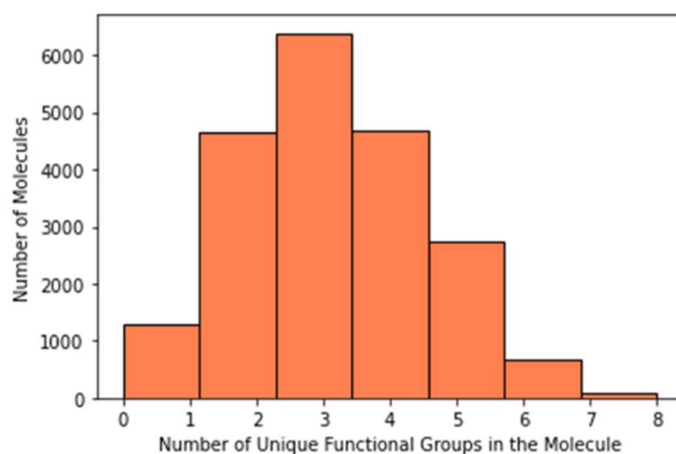


Figure 20. Histogram depicting the number of unique functional groups (duplicate functional groups within a molecule are not counted) present in each molecule from the NIST database. The largest distribution is molecules that contain 3 unique functional groups. Because most molecules contain multiple functional groups, they can be used to represent the positive case for multiple functional group models.

After preprocessing the data, it was separated into training and testing data sets (**Figure S1**). Once the model was trained, the test data was then used to determine how well the models performed on previously unseen data. The number of withheld test spectra was different for the two different parts of the project. When comparing the CNN and the LR based approaches only 10 spectra from each class were withheld. This was limited by computational expense. The workstation that was utilized to run all the training and

analysis was insufficient to run more than 10 test samples at a time. When focusing on using only the LR based models, 50 test spectra were withheld from each class before training. This allowed for further analysis of the accuracies of the models. After the testing data had been removed the remaining data was parsed into an 80:20 split of training and internal validation.¹⁹

3.2.2 Supplemental Experimental Data Collection

Mass spectra for multiple compounds were collected for further model analysis on experimental data outside of the NIST dataset. The data was collected from an Agilent 8890 GC coupled with a 5977B MSD.

3.2.3 Model Training and Testing

CNN and Inception V3

The architecture for our CNN in this work was a retraining of Inception V3,⁷³ a computer vision model. Inception V3 was trained on and has attained a greater than 78.1% accuracy on the ImageNet dataset (a large data set of millions of images with thousands of different words or word phrases labeled to them, a common test dataset in the computer vision realm⁷⁴⁻⁷⁶). ImageNet is certainly very different than a dataset consisting of 2D representations of mass spectra however the process of transfer learning on unrelated datasets has shown success in the literature.⁷⁷⁻⁷⁹ The Inception architecture has been used explicitly in the past for spectral processing applications.⁸⁰ Image processing CNNs have been used in other mass spec studies, for example, in 2019 Tran and colleagues developed DeepNovo-DIA which utilized intensity vectors to train a model to identify peptides.⁸¹ This history in the literature coupled with this approach's success with image based IR data in

our prior publication drove our decision to utilize retraining Inception V3.⁷² These models were trained using a learning rate of 0.1 and training step ranges between 200 and 20,000 steps.

Logistic Regression Through SciKit Learn

The LR models were developed using SciKit Learn's logistic regression classifier. In our utilization we use the newton-cg solver. LR was chosen as our alternative ML approach due to its simplicity. Using a less computationally complex, and specifically binary classifying, model allows for further analysis of where the inferences and assignments of the models are coming from. As we will show later this simplicity allows us to adjust the dataset and evaluate how those changes affect model outcomes.

LR is typically used as a binary classifier.^{82,83} This is because of the mathematics behind the architecture, the training of the models are working to identify the classification by maximizing the distance between the classes. This approach is very similar to support vector machines as both models maximize separation instead of minimizing an error function. This restriction of being a binary classifier coupled with using the entire mass spectrum as features are our reasoning for choosing to generate each model for the purpose of either identifying one specific functional group or one functional group classification. As we will explore later, specializing each of the models allows for the greatest model fit for that functional group as well as providing an avenue in which we can also describe how that greatest fit for each functional group was achieved.

3.3 Results and Discussion

3.3.1 Acquiring the Dataset

The mass spectra were web scraped from the NIST webbook using a web scraping implementation, details of which are described in our previous publication.⁷² In short, a web scraping script was written to individually download the mass spectral files from each of the NIST webbook pages that are labeled by CAS number. We obtained a total of 21,166 mass spectra. The files that were downloaded were in a JCAMP-DX file format. These files were then converted from JCAMP-DX into csv. The process does remove the associated metadata; however, this information was not necessary for our analyses. Once the files had been converted to csv further preprocessing could be completed. More information regarding the preprocessing steps has been reported in the SI.

3.3.2 Comparing Convolutional Neural Networks and Logistic Regression Feasibility

Both CNN and LR architectures were used to train functional group specific models and models to look at the functional group classifications. CNN was initially chosen due to its success in identifying functional groups using an IR dataset collected from the NIST database in our previous publication.⁷² Both approaches were each trained on a unique dataset which was a subset of all the data web scraped from NIST. Once all the models were trained it was possible to look at the final training accuracies to determine how well the final models fit the data sets.

There are two metrics that we utilized to describe how well the models were performing. The first of these is final training accuracy. This metric describes the final

ability to fit a segmented subset of the testing data after all the training steps have been completed. For our models we do a 80:20 split of training and internal validation data which is cited as being the most beneficial split.⁸⁴ The training accuracy is a description of how well the data can fit the data that it has been trained with. The second metric of interest is the final testing accuracy. This metric arises from how well the model can manage novel data. This metric is determined by analyzing previously withheld data using the models. Before generating the training datasets certain spectra are removed from the total dataset and withheld for testing the final model accuracy. This metric is critical in understanding how we can expect our models to perform with data in the future. **Figure 21** shows the final training and testing accuracies for four different functional groups' specific functional group models. Here we compare the training accuracy and testing accuracy of the specific functional group models for mass ranges 0 – 250 and 0-500 m/z. Model training and testing accuracies for all nineteen functional groups explored are shown in the SI.

Based only on the training accuracies, it appears that the models generated through CNNs should show a greater final accuracy in the specific case than the LR based models. The training accuracy values however do not tell the entire story. This highlights one of the main erroneous assumptions that is commonly made about ML. A model with an incredibly high fit of the training data is not necessarily better at describing novel samples. This metric is better described through the testing accuracy. This trend holds true for both the functional group specific (**Figure 21 A and C**) and functional group generalized models (**Figure 21 B and D**).

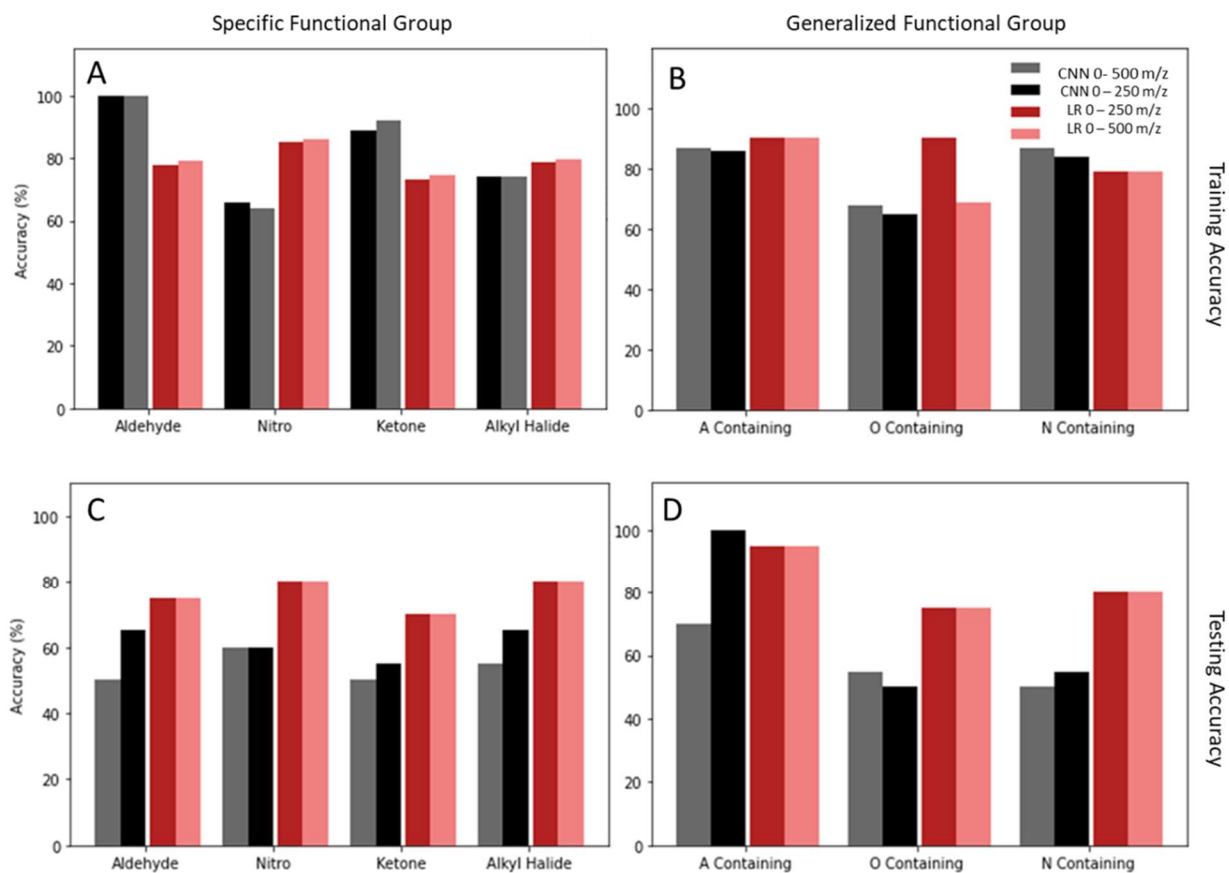


Figure 21. The results of the training and testing for four specific functional groups and the three functional group classifications are shown above. (A and B) Show the final training accuracy, accuracy of identifying the training portion of the data after the final training step has passed for both the functional group specific and functional group generalized models, respectively. For example, these plots would suggest that the CNN based approach should be better at correctly identifying the Aldehydes and the Ketones and that the LR based approach should have an edge on the Nitro group and the Alkyl Aldehydes. This, however, does not tell the full story. (C and D) The final test accuracies for the functional group specific and functional group generalized models respectively. The testing accuracy of the models is the accuracy of the models when presented with new

previously unseen data shows that a high training accuracy does not correlate necessarily with a high final testing accuracy.

A question that arose during this analysis was why did the transfer learning work so well with the IR data and so poorly with the MS data.⁷² The reasoning for this discrepancy likely falls under the differences in the atomic processes that are described with that technique. For IR data, because it is vibrational spectroscopy, we see the signals taking broader peaks that are influenced by bonding environment. This means that phase and having other molecular species in solution can lead to shifting those vibrational peaks. These broad and shifting peaks are both benefited by the transfer learning process. The broad peaks allow them to not be computationally removed when the mathematical convolutions occur. In fact, these convolutions make the model less sensitive to peak shifting on the range of tens of wave numbers. These aspects make transfer learning promising for vibrational techniques. On the other hand, comparing mass peaks that are only a couple of mass units apart from each other are likely describing entirely different fragmentation patterns or isotopic ratios. MS data also has incredibly narrow peaks that can be missed entirely if they are low intensity during the mathematical convolutions. These factors likely are why transfer learning using Inception V3 was successful with the IR based data and unsuccessful with the MS data. Upon the determination that the LR based models performed better on correctly identifying new data compared to the CNN models we decided to use the LR based models for the remainder of this study.

3.3.3 Logistic Regression's Ability to Manage Specific Functional Group Classifications

The choice to switch to LR arose from wanting to utilize binary classifiers. By simplifying each model to a binary classifier, it is more feasible to fully explain the model output. Given our dataset, it is easier to optimize one model per functional group than one model predicting on all functional groups. For example, there are thousands of aromatic-containing spectra and less than 400 amide samples. This would impart artificial bias that would have to be mathematically manufactured to avoid. Training one model would likely lead to functional groups penalization because of less examples and ultimately not being identified as consistently or frequently.

There is a large variation in the final training and testing accuracies for each of the different functional group models. This is to be expected due to the large variation between the fragment fingerprint for each functional group. For a total 17/20 of the models had a final testing accuracy of over 70% and 13/20 of our models had a final test accuracy of over 75%. **Figure 22** shows the final training and test accuracies of all the models.

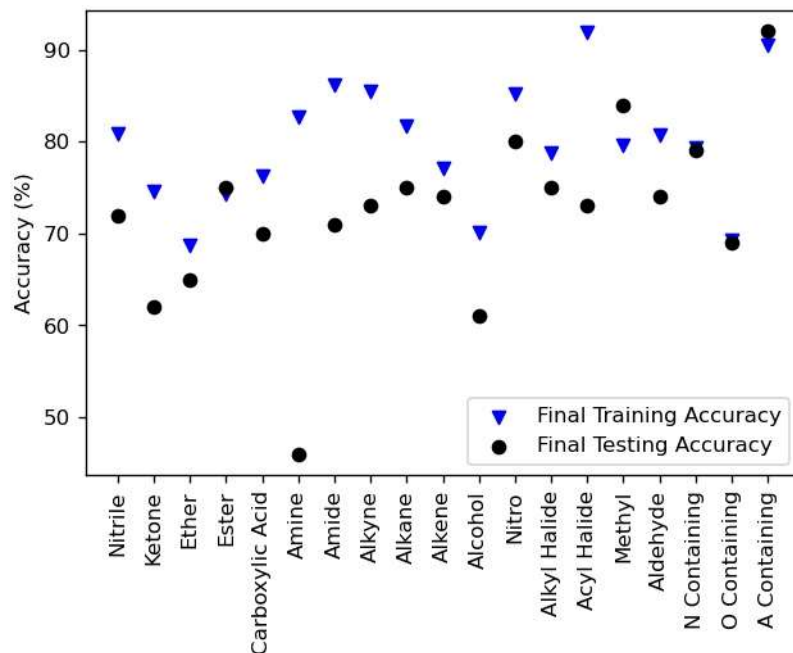


Figure 22. Scatter plot depicting all the final training and testing accuracies of each of the 20 different models. These final accuracies are highly variable with respect to the functional group that they are to be classifying.

The highest performing models, in terms of final testing accuracy, were the nitro, methyl, and the aromatic (A) containing models. This makes sense because with each of these models there are fragments that we can point to that would assist the model in its assignments. The nitro model could utilize the NO^+ and NO_2^+ fragments. The methyl model can look for the CH_3^+ ion and the A containing model can look for the loss of a benzene ring at 78 m/z.

Conversely, the poorest performing models are those of ketone, amine, and alcohol. These models likely struggle since the current methods of identifying these functional groups rely on looking at mass losses and looking for the products of secondary processes

including rearrangements and cleavages of certain areas of the molecule. These processes include α and β cleavages, McLafferty rearrangements, and radical losses among others.

Similarly, when looking at the generalized functional groups as well. The N containing and the O containing models both performed better than the O containing model. Albeit the O containing model still had a final testing accuracy of approximately 70%. This likely has to do with the fact that there is clear logic for identifying both odd numbers of nitrogen and aromatics in mass spectra. For the odd nitrogen spectra, we can look for odd numbered peaks suggesting the presence of nitrogen and we can look for a mass at 78 m/z to look for benzene, a common aromatic ring that shows up in organic molecules.

3.3.4 Identifying Mass Peaks that Guide Model Assignments

Feature selection and feature engineering are a common practices in the development of ML models and there are a large variety of methods to determine which features generate the best model outcomes.⁸⁵⁻⁸⁹ Feature selection differs from feature engineering, feature engineering works to reduce data dimensionality through convolving or creating statistical representations of the data through processes like principal component analysis or linear discriminant analysis among others⁹⁰ and feature selection works to reduce the raw data down to the most important features within.^{89,91,92} Both processes can be done manually or automatically via a statistical method.⁹³ Using feature engineering and feature selection processes provides different benefits to the modeling process.

To evaluate and explain the logic behind the model's assignments we looked at the coefficients that the model used in its final iteration. For each feature, in our case each mass, there is an associated coefficient describing the weight that that mass is used to determine if the functional group is present for that class. Positive peaks correlate to an increased likelihood that that functional group is present and negative peaks correlate to the increased likelihood that that functional group is absent. The larger the intensity, in either direction, the higher the correlation between that mass and the class that it is referring to. Figure 5 shows the overlapped final coefficients for both the generalized functional group models (**Figure 23A**) and the specific functional group models (**Figure 23B**)

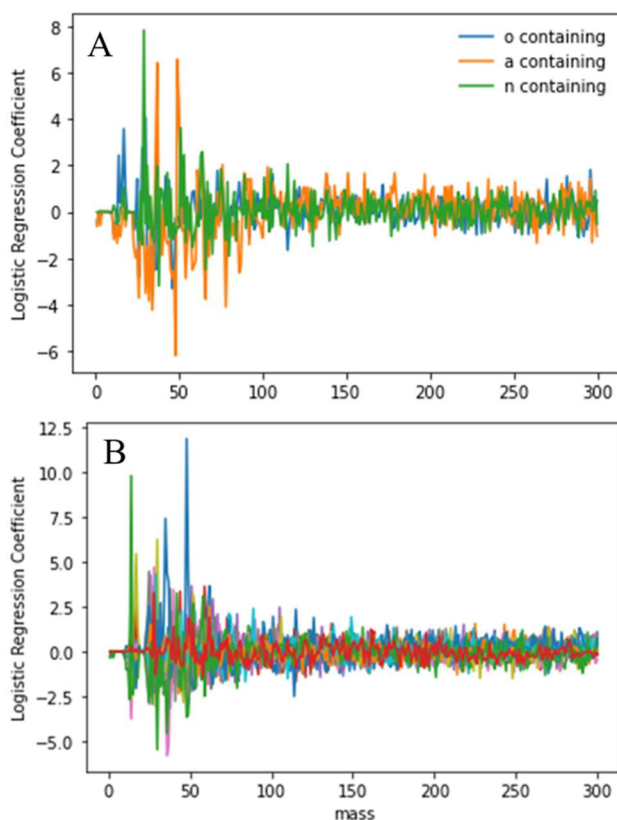


Figure 23. Model coefficients for each of the different trained models as a function of mass fragment. (A) Depicts the coefficients for the generalized functional group models and (B)

does the same with the specific functional group models. All coefficient plots for the individual models are presented in the SI.

When looking at all the aggregated coefficients it looks like the most impactful mass region to the analysis is below 100 m/z. This suggests that the model would perform similarly well if those were the only features given to the training set. This is an important conclusion that suggests that for this kind of analysis having a large mass range of available data is not necessary as long as the low mass range (less than 100 m/z) is thoroughly sampled.

To further understand how the different models were doing their assignments we developed a method to look at the impact of each peak on the final training and testing accuracies. To analyze the features, we trained each model 300 times, in each iteration of training one mass was removed. These final training and testing accuracies were compared to the accuracy of the model when it had access to all 300 mass units. This was used to identify peaks that were beneficial to the model's ability to identify functional groups and those that were hindering the models in making their assignments. The peaks that were beneficial led to a decrease in model accuracy when removed, the larger the discrepancy the more impactful the peak. On the other hand, peaks that were causing more false assignments, when removed led to an increase in model accuracy. Looking at the most beneficial peaks for the generalized functional group classification models leads to some interesting and promising results. **Table 2** has these values for the generalized functional group classifications.

Table 2. Mass values of the top 5 most impactful positively correlated peaks for each of the functional group generalized models. These were determined by comparing the testing accuracies of the model when it had access to all 300 mass units to when that mass unit of interest was removed. The % effect shown in the right-most column is negative because when those masses were removed the model experienced a reduction in the final testing accuracy. The mass values for the nitrogen containing model are all odd mass values and the mass values for the oxygen containing and the aromatic containing spectra are all even suggesting the utilization of the odd nitrogen rule without explicit training on that detail.

Functional Group Classification	Masses that Reduce Model Accuracy When Removed	
	Mass Value (m/z)	% Effect
A – Containing	78	-0.5
	42	-0.4
	66	-0.3
	50	-0.2
	68	-0.2
N – Containing	29	-2.5
	105	-0.7
	43	-0.6
	38	-0.5
	53	-0.4
O Containing	28	-0.4
	42	-0.4
	30	-0.6
	26	-0.8
	46	-1.0

In **Table 2** all the most impactful mass peaks for the N containing model are odd mass values whereas the most impactful peaks for the O and A containing models are even mass values. This suggests that even without explicitly “teaching” the model that there is an odd nitrogen rule the model was able to come to that conclusion on its own. We can also

look at the most impactful peak for the A containing model and see that it is 78 m/z which can be attributed to the mass fragment of benzene. However, we can also see that removing 78 m/z only leads to a 0.5% reduction in the final training accuracy of the model. This means that although there may be peaks that are important for assigning functional groups the model does not use a single peak or even a small set of peaks to make an assessment. The next step in our analysis shifted to the impacts of the number of available features on final accuracies.

3.3.5 Effects of Mass Range on Model Accuracy

To evaluate whether more data leads to higher accuracies for these models, we adjusted the dataset. We trained the models with 100, 300, and 500 mass units. We decided to reduce this to 100 mass units because the majority of the previously identified impactful mass fragments occurred at less than 100 m/z. We also increased to 500 mass units so that we can encompass more of the high mass range fragments. Both are compared to our 300 mass units' models for a basis.

For both decreasing the mass range from 300 m/z to 100 m/z and increasing the mass range from 300 m/z to 500 m/z we see an inconsistent response in the final testing accuracy with respect to the different functional groups. In **Figure 24** we observed no consistent trend in the mass range effect on final test accuracy. These results are consistent with what we observed in our analysis of the model coefficients suggesting that at mass ranges greater than 100 m/z the features aren't being as heavily utilized as they are at smaller masses.

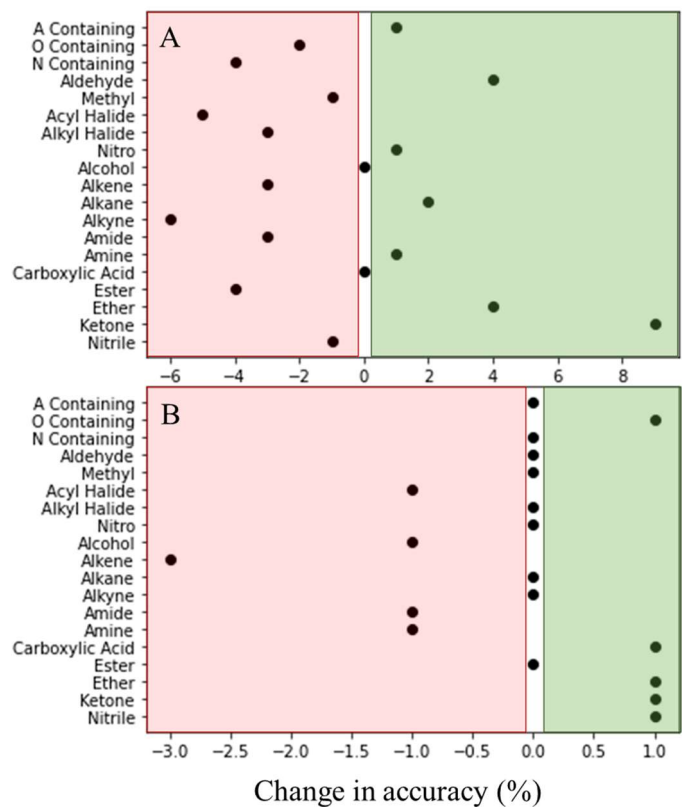


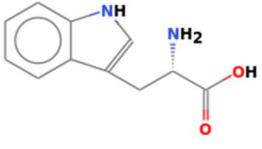
Figure 24. Scatter plots depict the effect of decreasing the utilized mass range (A) from 300 mass units to 100 mass units and increasing the utilized mass range (B) from 300 mass units to 500 mass units on the final testing accuracy of the models. The presence of points that are positive on the x axis (shaded in green, right-most box) show a net benefit in accuracy whereas a negative x value (shaded in red, left-most box) indicates a worsening accuracy.

3.3.6 Specific Examples of the Applications of this Approach

After exploring some of the parameters that affect the accuracy of these models, we then tested model success on a real-world application. When mass spectrometry data is returned, or downlinked, to Earth from planetary science missions, tens of thousands of mass spectra

may have been collected. Yet only a small subset of spectra may be scientifically significant. For example, a common target to identification of life is amino acids. To mimic this process, we examined the NIST mass spectrum of tryptophan to see if it would set off the correct models. **Table 3** shows the results of selected models on the ability to correctly identify the NIST spectra of Tryptophan. This example works to show how these tandem models may be beneficial in screening large amounts of data to look for specific spectra of interest.

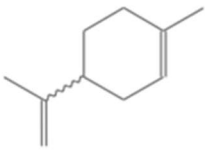
Table 3. Results of Selected Models on the Ability to Correctly Identify the NIST Spectra of Tryptophan

		
	Tryptophan	
	Presence In Molecule	Model Predicted
Carboxylic Acid	Present	Correct
Amine	Present	Correct
Aromatic	Present	Correct
Alcohol	Absent	Correct
Ketone	Absent	Correct
A Containing	Present	Correct
N Containing	Present	Correct
O Containing	Present	Incorrect

Aside from the O containing model, the LR model correctly predicted the present functional groups and functional group classifications for the tryptophan mass spectrum. This shows promise in these models being a useful tool for screening large numbers of mass spectra. In the example of planetary science missions, this process can be done onboard the spacecraft to help assist in the process of prioritizing spectra to downlink. It could also be used on data after it has been transmitted to prioritize spectral analysis.

To further benchmark the success of our models we also analyzed experimental data external to the NIST dataset. The spectra for limonene, pyridine, and 2 furan methanol were preprocessed in the same way as the NIST data to ensure that every mass had an associated intensity. These spectra were then presented to each of the models. **Table 4** shows the model assignments for limonene. Each of these compounds scored ~80% accurate for the identification of their functional groups. When making errors the models tended to overestimate the number of functional groups rather than underestimate.

Table 4. Results of the models on the ability to correctly identify experimental spectra of Limonene. The experimental spectra were preprocessed in the same way as the NIST data used for training.

		
	Limonene	
	Presence In Molecule	Model Predicted
Alkane	Present	Present
Alkene	Present	Present
Alkyne	Absent	Present
Methyl	Present	Present
Alcohol	Absent	Present
A Containing	Absent	Absent
N Containing	Absent	Absent
O Containing	Absent	Present

3.4 Conclusion

We present an investigation of multiple ML methods and parameters for mass spectral functional group analysis using minimal spectral preprocessing. Our results indicated that the CNN (Inception V3) did not perform as well as the LR models. We determined that the functional groups aromatics, nitro, and methyl are well defined though LR models, whereas alcohol, ketone, and amine functional groups are more difficult for LR models to define based on their fragmentation patterns alone. The most impactful peaks affecting model accuracy were determined by iteratively training models and removing one mass value in each model these results were echoed in looking at the final model feature coefficients. We observed that nitrogen-containing functional group models learn the odd-nitrogen rule. We evaluated the effect of mass range to determine if model accuracy is

improved with additional spectral information; these results vary between functional groups. Intuitively, EI fragmentation of small molecules will result in mass values below 100 m/z. Our model coefficients suggest that a mass range of 0 – 100 m/z is most beneficial for describing functional groups. The application of LR models to new sample mass spectra is evaluated on an example target molecule of interest, tryptophan as well as experimental data from outside of the NIST database. The success of these example analyses highlights the promise of ML approaches for screening a large volume of mass spectral data.

Future directions should further develop a methodology for approaching an ideal ML approach. For example, feature optimization for each model would achieve the highest possible final testing accuracy. Further validation of the models on experimental data outside of the NIST database is also necessary. Exploration of the LR method applied to other fragmentation patterns would enable implementation of generalizable ML more broadly in the field of mass spectrometry. The LR ML method explored herein provides a benchmark for application to space exploration, ultimately improving the analysis capabilities through identification of chemically interesting spectra.

Chapter 4. Saccharide concentration prediction from proxy sea surface microlayer samples analyzed via infrared spectroscopy and quantitative machine learning.

Solvated organics in the ocean are present in relatively small concentrations but contribute largely to ocean chemical diversity and complexity. Existing in the ocean as dissolved organic carbon (DOC) and enriched within the sea surface microlayer (SSML), these compounds have large impacts on atmospheric chemistry through their contributions to cloud nucleation, ice formation and other climatological processes. The ability to quantify the concentrations of organics in ocean samples is critical for understanding these marine processes. The work presented herein details an investigation to develop machine learning (ML) methodology utilizing infrared spectroscopy data to accurately estimate saccharide concentrations in complex solutions. We evaluated multivariate linear regression (MLR), K-Nearest-Neighbors (KNN), Decision Trees (DT), Gradient Boosted Regressors (GBR), Multilayer Perceptrons (MLP), and Support Vector Regressors (SVR) toward this goal. SVR models are shown to best predict the accurate generalized saccharide concentrations. Our work presents an application combining fast spectroscopic techniques with ML to analyze organic composition proxy ocean samples. As a result, we target a generalized method for analyzing field marine samples more efficiently, without sacrificing accuracy or precision.

4.1 Introduction

The sea surface microlayer (SSML) is a multifaceted, deeply complex region of the ocean.¹⁻⁷ As the interface between the Earth's atmosphere and ocean, the SSML performs

vital functions that affect climate^{5,8-10} and ice formation.^{4,11-13} Because of unique interfacial anisotropy,¹⁴⁻¹⁷ the physical and chemical properties of the SSML are of interest for their divergence from bulk water behavior. Generally, the SSML is enriched with lipids, proteins, and saccharides (also referred to as sugars or carbohydrates) which contribute to the total dissolved organic carbon (DOC).¹⁸⁻²² Understanding the chemical composition of the SSML provides insight into the biological activity and productivity within the SSML and enables predictions of cloud condensation⁹⁴ or ice nucleation,⁴ ultimately aiding climatological models.⁹⁵⁻⁹⁸ Recent analyses of saccharide concentrations in SSML have shown concentrations of about 500 nM from eight unique compounds.²⁰ The dynamic nature and chemical complexity of the SSML make monitoring the region difficult, and yet increasingly necessary.

For the described work, glucose and sucrose were chosen as analytes of interest as they are two of the most abundant saccharides found in ocean samples.⁹⁹ This approach focuses mainly on the quantification of saccharides due to their importance in many marine processes. For example, saccharides are common feedstocks for the ocean ecosystem^{100,101} and can contribute globally to atmospheric processes such as cloud nucleation through transport from the SSML into aerosols.^{99,102} Understanding a generalized saccharide concentration is important to understanding the total ocean chemical diversity and ecosystem health through these processes.

The presented work is motivated by the need for fast, accurate analysis of SSML samples to establish a method that enables exponentially more SSML chemical measurements. Traditional methods to analyze SSML samples are typically limited to mass

spectrometry,^{5,103,104} which requires extensive organic, solid-phase extraction processes. Nevertheless, these methods have provided invaluable information on SSML (and sea spray aerosol) chemical composition. To reduce the sample preparation process and expedite analysis of results, we developed methods that utilize infrared (IR) spectroscopy methods, specifically, attenuated total reflectance Fourier transform infrared (ATR-FTIR) to estimate the saccharide concentration via machine learning (ML) implementations. IR methods provide information on chemical composition and concentration by probing the vibrations of chemical bonds, rather than relying on mass fragmentation. Identification and quantification of specific chemical classes from IR spectra is carried out by analyzing peaks characteristic to specific chemical bonds.¹⁰⁵ We note that the limit of detection for ATR-FTIR spectroscopy is higher than for mass spectroscopy, however the speed of analysis for this method is superior.

ML provides a unique avenue to explore relationships among data that cannot be otherwise deduced. The applications to improve or expand chemical systems via ML are broad and present throughout all chemistry fields. Materials design,^{106,107} novel drug discovery,^{108,109} catalyst optimization,^{110,111} and clean energy production^{112,113} are some of the many fields where knowledge has expanded because of ML. Advances in molecular dynamics in combination with machine learning have also paved the way for bridging the connection between molecular structure and physical characteristics.^{114,115} Recent work emphasizes the improved application of FTIR spectroscopy, and more broadly vibrational spectroscopy, for qualitative and quantitative assignment, especially when combined with ML models.^{116,117} Takamura and colleagues explored methods to identify donor biological

sex from urine samples.¹¹⁸ They presented several ML applications, including partial least-squares discriminant analysis with and without a genetic algorithm, to explore the chemical information contained in their FTIR spectra. They found that the increased computational complexity of an artificial neural network resulted in comparable results to their discriminant analysis model's predictive power. Butler and coworkers presented successful use of support vector regressors (SVR) in predicting brain cancer from ATR-FTIR spectra.¹¹⁹ Their high-throughput approach featured high sensitivity and specificity in the prediction of benign versus malignant samples.

SVRs have also been employed in classification of Raman spectra to identify Alzheimer's Disease in mice; a relevant features map is utilized to identify pertinent peaks that are from molecules known to be associated with the disease. A study from 2022 reports comparable classification accuracy of microplastic Raman microscopy samples from k-nearest neighbors (KNN), multilayer perceptron (MLP), and random forest (RF) models.¹²⁰ These literature examples highlight the diverse applications of ML and develop techniques that expand the applications of chemistry, as we present herein.

This work utilizes ML methods of increasing complexity to evaluate the training data and investigate new data, including field samples with unknown composition. The utilized models in this work are multivariate linear regression, K nearest neighbors, decision trees, gradient boosted regression, multilayer perceptron, and support vector regressors. This diversity in model approach explores the effects of computational complexity, i.e. single models vs ensemble models, and a variety of regression solving techniques.

Fitting data to a linear regression model is common for absorbance data, such as fitting to the Beer-Lambert Law to determine physical constants or identify concentrations of unknown samples.¹²¹ Absorbance FTIR spectra generally follow a linear relationship of intensity with respect to concentration, which is advantageous for determining new sample composition. Recent work has utilized multiple linear regression to identify heavy metals, including investigating the effect of surface chemistry on vanadium¹²² and lead¹²³ toxicity. However, the simplicity of the method ultimately restricts the model's usefulness in more complex, dynamic systems. The largest difference between Beer-Lambert Law linear regression and multivariate linear regression is that all features (in this work, wavenumbers) are used simultaneously to make the multivariate model's assignments.¹²⁴ This multivariate linear regression (MLR) will act as a benchmark that can be used to compare the other listed models to.

In contrast, SVR fits training data to the best function by minimizing the distance of each value from the fitting equation to be able to predict continuous values. Not all data is appropriate for SVR, but in cases where concentration is being predicted and is linearly correlated with absorbance, it can be a well-suited model. A 2020 report by Mohammadi and colleagues presented an application of SVR to predict different functional group fractions in crude oil.¹²⁵ As another example, ATR-FTIR and SVR were employed by Chen et al. 2022 to predict bio-oil characteristics quickly.¹²⁶

The work described herein provides a discussion on an improved approach to monitoring the SSML. We explore ML approaches to achieve precise and accurate quantitative analysis of simplified proxies of glucose and egg serum albumin (ESA).

Glucose is used as our saccharide proxy for training data as it is commonly observed in field measurements and saccharides are frequently reported as a concentration of glucose.^{103,127,128} We also use ESA in our training set because ESA, our SSML protein proxy, has been shown to have surface activity and form insoluble monolayers on aqueous interfaces, despite being a water soluble protein.^{129–131} While an unlikely protein to find in field samples, ESA provides a complex matrix of amino acids that are abundant in the ocean's water column.^{5,7,132–134} The use of ML in conjunction with vibrational spectroscopy enables greater exploration of chemical space and identifying connections between data. Our results present, to our knowledge, a first account of predicting saccharide concentration from FTIR spectra of ocean proxy samples using ML.

4.2 Methods

4.2.1 Training Solution Preparation, Data Collection, and Data Preprocessing

All chemicals were used as received and all solutions requiring water were prepared using ultrapure water (18 mΩ) from a MilliQ system. For Simplified Proxy (SP) training spectra, stock solutions of 1M glucose (Sigma Aldrich, ≥99.5% (GC)) in ultrapure water and 5 mg/mL egg serum albumin (ESA) (Sigma Aldrich, 62-88%, agarose gel electrophoresis) in ultrapure water were prepared. The solution matrix was produced by dispensing the relevant amount of each stock solution via auto pipette and diluting with the requisite amount of water. Briefly, we selected this system and concentrations to have reasonable complexity.

Both the protein and saccharide have IR absorbances from 1800 to 900 cm⁻¹. The peaks were well resolved, with minimal convolution. Inorganic salts were excluded in our

matrix, but we provide spectra of the O-H stretching region in the SI to emphasize the limited effect that they have on the IR spectra. Concentrations were selected based on literature precedent from field study results.^{97,98,104} Solutions were measured in triplicate via ATR-FTIR spectroscopy (PerkinElmer Spectrum 3) with a single beam KRS-5/diamond ATR assembly. Spectra were acquired in the “SingleBeam” mode without the use of a continuous reference and were detected using a liquid nitrogen cooled HgCdTe (MCT) detector over 32 scans (approximately one minute) from 4000 to 450 cm^{-1} with a resolution of 1 cm^{-1} . Spectra were converted to absorbance with a water-only background spectrum (R_0) using the established relationship of $-\log(R/R_0)$. Baseline correction was done using a linear fit model to correct for inconsistent baseline between measurements. Water-only backgrounds were obtained every 5 sample measurements. Triplicate measurements were used as individual spectra, rather than an average of the three, to provide more machine learning training and testing data (**Figure 25**).

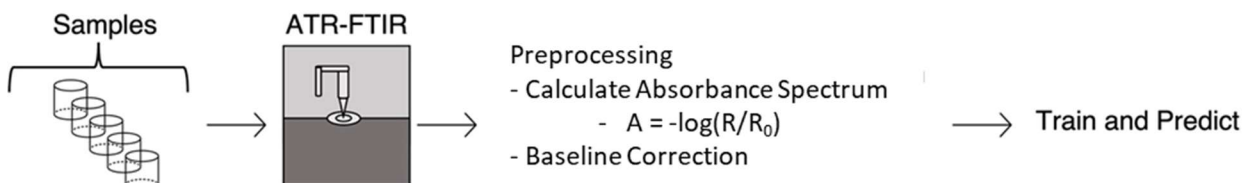


Figure 25. Schematic flow chart of data collection process to the ML pipeline.

4.2.2 Lab Generated Simplified and Ocean Proxy Sample Preparation and Sampling

To test the models' accuracies with increasing chemical complexity, ocean proxy (OP) samples were made in the lab with a greater diversity of chemical constituents than the simplified proxies. For these test data, stock ocean proxy-solution was prepared to

have 0.1 M sucrose (Sigma Aldrich, $\geq 99.5\%$ (GC)), 0.1 M glucose, 0.5 mg/mL ESA, 3.323 mg/mL bovine serum albumin (BSA) (Sigma Aldrich, $\geq 98\%$, heat shock fraction, pH 7), and 0.1 M 1-butanol (Sigma Aldrich, 99.9%) (**Table 5**). Two additional solutions were prepared via dilution of the stock. The higher concentration dilution was 7.5 mL of stock and 2.5 mL of water and the lower was 5 mL of stock and 5 mL of water. The three solutions were analyzed using the data collection and preprocessing described above.

Table 5. Concentrations of all species in the lab-made ocean proxy samples for evaluation of model accuracy on more chemically diverse conditions.

	<i>Ocean Proxy</i> <i>A</i>	<i>Ocean Proxy</i> <i>B</i>	<i>Ocean Proxy</i> <i>C</i>
<i>Concentration of Sucrose (M)</i>	0.10	0.075	0.05
<i>Concentration of Glucose (M)</i>	0.10	0.075	0.05
<i>Concentration of Saccharide (M)</i>	0.20	0.15	0.10
<i>Concentration of ESA (mg/mL)</i>	0.50	0.38	0.25
<i>Concentration of BSA (mg/mL)</i>	3.32	2.49	1.66
<i>Concentration of 1-Butanol (M)</i>	0.10	0.075	0.05

4.2.3 Machine Learning Methods

All machine learning (ML) methods were implemented using Python scripts and SciKit-Learn packages. These are available online at:

https://github.com/Ohio-State-Allen-Lab/Saccharide_Quantification_2024.

4.2.4 Preprocessing

All data, which includes the entire training set of simplified proxy (**SP** – containing only ESA and glucose) and the ocean proxy (**OP** – containing ESA, glucose, BSA, and 1-

butanol) samples were standardized using the SciKit-Learn StandardScaler function. This function subtracts the mean of each feature (wavenumber) and divides each feature by the respective standard deviation. The StandardScaler function was first fit using only the SP data, then this fit was applied to both the SP and OP datasets. This was done to avoid the StandardScaler function using the SP dataset information in the OP samples. If the StandardScaler function was fit on the SP and OP datasets together, it would incorrectly inflate the final ability of these models to identify the OP concentrations.¹³⁵ After standardization, the OP data was separated from the data that would then be used for training. The data was then split 70::30 into training and validation/test sets. The latter of which was then split 50::50 into validation and testing datasets. A random state was set to split the data the same way every time into the training, validation, and test datasets to ensure consistency. The training and validation sets were used to train each of the models (210 spectra for training 45 for validation). The withheld test data (45 spectra) were then used to further explore the models' accuracy on previously unseen data that was similar to the data the models were trained on.

A total of 6 machine learning methods were utilized in this work. They will be described here in order of increasing computational complexity.

4.2.5 Multivariate Linear Regression (MLR)

In MLR, all features are fit with a hyperplane in which the dimensionality is determined by the number of features and each feature is has associated weights. This hyperplane is then used to identify concentrations of new samples in the same way that a line would be used for regression with only one feature. Multiple linear models including

Lasso, ElasticNet, and Orthogonal Matching Pursuit were tested, but the best performing estimator was the Ridge regressor. This method tends to perform well when there are a large number of features compared to the number of spectral samples.¹³⁶

4.2.6 K-Nearest Neighbors (KNN)

KNN is a method of supervised learning that uses the proximity of previously explored data to make predictions by looking at the distance (the calculation of this distance is variable depending on model parameters) between the neighbors and the training datapoint and using that to adjust the predictions.¹³⁷ In this work, we use the default Minkowski metric for distance which calculates the standard Euclidian distance between points in multivariate space. Different numbers of neighbors between 2 and 10 were tested and the model performed the highest when 5 were used.

4.2.7 Decision Trees (DT)

DTs work to separate the large dataset into smaller pieces repeatedly based on optimized features to be used as split points.¹³⁸ These smallest components, or leaves, then are used to identify predictions for new data. The model utilized in this work terminated splitting once two features were unable to be split further. The model then worked to minimize squared error between training predictions and true values. The original splits were randomized.

4.2.8 Gradient Boosted Regression (GBR)

GBR is an example of an ensemble algorithm that allows for the use of many smaller models, in this context, decision trees.¹³⁹ This method is more computationally complex than a single DT and can identify more complex patterns. The model presented

here utilizes a Huber loss function and 2,000 estimators with a learning rate of 0.5 and a max depth of 1.

4.2.9 Multilayer Perceptron (MLP)

MLP is an example of an artificial neural network, a framework of interconnected nodes referred to as neurons.¹⁴⁰ Each neuron has associated weights, which are adjusted with each training step through a mathematical process of backpropagation. The model presented in this work uses a tanh activation function, an Adam solver, and 500 training steps.

4.2.10 Support Vector Regression (SVR)

SVR utilizes the power of high dimensionality data to identify patterns.¹⁴¹ By transforming the data into a higher dimensionality space, it allows for the fitting of the model with different mathematical approaches. The kernel describes the transformation used to transform the data into the high dimensionality hyperplane. This model utilizes a radius bias function (RBF) as the kernel for fitting the dataset.

4.2.11 Model Analysis

To evaluate the models after training, error was also calculated at three different places within the training and testing process. The error calculated is root mean squared error (RMSE). First, the RMSE for the training data is evaluated by comparing the predicted values to the true values with each model. This describes how well the model was able to fit the training data. Next, the validation error was calculated to predict the model's accuracy on new data. Finally, the testing error focuses on the ability of the model to evaluate data that it has not previously been exposed to.

We also evaluate the prediction of the models on the OP samples to determine how well they perform on data that is chemically different than the data that the models were trained on. This is done by determining the estimation accuracy by comparing the amount of saccharide predicted by each model compared to the true combined saccharide concentration. If the model exactly predicts the concentration, this amount would be 100%. Scores of less than 100% and more than 100% represent under and over prediction respectively. This highlights the degree and directionality of the prediction error in the final estimates of OP data.

4.3 Results and Discussion

4.3.1 Evaluating Feasibility of Using IR Spectra to Quantify Saccharide Concentration

The chemical complexity of SP and OP samples is explored with ATR-FTIR spectroscopy and quantitative ML approaches to develop a simple and accurate method of analysis. The FTIR spectra provide chemical information about the sample components and their concentrations, which have a linear correlation with absorbance. The correlation diverges from a linear relationship at high absorbance values, which is not of concern in the presently studied concentration ranges. A single figure containing all the acquired spectra is presented in the SI (**Figure S3**). Glucose has many vibrational modes that can be used for analysis (**Figure S4**).

Heat maps can be used to visualize the SP dataset in its entirety. The data was sorted with respect to the concentration of glucose and then plotted against the wavenumber and the intensity at that wavenumber for a given spectrum. This allows for the visualization of

the entire dataset in the context of changing glucose concentration and is presented as a heat map in **Figure 26**. A band of increasing intensity can be seen between 1200 and 1000 cm^{-1} correlating to the increasing concentration of glucose in solution, specifically with the C-C and C-O vibrational modes. The presence of this band supports the ability of the machine learning models to have representative features that will allow for the concentration analysis of glucose.

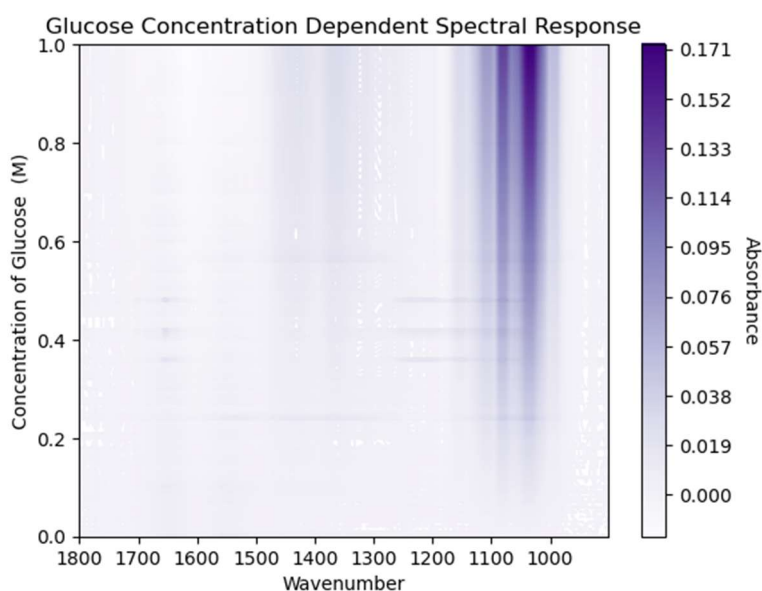


Figure 26. Heat map of the ATR-FTIR dataset as sorted by the concentration of glucose (0 – 1 M). The band of intensity growing in between 1100 and 1000 cm^{-1} corresponds to the increasing C-O stretching within the IR fingerprint region from the increased concentration of glucose. We do not see a strong spectral signature for the ESA relative to

that of glucose also in solution (0 – 5 mg/mL) where we would expect the amide bands to exist between 1700 and 1500 cm^{-1} .

To evaluate each model's ability to accurately predict within the training dataset, model accuracy will be calculated for the training on the simplified proxy (**SP**) dataset. This SP dataset contains only glucose (the analyte of quantification) and ESA (the chemical matrix). To explore if the models are able to expand outside of the explicit training, these models will then be tested on the ocean proxy (**OP**) dataset. Beyond the ESA and glucose within the SP dataset, the OP dataset also contains sucrose, BSA, and 1-butanol. Each of the model's predicted values will be compared to the additive concentration of glucose and sucrose to make a generalized saccharide concentration (**Figure 27**).

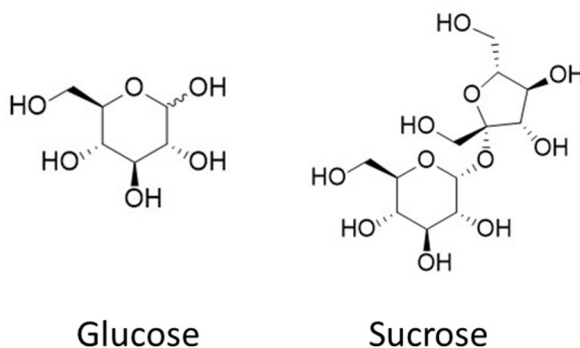


Figure 27. Molecular structures of both glucose (left) and sucrose (right). Both saccharides contain similar vibrational bonds and vibrational environments in regions of the structure. The simplified proxy (SP) dataset contains only glucose and egg serum albumin whereas the ocean proxy (OP) dataset contains both glucose and sucrose in solution with egg serum albumin, bovine serum albumin, and 1-butanol.

4.3.2 Evaluating Machine Learning Models' Fit of the Simplified Proxy (SP) Dataset

After training, the accuracy of each model's ability to identify the concentrations of the test and validation sets was evaluated to explore the influence of the chosen model to evaluate the SP dataset through analyzing the RMSE error. Ideally, there wouldn't be any effect and the error would be consistent regardless of concentration range. **Figure 28** visualizes these results. DT (**Figure 28. C**) had the smallest associated RMSE and did not exhibit an increased error in low concentrations. KNN, GBR, MLP, and SVR (**Figure 28. B, D, E, and F respectively**) all experienced increased error at low concentrations. R^2 values for each model have also been calculated and are presented in the SI (**Table S12**).

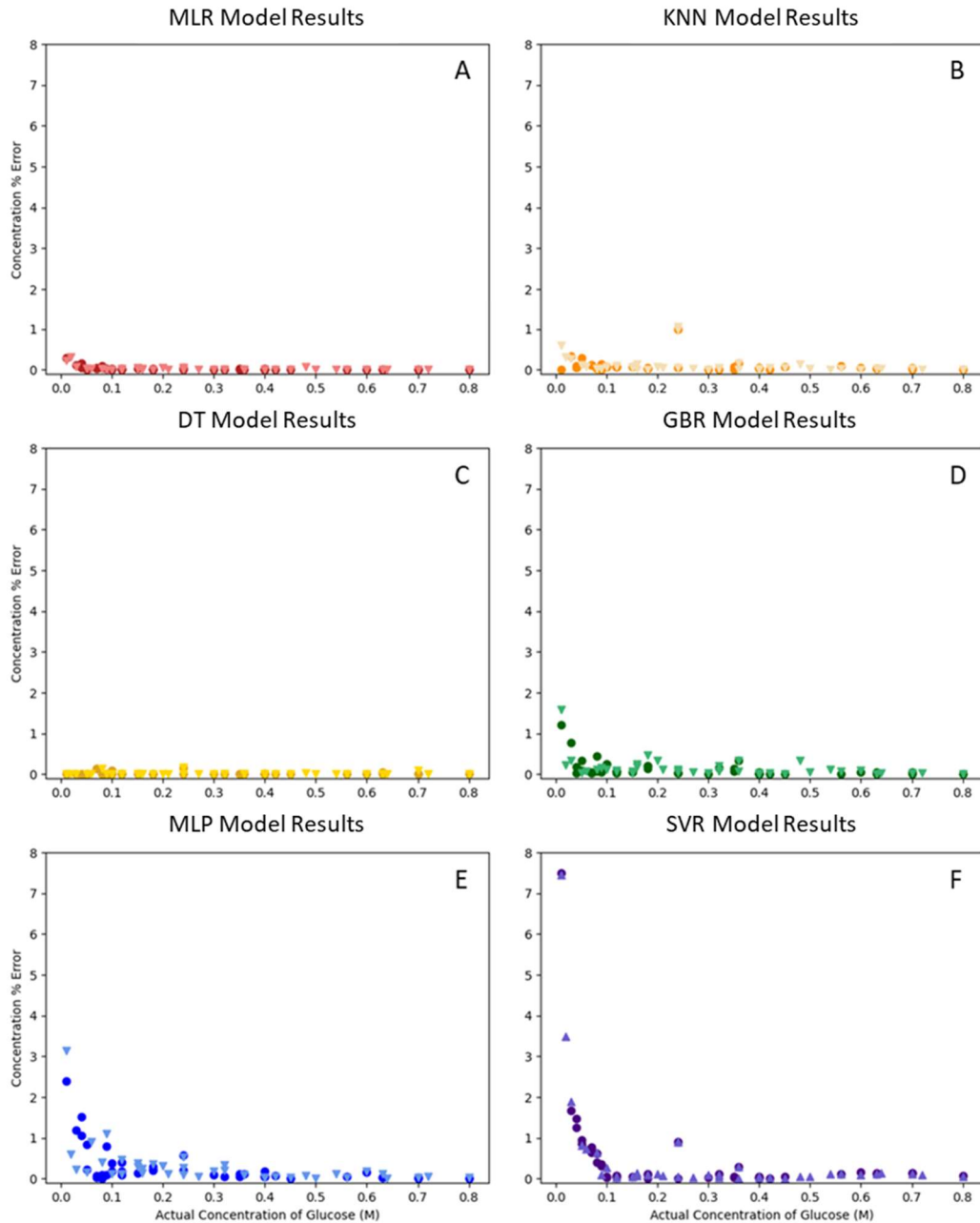


Figure 28. Scatter plots depicting the accuracy of each of the utilized machine learning models on the simplified proxy (SP) dataset. The y-axis represents the difference between the model assigned and the actual concentrations of the testing dataset divided by the actual concentrations multiplied by 100% (circles) and the withheld validation dataset (triangles).

The gradient boosted regression, multilayer perceptron, and support vector regression models do experience an increased error at low concentrations.

To perform a more in-depth error analysis, each of the model's RMSE was calculated between each step of the training by evaluating the training, validation, and test sets' final accuracies. All of the models had smaller than 70 mM in error amongst the different steps. These results have been visualized in **Figure 29**.

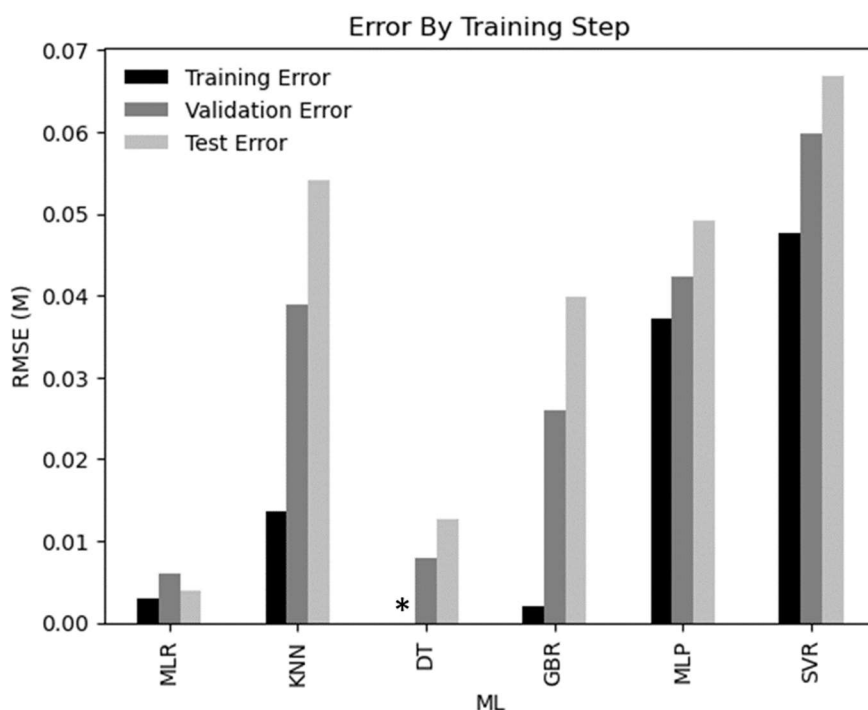


Figure 29. Bar graphs depicting the associated root mean squared error (RMSE) in each part of the training process for the simplified proxy (SP) dataset. All models have a final testing error of less than 0.07 M, but the MLR performed the best in this evaluation. The asterisk indicates that for the decision trees the training error was 0.00 M.

4.3.3 Evaluating Machine Learning Models' Fit of the Ocean Proxy (OP) Dataset

The saccharide concentrations of the OP samples were then estimated using these same ML models. The “true” saccharide concentrations are defined as the sum of the concentrations of glucose and sucrose. This additive concentration, coupled with the increased complexity of the matrix extends these proxies beyond the chemical space that the models were originally trained on. For the purpose of identifying a generalized saccharide concentration, it is important to select for the models with the highest estimation accuracy when comparing the estimated and true concentrations without disproportionately valuing low or high concentration samples. A model performing poorly here doesn't suggest that the model is poorly trained, just that it doesn't have the capacity to generalize that far beyond the training. For example, MLR had the lowest RMSE error in validation and test datasets as seen in **Figure 29** for the simplified proxies. The MLR, however, only has an estimation accuracy of 50-60% on the OP data, underestimating the combined saccharide concentration by approximately half. This suggests that the MLR model is highly fit to glucose and does not generalize to sucrose, which for other chemical contexts would be ideal.

The highest accuracy in identifying the combined saccharide concentrations came from the SVR and GBR models. They were both able to assign 2/3 of the solutions within 20% of the true concentration of combined saccharide. SVR showed less spread in its predictions but tended to overestimate. The lowest concentration of saccharide was not correctly identified but also existed outside of the range of concentrations where SVR was

performing well (**Figure 28**). GBR did not consistently over or underestimate, but it had a large spread in prediction accuracy. These results are shown in **Figure 30**.

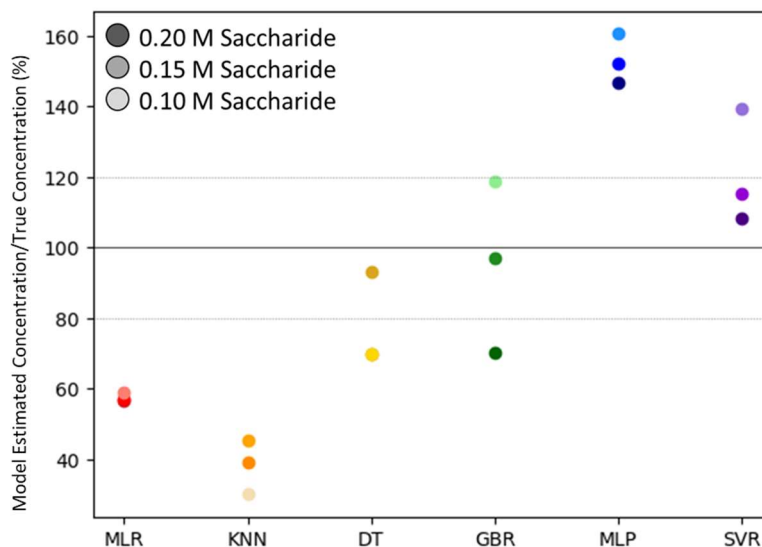


Figure 30. Predicted concentration divided true concentration of combined saccharide for ocean proxy (OP) saccharide concentrations. Solid line at 100 represents 100% meaning that the predicted concentration equals the predicted concentration. The dotted lines represent +/- 20%. The darkest markers in each column represent the highest concentration of saccharide in OP (0.20 M) and the lightest represent the least concentrated (0.10 M). The models have varied levels of success at identifying samples that are far removed from the original training set. The highest performing models were GBR and SVR.

4.3.4 Summary of Discussion

Our quantitative results indicate SVR and GBR are the most promising models to explore for identifying concentrations of saccharides within ocean samples. They are both able to estimate the combined saccharide concentrations within 20% for 2/3 of the complex OP samples. This accuracy would likely be increased if the data that the models were trained on were more chemically similar to the OP dataset as that training would be more relevant to the OP data.

4.4 Conclusions

To develop efficient, less-expensive analytical techniques for analysis of the SSML, several ML methods were applied to ATR-FTIR spectra and used to determine saccharide concentration and chemical composition of aqueous samples. Our results indicate that SVR and GBR models are viable for complex solutions, especially considering the training sample data is relatively simple. The research presented herein provides a unique approach to studying the contributions to the DOC and as a result the SSML utilizing the advanced computational tools available and reduces the time needed to perform analyses of marine samples. Further work should focus on finding an optimal training data set, investigating quantifying other organic concentrations, and intercalating other spectroscopic or spectrometric data, to name a few. An improved understanding and quantification of the marine organics is achievable, wherein more frequent measurements and analysis can occur, ultimately providing more information about the productivity of the marine organics and thus their effects on our atmosphere and climate.

Chapter 5. Multi Analyte Concentration Analysis of Marine Samples Through Regression Based Machine Learning

Marine systems are incredibly chemically complex. An understanding of the chemical compounds that make up the chemical diversity in these samples is critical to understanding ecological and ocean health metrics. Using Raman spectroscopy in tandem with machine learning combines a low-cost highly transportable analytical technique with a powerful and rapid computational approach that can aid in marine analysis. Here we use Raman and machine learning to identify concentrations of three chemically relevant compounds in three distinct classes. Saccharides are represented by glucose, fatty acids by butyric acid, and proteins are represented by amino acid proxy through glycine. Eight machine learning models (gradient boosted regressors, random forests, histogram gradient boosted regressors, decision trees, k nearest neighbors, support vector regression, multilayer perceptrons, and multivariate linear regression) were tested for their accuracy in identifying the concentrations of glycine, glucose, and butyric acid in marine samples. Support vector regression was able to best identify all three concentrations of glycine, butyric acid, and glucose. Butyric acid was similarly well described through gradient boosted regression and histogram gradient boosted regression. In this work Raman, though it has a lower sensitivity than mass spectrometry, can still be used to identify mM concentrations of organics in complex aqueous matrix. The described methodology has the potential to significantly advance rapid field analysis of marine samples.

5.1 Introduction

Marine organic composition drives many of the methods in which the ocean interacts with the other chemical systems of Earth. The organic compounds have the ability to influence atmospheric chemistry when partitioning to the surface of the ocean and contributing to sea spray aerosols.^{102,142–149} They can also act as feedstocks or markers of metabolism within biological systems like algal blooms.^{150–152} Measuring marine organic compounds also improves the ability to detect and remediate potential marine disasters like oil spills.^{153,154} The largest challenges in attaining a large scale understanding of ocean chemistry arises from the incredibly diverse array of compounds present in marine samples.^{148,149,155–157}

Vibrational spectroscopy has been used extensively in the literature to describe marine chemistry and aqueous environments.^{147,158–160} Raman spectroscopy has been used, in particular, in deep ocean probes due to the durable instrumentation and ability to analyze aqueous environments without major disruption from the vibrational signature of water itself.^{161,162} Raman spectroscopy has also been used in the past as a method to identify chemical markers to understand physical properties (e.g. chemical kinetics, thermochemistry, and chemical building blocks) and biologic activity in marine systems.^{158,163} In our prior work, we have used Raman spectroscopy to identify ion pairing in aqueous solutions of NaCl and KCl.⁴³ Ion pairing is detected by observing how the vibration of water is affected by being in different solvation shells of ions. Detecting ion pairing requires a high signal-to-noise ratio as these interactions may only make small perturbations in the OH symmetrical and asymmetrical stretching regions of the spectra.

Utilizing machine learning in tandem with vibrational spectroscopy has been of high interest in recent years,^{164,165} particularly in the areas of real-time reaction monitoring¹⁶⁶ and medical diagnostics.^{51,167–169} The vibrational fingerprints of different analytes of interest are proving to be powerful features for machine learning models. De Medeiros Back and colleagues published a paper in 2022 describing methodology utilizing vibrational spectroscopy to identify microplastics in the Mediterranean Sea. They found that support vector classification showed the best performance out of the machine learning methods that they evaluated. Our group has also successfully utilized machine learning and vibrational spectroscopy to identify organic concentrations in complex chemical matrices.¹⁷⁰ In the prior work, attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR) data was used to evaluate the ability of six different machine learning methods to identify concentrations of glucose in a complex matrix of differing concentrations of egg serum albumin. Ultimately, we found that support vector regression had the highest accuracy in identifying glucose. To further analyze the extent of the expandability of the training, more chemically complex samples (containing sucrose, glucose, egg serum albumin, bovine serum albumin and 1-butanol) were created. It was found that the model could not only identify the concentration of glucose alone, but a sum concentration of saccharide (glucose and sucrose).

Here, we expand upon our results in our previous work by evaluating three different analytes' concentrations simultaneously, rather than just one. Our three analytes of interest have been curated due to their relevance and impact on marine chemistry and the marine ecosystem. We evaluate a total of eight machine learning models to this end. Each of these

models was trained using two different datasets. The first dataset, the spiked lab (**SL**) sample dataset, is created with ultrapure water and spikes of various concentrations of the three analytes. This dataset focuses on giving the models access to highly resolved calibration spectra with little matrix effect. The second dataset, the spiked marine (**SM**) sample dataset, utilizes the same spikes as in the previous dataset but instead of using ultrapure water, unspiked marine (**UM**) samples are used. This dataset provides real-world samples and works to highlight the effect of the matrix (salts, other organics, potential fluorophores) as well as secondary chemical effects of the complex chemical environment. This work presents, to the best of our knowledge, the first multi-output machine learning models to understand the organic components of ocean chemistry quantitatively.

5.2 Methods

5.2.1 Solution Preparation

Butyric acid, glycine, and L-histidine were obtained from MilliporeSigma and glucose was obtained from Sigma Aldrich. All compounds have a purity of higher than 98%. 1 L of solution was made with each analyte compound at concentrations of 303, 262, 145, and 300 mM for glucose, butyric acid, histidine, and glycine respectively with ultrapure water (Milli-Q Advantage A10, resistivity 18.2 M Ω). These stocks were used as the spikes to make the SL and SM datasets as described in the Results and Discussion section.

5.2.2 Raman Spectroscopy

Raman spectra were collected using a custom-built Raman spectrometer. This instrument contains a diode-pumped 532 nm CW laser containing built-in laser line (± 0.5

nm) and polarization filters (>100:1) (CrystaLaser). The excitation source is directly coupled to a custom-built fiber-optic polarized Raman probe system (InPhotonics) allowing 235 mW power at the sample with a spectral range of 90–4200 cm^{-1} . The output, both polarized and depolarized scattered light, is collected by two independent fiber-optic terminated ports. The two polarization output ports are fiber coupled directly to a spectrograph through a 50 μm slit with a 1200 g mm^{-1} grating with a 750 nm blaze, which is calibrated to Ar/Ne emission lines (IsoPlane 320, Princeton Instruments), and is detected with a liquid-nitrogen cooled CCD detector (Pylon, 1340 \times 400 pixels, Princeton Instruments). Each 200 μm core fiber is directly coupled to the spectrograph and allows for the simultaneous collection of the perpendicular (HV, depolarized) and parallel (VV, polarized) spectra. Measurements of all of the concentrations were performed at a room temperature of 21 ± 2 $^{\circ}\text{C}$. Spectra were collected by signal averaging 50 frames each with a 0.4 s integration time. Only the parallel (VV, polarized) spectra were used for analysis.

5.2.3 Paper Spray Ionization Mass Spectrometry (PSI-MS)

The mass spectrometry (MS) method used herein for all calibrations consisted of a paper spray ionization platform which has been utilized as a valuable ambient ionization MS method for direct, targeted, and rapid analysis of analytes within a native sample.^{171,172} In PSI-MS sample is deposited directly onto an untreated Whatman one paper triangle substrate produced in-house. All samples (i.e. SL and SM samples) were deposited on the paper substrate (10 μL sample size) and allowed to dry completely before the application of methanol extraction solvent. Ionization was facilitated by the application of a high DC voltage (6 kV) to the ionization apparatus, thus inducing an electrospray ionization

mechanism from the paper substrate. Methanol extraction solvent was applied directly onto the paper substrate with the paper triangle secured from the rear via a copper clip. Paper substrates were held at a 5 mm distance from the inlet of the mass spectrometer which was held at 250 °C. Spectra were recorded over a total acquisition time of two minutes with 0.25 minutes analyte and internal standard averaging for all calibrant and UM solutions. The MS was operated in positive-ion mode for butyric acid, glycine, and histidine analytes with analysis of protonated pseudomolecular ion and negative-ion mode for glucose for analysis of the chloride adduct pseudomolecular ion. Protonation of butyric acid was facilitated via the high DC voltage ionization mechanism.¹⁷³ Protonation of glycine and histidine was assisted via addition of 0.1 % formic acid. Glucose chloride adduct formation was assisted via addition of 10 mM ammonium chloride.¹⁷⁴

Mass spectra were obtained using a Thermo Fisher Scientific Finnigan ion trap mass spectrometer (San Jose, CA). All MS parameters were held constant throughout with 3 microscans and 100 ms injection time. All spectral averaging was performed for 0.25 min. Tandem MS was performed via collisional induced dissociation (CID) for structural analysis using collision energies ranging from 20-40 manufacturer's units and were optimized for each unique chemical system. Data processing was performed using Thermo Fisher Scientific Xcalibur 2.2 SP1 software.

5.2.4 Mass Spectral Quantification - Internal Standard Calibration Curve

Using the PSI-MS platform, we sought to quantify each analyte in the UM samples and constructed internal standard calibration curves (**Figure S9**). This was done using standard solutions of each analyte made in neat water (13-100 mM) with appropriate

internal standards (800 mM). We placed 50 μ L of the prepared internal standard solution into 2 mL of the standard solution to prepare a 16 mM solution for analysis. We then took 10 μ L aliquot of the 16 mM solution and place this on the paper triangle, allowing for 1 minute of dry time before extraction solvent application onto the paper and applying a 6 kV high DC voltage for subsequent analysis in the positive ion mode (butyric, glycine, and histidine) and negative ion mode (glucose). Tandem MS (MS/MS) mode was implemented for analysis, using the appropriate transitions for each compound and its corresponding internal standard (**Figure S9**). We monitored the ratio of the intensity of the analyte-to-internal standard (A/IS) as a function of the analyte concentration – consistent with MS based calibration. **Figure S9** shows the linearity achieved with R^2 values that fall within the 0.99 range. With these results, we moved forward with the quantitative analysis of the selected compounds using the PSI-MS set-up with UM samples. Under analogous conditions to calibration, the UM samples were analyzed, and their spectrum confirmed the presence of butyric acid, glucose, glycine, and histidine in the ocean water samples via MS/MS.

5.2.5 Field Collection for SM and UM Samples

Water was collected from two locations in Cocoa Beach, Florida in January 2023. Sampling site one was the Atlantic Ocean and site two was the Banana River within the Indian River Lagoon System. The Banana River is a brackish waterway connected via ocean inlet with mangrove shorelines; the conditions provide a unique aqueous environment on the west side of the Florida barrier islands. Samples are categorized as sea surface microlayer (**SSML**) and bulk sea/river water (**BW**). We operationally define the

SSML as the top 1 mm of the sampled water and BW as the top 1 m of the sampled water. All samples were stored at room temperature and shipped; once received, samples were stored at 2°C until analyzed.

BW samples from Cocoa Beach, Florida were collected. Briefly, sea samples were collected within 10 meters of the ocean shoreline (28.314885 N, 80.607818 W) and river samples were acquired approximately 2 meters from land (28.309917 N, 80.614893 W) on January 10th and 11th 2023. All samples were collected and stored in mason jars with plastic lids instead of the traditional metal lids to avoid contamination through metal corrosion.

BW was collected by first copiously rinsing a jar, replacing the lid, submerging the covered jar, and finally removing the lid underwater. Jars were filled to avoid head space. SSML water was collected according to methods detailed by Harvey and Burzell.¹⁷⁵ Briefly, a clean hydrophilic glass plate (Millipore Sigma, unframed, H × W × D 200 mm × 260 mm × 4 mm) was submerged perpendicular to the surface to about the top inch, the plate was then withdrawn from the water at a rate of approximately 20 cm/s. Adsorbed water and organics were collected via silicone squeegee into a copiously rinsed glass jar.

5.2.6 Data Preprocessing

There was a large degree of observed Raman fluorescence in the SM and UM sample datasets. This presented itself as broad band elevated baselines (**Appendix A – Figure S1**). Fluorescence was expected from the large number and variety of naturally occurring organic compounds in solution. Multiple methods of preprocessing were evaluated to see if this baseline variation could be corrected and if that correction led to higher model accuracies. To ensure that all data was treated the same way, all preprocessing

was completed on the SM , UM, and the SL datasets even though the fluorescence was not observed in the SL data (**Appendix A – Figure S1**). The highest accuracies came from taking the average of the Raman spectra from 1283 to 2640 cm^{-1} . This average was then subtracted from all intensities from 346 to 3117 cm^{-1} . This baseline corrects some of the observed Raman fluorescence in the SM and UM dataset. Next, the entire spectrum is normalized with respect to 3343 cm^{-1} which is correlated with the isosbestic point between the symmetric and asymmetric O-H stretching bands. This further corrects for the fluorescence.

After preprocessing, the data was then split into training, testing, and validation datasets in ratios of 70:15:15 respectively. A random state, a variable within the sklearn train test split function, was assigned to ensure that the data was split the same way for each Jupyter Notebook, so all the models have access to the same data in the same splits. The 15 validation spectra were removed, in part, to ensure that when we performed a sample dropout test, the difference in accuracy could be associated directly with the sample's representation in the dataset and not to the size of the dataset analyzed. This dropout test ensures that the models weren't simply using the dilutions of the field samples to make their assignments.

5.2.7 Python Scripts

All python scripts have been made available via Jupyter Notebooks on GitHub (https://github.com/Ohio-State-Allen-Lab/multi_compound_marine_regression).

5.2.8 Regression Methods

Eight total regression methods were tested for accuracy in identifying the concentrations of the UM samples. Six of these models were evaluated in our previous work on the saccharide and egg serum albumin dataset. The remaining two were added once it was seen that ensemble algorithms were performing well on the SM and SL datasets.

Decision Trees (DT)¹³⁸

Decision trees (DT) utilize iterative binary splits of the data to identify concentrations of new data. A fitting criterion of absolute error was used with a best splitter to separate the data into leaves that had a minimum of 5 samples.

Random Forest (RF)¹⁷⁶

Random forests (RF) utilize many decision trees to improve model accuracy. In this context, 100 trees were trained independently of each other (non-bootstrapped) by minimizing squared error. All of the trees were then used simultaneously to make model assignments.

Gradient Boosted Regression (GBR)¹³⁹ and Histogram Gradient Boosted Regression (HGBR)¹⁷⁷

Gradient boosted regression (GBR) and histogram gradient boosted (HGBR) models are made similarly to random forest models in the fact that the base architecture is a decision tree. However, the difference is that as new trees are trained in GBR, models learn from the previous trees. For this context, 100 trees are used reducing a loss of squared error. A learning rate of 0.5 was used with a max depth of 1. HGBR utilizes a histogram estimator to improve the speed of computation.

K Nearest Neighbors (KNN)¹³⁷

K nearest neighbors (KNN) models utilize the distance from previous datapoints to estimate quantifications for new samples. The presented models utilize the 5 nearest neighbors to make their assignments.

Support Vector Regression (SVR)¹⁴¹

Support vector regression (SVR) models work to optimize high dimensionality hyperplanes to fit datasets with many features. The kernel being utilized in the presented models is a radial bias function.

Multi-Layer Perceptron (MLP)¹⁴⁰

Multi-Layer Perceptron (MLP) models are examples of neural networks. These models utilize a combination of weights and biases that exist in pairs called neurons. These neurons are tuned throughout training steps to minimize error. The presented models were trained for 5,000 training iterations, with a rectified linear unit (ReLU) activation function and an Adam solver.

Multi-Variate Linear Regression (MLR)¹³⁶

Multi-variate linear regression models fit each feature (in this case, each wavenumber) with a linear function. The function for every feature is used simultaneously to make model assignments. The presented models use a Ridge linear model to fit the features.

5.3 Results and Discussion

5.3.1 Selection of Representative Analytes

Describing the vast chemical complexity of ocean samples in just a few analytes of interest is incredibly challenging. This current work aims to focus on a saccharide, a fatty

acid, and a proxy for proteins. Due to time and technique constraints, only one to two representatives could be chosen for each class of molecule. As for concentration range, the total concentration sum was < 300 mM,¹⁷⁸ arising from estimated total organic carbon (TOC) for marine samples. This average varies globally depending on marine system, time of year, and local ocean productivity.^{179–182} This adds the constraint that the analytes of interest should be soluble in room temperature water at a concentration of close to 300 mM.

Marine proteins vary greatly with type and size. These variables make defining the concept of a total concentration challenging. To standardize and simplify this analysis, this study looks at amino acids rather than a specific protein. Glycine and histidine were chosen as analytes of interest. These amino acids have been defined as potential markers for gluconeogenesis (non-sugar metabolism) and antifungal properties among others.^{183–186} Glycine has also been noted as partitioning into sea spray aerosols and being transported into cloud water.¹⁸⁷ Amino acids have been reported to make up 11% by mass of the dissolved organic carbon within submicron sea spray aerosol particles.¹⁸⁸ Note although histidine was chosen as a representative analyte it was not found to be in UM samples above the LOQ of the utilized mass spectral calibration and thus could not be analyzed through our Raman and ML combined approach.

For fatty acids, the analyte needed to be marine relevant and not have a strong partitioning to the aqueous surface. This second criterion limited the options to fatty acids with a carbon chain length of three or less. Butyric acid has a carbon chain length of three and has been noted as one of the most abundant short chain fatty acid in algal bloom

metabolic processes.^{150,151,189,190} Butyric acid can also be an indicator of ocean oxygenation.¹⁹¹ As algal bloom populations collapse, the dissolved oxygen is depleted, causing negative impacts to ocean health.^{192,193} This lack of oxygen also increase ocean acidification.¹⁹⁴

The chosen analyte representative for saccharides is glucose. This saccharide is one of the most abundant of the saccharides in marine systems.⁹⁹ It is also a common feedstock for small scale marine life like algae and has been used in the past as a biomarker of algal bloom presence and stage.^{100,101} Glucose, along with other saccharides have also been known to partition into aerosols^{99,102} where they can act as potential ice nucleators.¹⁹⁵

5.3.2 Sample Organization

After selection of the analytes of interest, a methodology was developed to make unique combinations of organic concentrations to generate the datasets for training. Four distinct calibration curves, two with 10 datapoints and two with 5 datapoints, were utilized in the method. The calibrations and the sample combinations that are developed make up a single array and each dataset contains two sample arrays. Each sample array contains 50 samples. This is done in different ways for the SL samples and the SM samples (**Figure 31**).

For the SL samples, the first sample array has anti-correlated calibration curves (**Figure 31** SL samples rows 7-11). This means that the analyte concentration gradients on opposite sides of the sample array are changing inversely to one another. This ensures that the models are penalized for trying to correlate any of the concentrations during the training. For the second sample array, the opposing analyte concentrations change

proportionally to one another (**Figure 31** SL samples rows 12 – 16). This second array was to ensure that there wasn't an inverse correlation that could be picked up by the model either.

Spiked Lab (SL) Samples											Spiked Marine (SM) Samples											
	A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J		
7	28 mM			Glucose Calibration Curve						101 mM	100 mM	1	2:18 Ocean Water to Lab Water Dilution									
8	13 mM			Anti Correlated Spike Combinations							Histidine Cal.	2	5:15 Ocean Water to Lab Water Dilution									
9												3	10:10 Ocean Water to Lab Water Dilution									
10											Glycine Cal.	4	15:5 Ocean Water to Lab Water Dilution									
11	48 mM	87 mM	Butyric Acid Calibration Curve							24 mM		5	18:2 Ocean Water to Lab Water Dilution									
12	76 mM			Glucose Calibration Curve						19 mM	19 mM	7	Glucose Calibration Curve									
13	16 mM			Correlated Spike Combinations							Histidine Cal.	8	Anti-Correlated Spike Combinations									
14												9										
15											Glycine Cal.	10										
16	76 mM	65 mM	Butyric Acid Calibration Curve							16 mM		11	Butyric Acid Calibration Curve									

Figure 31. Sample organization for model training datasets. The SL sample dataset (I) contains two sample arrays one in which there are anti-correlated concentrations (the species on opposite sides of the array have inverse calibration curves), and in the second the calibration curves move in the same direction. The SM sample dataset (II) contains first a dilution series of the field samples to ensure that the calibration curves were done lower than the concentration of the UM samples and then an anti-correlated array of spikes. The row numbers show the solution array being used 1-5 is dilutions, 7-11 is anti-correlated calibration curves, and 12-16 is the correlated calibration curves. Not pictured: 6 represents the UM samples that are withheld as the final validation set for the trainings.

For the SM samples, the setup involved associating each column of the sample arrays with a marine sample (

Table 6). This allowed for a dilution series to be made for the first sample array (**Figure 31** SM samples rows 1-5). Due to the UM samples already having unique concentrations this dilution series took the place of the anti-correlated sample array in the SL dataset. The second sample array contained the same organic spikes that the correlated calibration data of the second SL sample array (**Figure 31** SM samples rows 7-11). Together, these ensured that the concentrations of the UM samples would be within the calibration.

After analysis of the marine samples through mass spectrometry, it was determined that that the marine sample concentrations of histidine were below the limit of quantification (LOQ) of our mass spectral calibration. This suggests that the marine concentrations are in the μM range or below and thus would be beneath the limit of detection for our Raman system. As a result, the histidine spikes are in the samples and are part of the solution prep, however they are not represented in the analysis as there is no “true” value to compare to model results for accuracy.

Table 6. Marine samples associated with the UM and SM datasets. Concentrations of glycine, butyric acid, and glucose were calculated through mass spectrometry and will be used as the “true” values of concentration for these samples. Histidine concentrations were all beneath the LOQ for the mass spectral method.

SAMPLE COLUMN	WATER SAMPLING LOCATION	GLYCINE (mM)	BUTYRIC ACID (mM)	GLUCOSE (mM)	HISTIDINE (mM)
A	Atlantic Ocean - BW	6.01	26.81	14.20	<LOQ
B	Banana River - SSML	2.94	22.23	6.37	<LOQ
C	Banana River - BW	1.24	26.26	12.95	<LOQ
D	Atlantic Ocean - SSML	<LOQ	21.82	20.19	<LOQ
E	Atlantic Ocean - BW	11.27	48.11	29.85	<LOQ
F	Saltwater Aquarium - BW	3.61	25.91	11.74	<LOQ
G	Atlantic Ocean - SSML	3.79	23.31	10.94	<LOQ
H	Banana River - BW	6.65	21.72	40.57	<LOQ
I	Banana River - SSML	2.23	24.82	14.90	<LOQ
J	Atlantic Ocean - BW	8.95	21.92	17.82	<LOQ

The concentration combinations within the spike-containing sample arrays are created by taking the row and column and spiking those concentrations into either lab or marine water depending on the dataset (**Figure 32**). This generates 50 unique combinations for each sample array.

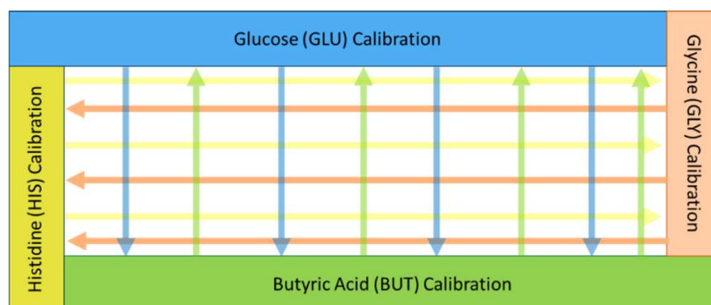


Figure 32. Diagram depicting the process of mapping the calibration curves to make unique combinations of concentrations for each spike-containing sample array.

After training all the models, initial assessments on internal accuracy were made.

Figure 33 depicts all the error associated with each chemical species (glycine, butyric acid, and glucose) for each of the machine learning methods trained on SL data (left) and SM data (right). The errors associated with the SM models are, on average, higher than the models trained on the SL data. Within each set, the ensemble methods (GBR, RF, and HGBR) tend to perform better than the single models. There doesn't tend to be an immediately visible trend between error and chemical species, suggesting that different models are able to optimize different chemical species more effectively.

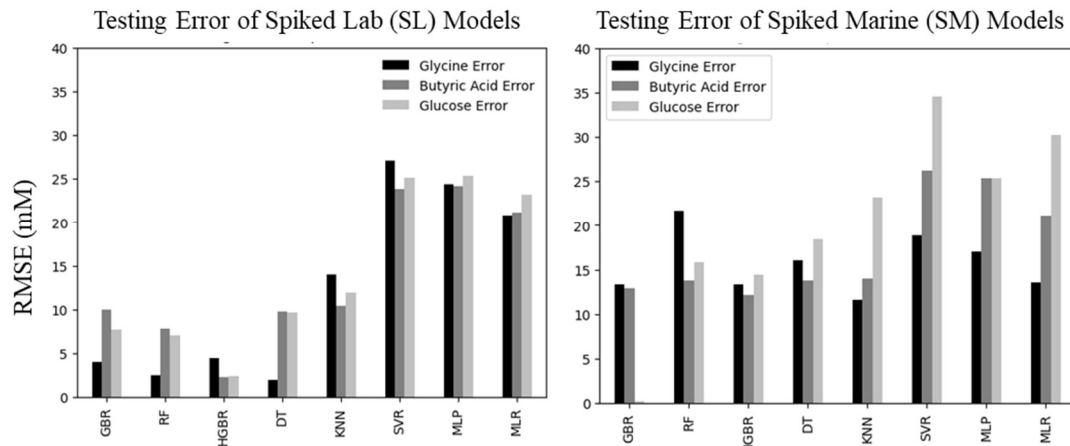


Figure 33. Test stage root mean squared error (RMSE) values for each combination of ML approach and chemical species.

All the models can then be used to predict the concentrations of the UM samples. **Figure 34** has the model assignments for each of the different compounds. The models trained with the LS data are on the left (circles) and the models trained with the SM data are on the right (triangles). The dotted lines show a boundary of +/- 20% of the highest concentration of that analyte in a single marine sample. The models trained on the SM models show much more clustering of assignments within this +/-20% region. It is also possible to identify models that are performing more poorly across the board these include MLP, MLR, and RF.

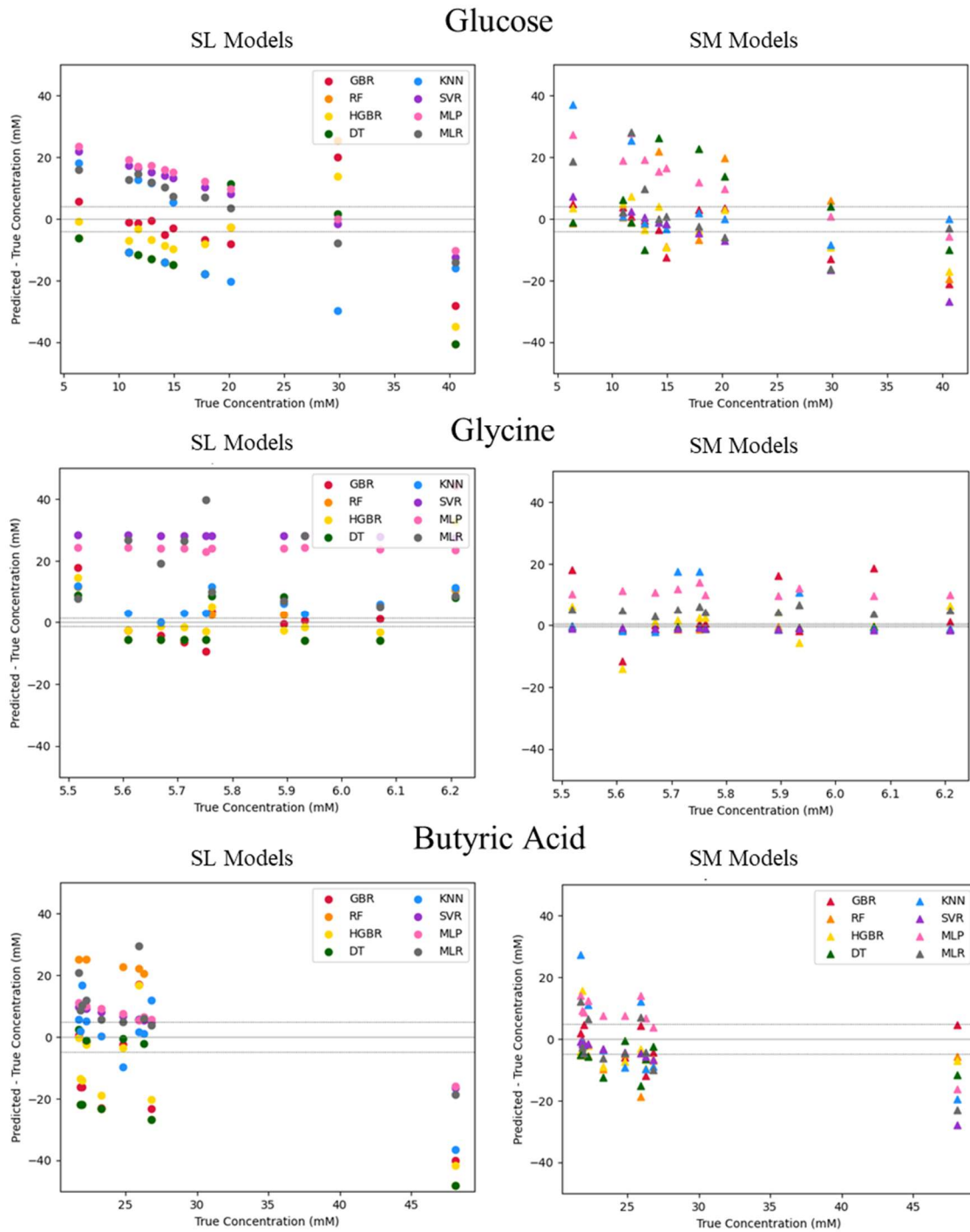


Figure 34. UM sample estimates from each ML approach on SM models (left - circles) and on SL models (right - triangles). Solid grey line denotes a difference between

actual and predicted concentrations of 0. The dotted lines represent +/- 20% of the most concentrated marine sample for the given chemical species (glucose, glycine, and butyric acid). The SM models show more clustering within these boundaries than the SL models suggesting that the SM models were more accurate at identifying the concentrations within the UM samples.

To improve the visualization of the models that are making assignments in the +/- 20% range, the number of assignments in this region were counted for each model and for each compound (**Figure 35**). This confirms that the SM models perform better than the LS models at identifying the UM samples. This increase in accuracy likely comes as a function of the increased similarities between the training data and the final validation data. These similarities include non-analyte organics which are leading to the observed Raman fluorescence and the solvated salts. These variables likely change from sample to sample, but they work to make the training data more chemically similar to the final validation data.

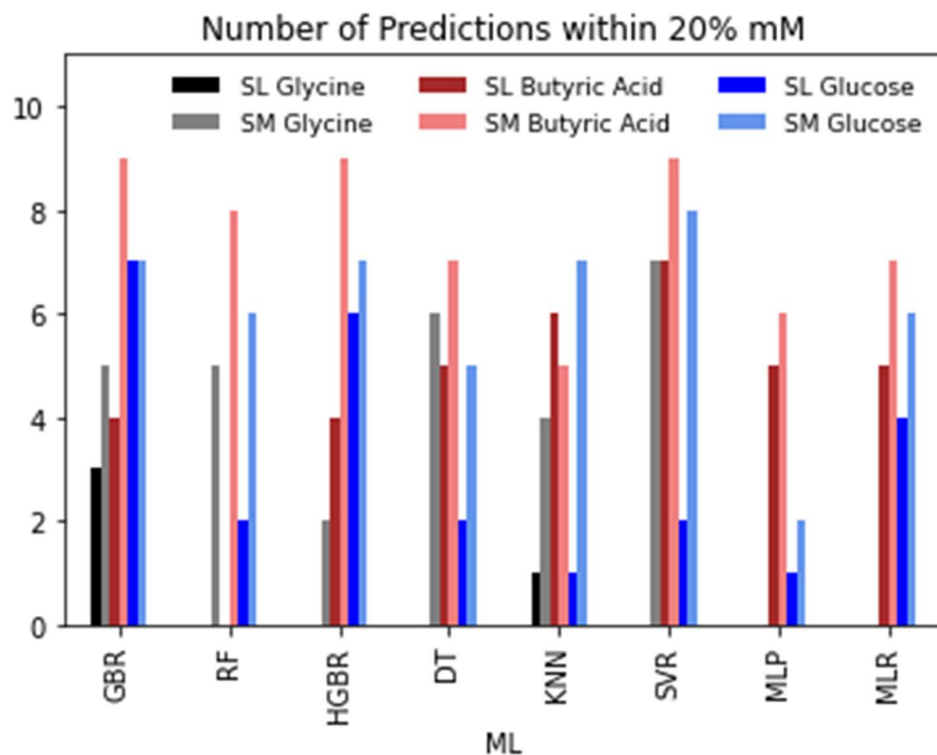


Figure 35. Counted values out of 10 for the correctly quantified UM samples within 20% of the max true values in a single UM sample. These counts are separated by ML approach and chemical species. Importantly, the SM models perform higher than the SL models in nearly every case. SVR achieved the highest accuracies for all three analyte concentrations.

SVR performed the best at identifying the concentrations of glycine, butyric acid and glucose assigning 7/10, 9/10, and 8/10 within 20% of the true value, respectively. Butyric acid was also well described through the GBR and HGBR methods (

Table 7).

Table 7. Highest performing models for each analyte compound.

Highest Accuracy Model for Each Analyte Compound		
Glycine	Butyric Acid	Glucose

Support Vector Regression (SVR)	Support Vector Regression (SVR) Or Gradient Boosted Regression (GBR) Or Histogram Gradient Boosted Regression (HGBR)	Support Vector Regression (SVR)
7/10 Marine Samples	9/10 Marine Samples	8/10 Marine Samples

As mentioned in **Figure 31**, the SM sample models (highest performing) are trained on dilutions of the marine samples. To further analyze the accuracy of these models, it is important to measure the marine sample accuracy if the model hadn't been trained on dilutions of that exact sample. To accomplish this, the highest performing models (**Table 7** –SVR (for glycine, butyric acid, and glucose), HGBR (for butyric acid), and GBR(for butyric acid) were trained another 10 times each. For each model, training one column of marine samples was dropped, (e.g., SM samples: column A) then the model was evaluated using the UM sample associated with that marine sample (for column A: sample A6). This allows for the analysis of the model if it was shown a truly new marine sample. This was then repeated with each of the remaining columns independently. **Figure 36** shows these results.

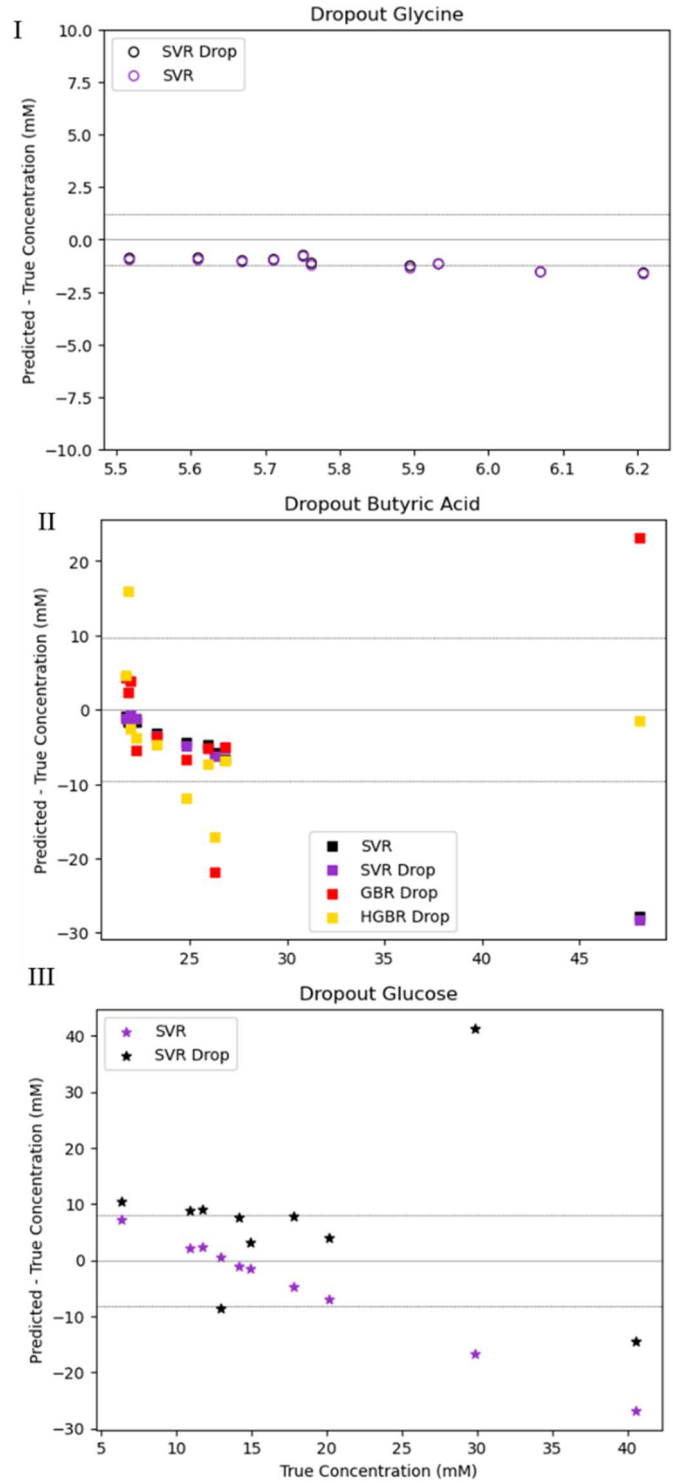


Figure 36. Marine sample analysis using dropout sample method. For each model training one column of samples was dropped (e.g., SM samples column A) then the model was

evaluated using the UM sample associated with that marine sample (for column A: sample A6). The dropped sample results are in black or grey and the original analysis is left in the color associated with that ML approach in Figure 34. The accuracy of models is well maintained for glycine (I) and butyric acid (II). The largest loss in accuracy was in the measurement of glucose. This variance, due to it mostly being overestimates, may be associated with the presence of other saccharides in these samples that cannot be determined using the stated mass spectral method.

The accuracy of analysis with and without sample dropout is maintained well in analyzing glycine and butyric acid. SVR performed the best out of the three possible butyric acid models in the drop out test. The model was able to achieve accuracy for 9/10 samples even with the sample dropout. GBR (8/10 correct) and HGBR (7/10 correct) both experienced reductions in accuracy in identifying concentrations of butyric acid while using sample dropout. The largest variance was found in the analysis of glucose where there is a trend in over estimation from the SVR model (**Table 8**). This perpetual overestimation may suggest that there are other saccharides in these field samples.¹⁷⁰ Our “true” value for glucose is limited to only glucose based on the limitations of our mass spectral method, which can only evaluate one stated analyte at a time. Machine learning models can accurately identify a generalized saccharide concentration through a sum of glucose and sucrose;¹⁷⁰ this is consistent with the vast majority of errors being positive, as observed here (**Figure 36**).

Table 8. Effects of dropout sample test on highest performing models for each analyte compound.

I		Before Sample Dropout		
Analyte	ML Approach	No. Estimates Below 20% Threshold	No. Estimates Within 20% Threshold	No. Estimates Above 20% Threshold
Glycine	SVR	3	7	0
Butyric Acid	SVR	1	9	0
	GBR	1	9	0
	HGBR	0	9	1
Glucose	SVR	2	8	0
II		After Sample Dropout (Net change)		
Analyte	ML Approach	No. Estimates Below 20% Threshold	No. Estimates Within 20% Threshold	No. Estimates Above 20% Threshold
Glycine	SVR	3	7	0
Butyric Acid	SVR	1	9	0
	GBR	1	8 (- 1)	1 (+ 1)
	HGBR	2 (+ 2)	8 (- 2)	1
Glucose	SVR	4 (+ 2)	4 (- 4)	2 (+ 2)

5.4 Conclusion

Eight machine learning models were tested for their ability to identify four different analyte concentrations in a complex marine matrix. Two different Raman spectral datasets of organic spike arrays were made on ultrapure water and on marine samples to approach the complex system in different ways. The results indicate that support vector regression had the highest accuracy in identifying all three analytes. Butyric acid was also well described through gradient boosted regression and histogram gradient boosted regression however these approaches performed more poorly than the support vector regression during the sample drop out test. In nearly every case the spiked marine (SM) dataset, in which the spikes were added to marine samples with their internal chemical complexity,

outperformed the spiked lab (SL) dataset. Upon testing sample dropout to remove potential internal correlation in concentrations from the dilution series making up half of the SM dataset, it was found that butyric acid and glycine were largely unaffected. When this dropout approach was used with glucose, it led to an increase in overestimating glucose concentrations which suggests that there are saccharides in solution that are contributing to the same vibrational modes. Future work should add samples to the SM dataset to help improve its stability, to lessen the reliance on any given marine sample. Other experimental methods to benchmark and confirm the concentrations of the marine samples with additional representative analytes should be developed to improve the scope of the “true” concentrations. Other complementary analyte models should also be added to improve the overall organic compositional analysis. With sufficient analyte models it may also be possible to look for correlations between analyte models which would suggest which compounds may lead to systematic errors when coexisting in solution. This work reveals that it is possible to achieve accurate estimates of selected organics in an increasingly complex chemical matrix using Raman spectroscopy with machine learning. This combination of Raman and ML stands to improve our rapid response and characterization of marine samples both in the lab and in the field due to the durability and transportability of Raman instrumentation and the ease of use and rapid computations of a pretrained machine learning model.

Bibliography

- (1) Carlson, D. J. Dissolved Organic Materials in Surface Microlayers: Temporal and Spatial Variability and Relation to Sea State. *Limnol. Oceanogr.* **1983**, *28* (3), 415–431. <https://doi.org/10.4319/lo.1983.28.3.0415>.
- (2) Cunliffe, M.; Engel, A.; Frka, S.; Gašparović, B.; Guitart, C.; Murrell, J. C.; Salter, M.; Stolle, C.; Upstill-Goddard, R.; Wurl, O. Sea Surface Microlayers: A Unified Physicochemical and Biological Perspective of the Air–Ocean Interface. *Prog. Oceanogr.* **2013**, *109*, 104–116. <https://doi.org/10.1016/j.pocean.2012.08.004>.
- (3) Engel, A.; Bange, H. W.; Cunliffe, M.; Burrows, S. M.; Friedrichs, G.; Galgani, L.; Herrmann, H.; Hertkorn, N.; Johnson, M.; Liss, P. S.; Quinn, P. K.; Schartau, M.; Soloviev, A.; Stolle, C.; Upstill-Goddard, R. C.; van Pinxteren, M.; Zäncker, B. The Ocean’s Vital Skin: Toward an Integrated Understanding of the Sea Surface Microlayer. *Front. Mar. Sci.* **2017**, *4* (MAY), 1–14. <https://doi.org/10.3389/fmars.2017.00165>.
- (4) Chance, R. J.; Hamilton, J. F.; Carpenter, L. J.; Hackenberg, S. C.; Andrews, S. J.; Wilson, T. W. Water-Soluble Organic Composition of the Arctic Sea Surface Microlayer and Association with Ice Nucleation Ability. *Environ. Sci. Technol.* **2018**, *52* (4), 1817–1826. <https://doi.org/10.1021/acs.est.7b04072>.
- (5) Cochran, R. E.; Laskina, O.; Trueblood, J. V.; Estillore, A. D.; Morris, H. S.; Jayarathne, T.; Sultana, C. M.; Lee, C.; Lin, P.; Laskin, J.; Laskin, A.; Dowling, J. A.; Qin, Z.; Cappa, C. D.; Bertram, T. H.; Tivanski, A. V.; Stone, E. A.; Prather, K. A.; Grassian, V. H. Molecular Diversity of Sea Spray Aerosol Particles: Impact of Ocean Biology on Particle Composition and Hygroscopicity. *Chem* **2017**, *2* (5), 655–667. <https://doi.org/10.1016/j.chempr.2017.03.007>.
- (6) Ault, A. P.; Moffet, R. C.; Baltrusaitis, J.; Collins, D. B.; Ruppel, M. J.; Cuadra-Rodriguez, L. A.; Zhao, D.; Guasco, T. L.; Ebben, C. J.; Geiger, F. M.; Bertram, T. H.; Prather, K. A.; Grassian, V. H. Size-Dependent Changes in Sea Spray Aerosol Composition and Properties with Different Seawater Conditions. *Environ. Sci. Technol.* **2013**, *47* (11), 5603–5612. <https://doi.org/10.1021/es400416g>.
- (7) Bertram, T. H.; Cochran, R. E.; Grassian, V. H.; Stone, E. A. Sea Spray Aerosol Chemical Composition: Elemental and Molecular Mimics for Laboratory Studies of Heterogeneous and Multiphase Reactions. *Chem. Soc. Rev.* **2018**, *47* (7), 2374–2400. <https://doi.org/10.1039/c7cs00008a>.
- (8) Abraham, J. P.; Baringer, M.; Bindoff, N. L.; Boyer, T.; Cheng, L. J.; Church, J. A.; Conroy, J. L.; Domingues, C. M.; Fasullo, J. T.; Gilson, J.; Goni, G.; Good, S. A.; Gorman, J. M.; Gouretski, V.; Ishii, M.; Johnson, G. C.; Kizu, S.; Lyman, J. M.; Macdonald, A. M.; Minkowycz, W. J.; Moffitt, S. E.; Palmer, M. D.; Piola, A. R.; Reseghetti, F.; Schuckmann, K.; Trenberth, K. E.; Velicogna, I.; Willis, J. K. A Review of Global Ocean Temperature Observations: Implications for Ocean Heat Content Estimates and Climate Change. *Rev. Geophys.* **2013**, *51* (3), 450–483. <https://doi.org/10.1002/rog.20022>.

- (9) Burrows, S. M.; Ogunro, O.; Frossard, A. A.; Russell, L. M.; Rasch, P. J.; Elliott, S. M. A Physically Based Framework for Modeling the Organic Fractionation of Sea Spray Aerosol from Bubble Film Langmuir Equilibria. *Atmospheric Chem. Phys.* **2014**, *14* (24), 13601–13629. <https://doi.org/10.5194/acp-14-13601-2014>.
- (10) Cheng, S.; Li, S.; Tsona, N. T.; George, C.; Du, L. Insights into the Headgroup and Chain Length Dependence of Surface Characteristics of Organic-Coated Sea Spray Aerosols. *ACS Earth Space Chem.* **2019**, *3* (4), 571–580. <https://doi.org/10.1021/acsearthspacechem.8b00212>.
- (11) Wilson, T. W.; Ladino, L. A.; Alpert, P. A.; Breckels, M. N.; Brooks, I. M.; Browse, J.; Burrows, S. M.; Carslaw, K. S.; Huffman, J. A.; Judd, C.; Kilhau, W. P.; Mason, R. H.; McFiggans, G.; Miller, L. A.; Najera, J. J.; Polishchuk, E.; Rae, S.; Schiller, C. L.; Si, M.; Temprado, J. V.; Whale, T. F.; Wong, J. P. S.; Wurl, O.; Yakobi-Hancock, J. D.; Abbatt, J. P. D.; Aller, J. Y.; Bertram, A. K.; Knopf, D. A.; Murray, B. J. A Marine Biogenic Source of Atmospheric Ice-Nucleating Particles. *Nature* **2015**, *525* (7568), 234–238. <https://doi.org/10.1038/nature14986>.
- (12) Ting Katty Huang, W.; Ickes, L.; Tegen, I.; Rinaldi, M.; Ceburnis, D.; Lohmann, U. Global Relevance of Marine Organic Aerosol as Ice Nucleating Particles. *Atmospheric Chem. Phys.* **2018**, *18* (15), 11423–11445. <https://doi.org/10.5194/acp-18-11423-2018>.
- (13) DeMott, P. J.; Hill, T. C. J.; McCluskey, C. S.; Prather, K. A.; Collins, D. B.; Sullivan, R. C.; Ruppel, M. J.; Mason, R. H.; Irish, V. E.; Lee, T.; Hwang, C. Y.; Rhee, T. S.; Snider, J. R.; McMeeking, G. R.; Dhaniyala, S.; Lewis, E. R.; Wentzell, J. J. B.; Abbatt, J.; Lee, C.; Sultana, C. M.; Ault, A. P.; Axson, J. L.; Martinez, M. D.; Venero, I.; Santos-Figueroa, G.; Stokes, M. D.; Deane, G. B.; Mayol-Bracero, O. L.; Grassian, V. H.; Bertram, T. H.; Bertram, A. K.; Moffett, B. F.; Franc, G. D. Sea Spray Aerosol as a Unique Source of Ice Nucleating Particles. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (21), 5797–5803. <https://doi.org/10.1073/pnas.1514034112>.
- (14) Carter-Fenk, K. A.; Dommer, A. C.; Fiamingo, M. E.; Kim, J.; Amaro, R. E.; Allen, H. C. Calcium Bridging Drives Polysaccharide Co-Adsorption to a Proxy Sea Surface Microlayer. *Phys. Chem. Chem. Phys.* **2021**, *23* (30), 16401–16416. <https://doi.org/10.1039/d1cp01407b>.
- (15) Yao, X.; Liu, Q.; Wang, B.; Yu, J.; Aristov, M. M.; Shi, C.; Zhang, G. G. Z.; Yu, L. Anisotropic Molecular Organization at a Liquid/Vapor Interface Promotes Crystal Nucleation with Polymorph Selection. *J. Am. Chem. Soc.* **2022**, *144* (26), 11638–11645. <https://doi.org/10.1021/jacs.2c02623>.
- (16) Neal, J. F.; Rogers, M. M.; Smeltzer, M. A.; Carter-Fenk, K. A.; Grooms, A. J.; Zerkle, M. M.; Allen, H. C. Sodium Drives Interfacial Equilibria for Semi-Soluble Phosphoric and Phosphonic Acids of Model Sea Spray Aerosol Surfaces. *ACS Earth Space Chem.* **2020**, *4* (9), 1549–1557. <https://doi.org/10.1021/acsearthspacechem.0c00132>.
- (17) Vazquez De Vasquez, M. G.; Carter-Fenk, K. A.; McCaslin, L. M.; Beasley, E. E.; Clark, J. B.; Allen, H. C. Hydration and Hydrogen Bond Order of Octadecanoic

- Acid and Octadecanol Films on Water at 21 and 1°C. *J. Phys. Chem. A* **2021**, *125* (46), 10065–10078. <https://doi.org/10.1021/acs.jpca.1c06101>.
- (18) Myklestad, S. M. Dissolved Organic Carbon from Phytoplankton. In *Marine Chemistry*; Wangersky, P. J., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2000; pp 111–148. https://doi.org/10.1007/10683826_5.
- (19) Lønborg, C.; Carreira, C.; Jickells, T.; Álvarez-Salgado, X. A. Impacts of Global Change on Ocean Dissolved Organic Carbon (DOC) Cycling. *Front. Mar. Sci.* **2020**, *7* (June), 1–24. <https://doi.org/10.3389/fmars.2020.00466>.
- (20) Jayarathne, T.; Sultana, C. M.; Lee, C.; Malfatti, F.; Cox, J. L.; Pendergraft, M. A.; Moore, K. A.; Azam, F.; Tivanski, A. V.; Cappa, C. D.; Bertram, T. H.; Grassian, V. H.; Prather, K. A.; Stone, E. A. Enrichment of Saccharides and Divalent Cations in Sea Spray Aerosol during Two Phytoplankton Blooms. *Environ. Sci. Technol.* **2016**, *50* (21), 11511–11520. <https://doi.org/10.1021/acs.est.6b02988>.
- (21) Gericke, A.; Hühnerfuss, H. Investigation of Z- and E-Unsaturated Fatty Acids, Fatty Acid Esters, and Fatty Alcohols at the Air/Water Interface by Infrared Spectroscopy. *Langmuir* **1995**, *11* (1), 225–230. <https://doi.org/10.1021/la00001a039>.
- (22) Li, Y.; Shrestha, M.; Luo, M.; Sit, I.; Song, M.; Grassian, V. H.; Xiong, W. Salting up of Proteins at the Air/Water Interface. *Langmuir* **2019**, *35* (43), 13815–13820. <https://doi.org/10.1021/acs.langmuir.9b01901>.
- (23) Klenner, F.; Postberg, F.; Hillier, J.; Khawaja, N.; Cable, M. L.; Abel, B.; Kempf, S.; Glein, C. R.; Lunine, J. I.; Hodyss, R.; Reviol, R.; Stolz, F. Discriminating Abiotic and Biotic Fingerprints of Amino Acids and Fatty Acids in Ice Grains Relevant to Ocean Worlds. *Astrobiology* **2020**, ast.2019.2188. <https://doi.org/10.1089/ast.2019.2188>.
- (24) Martin, A.; Mcminn, A. Sea Ice, Extremophiles and Life on Extra-Terrestrial Ocean Worlds. **2020**. <https://doi.org/10.1017/S1473550416000483>.
- (25) Khawaja, N.; Postberg, F.; Hillier, J.; Klenner, F.; Kempf, S.; Nölle, L.; Reviol, R.; Zou, Z.; Srama, R. Low-Mass Nitrogen-, Oxygen-Bearing, and Aromatic Compounds in Enceladean Ice Grains. *Mon. Not. R. Astron. Soc.* **2019**, *489* (4), 5231–5243. <https://doi.org/10.1093/mnras/stz2280>.
- (26) Klenner, F.; Postberg, F.; Hillier, J.; Khawaja, N.; Reviol, R.; Stolz, F.; Cable, M. L.; Abel, B.; Nölle, L. Analog Experiments for the Identification of Trace Biosignatures in Ice Grains from Extraterrestrial Ocean Worlds. *Astrobiology* **2020**, *20* (2), 179–189. <https://doi.org/10.1089/ast.2019.2065>.
- (27) Postberg, F.; Khawaja, N.; Abel, B.; Choblet, G.; Glein, C. R.; Gudipati, M. S.; Henderson, B. L.; Hsu, H. W.; Kempf, S.; Klenner, F.; Moragas-Klostermeyer, G.; Magee, B.; Nölle, L.; Perry, M.; Reviol, R.; Schmidt, J.; Srama, R.; Stolz, F.; Tobie, G.; Trieloff, M.; Waite, J. H. Macromolecular Organic Compounds from the Depths of Enceladus. *Nature* **2018**, *558* (7711), 564–568. <https://doi.org/10.1038/s41586-018-0246-4>.
- (28) Porco, C. C.; Dones, L.; Mitchell, C. Could It Be Snowing Microbes on Enceladus? Assessing Conditions in Its Plume and Implications for Future

- Missions. *Astrobiology* **2017**, *17* (9), 876–901.
<https://doi.org/10.1089/ast.2017.1665>.
- (29) Postberg, F.; Kempf, S.; Schmidt, J.; Brilliantov, N.; Beinsen, A.; Abel, B.; Buck, U.; Srama, R. Sodium Salts in E-Ring Ice Grains from an Ocean below the Surface of Enceladus. *Nature* **2009**, *459* (7250), 1098–1101.
<https://doi.org/10.1038/nature08046>.
- (30) Cable, M. L.; Porco, C.; Glein, C. R.; German, C. R.; MacKenzie, S. M.; Neveu, M.; Hoehler, T. M.; Hofmann, A. E.; Hendrix, A. R.; Eigenbrode, J.; Postberg, F.; Spilker, L. J.; McEwen, A.; Khawaja, N.; Waite, J. H.; Wurz, P.; Helbert, J.; Anbar, A.; Vera, J.-P. de; Núñez, J. The Science Case for a Return to Enceladus. *Planet. Sci. J.* **2021**, *2* (4), 132. <https://doi.org/10.3847/PSJ/abfb7a>.
- (31) Postberg, F.; Schmidt, J.; Hillier, J.; Kempf, S.; Srama, R. A Salt-Water Reservoir as the Source of a Compositionally Stratified Plume on Enceladus. *Nature* **2011**, *474* (7353), 620–622. <https://doi.org/10.1038/nature10175>.
- (32) Hsu, H. W.; Postberg, F.; Sekine, Y.; Shibuya, T.; Kempf, S.; Horányi, M.; Juhász, A.; Altobelli, N.; Suzuki, K.; Masaki, Y.; Kuwatani, T.; Tachibana, S.; Sirono, S. I.; Moragas-Klostermeyer, G.; Srama, R. Ongoing Hydrothermal Activities within Enceladus. *Nature* **2015**, *519* (7542), 207–210.
<https://doi.org/10.1038/nature14262>.
- (33) Waite, J. H.; Lewis, W. S.; Kasprzak, W. T.; Anicich, V. G.; Block, B. P.; Cravens, T. E.; Fletcher, G. G.; Ip, W. H.; Luhmann, J. G.; Mcnutt, R. L.; Niemann, H. B.; Parejko, J. K.; Richards, J. E.; Thorpe, R. L.; Walter, E. M.; Yelle, R. V. The Cassini Ion and Neutral Mass Spectrometer (INMS) Investigation. *Space Sci. Rev.* **2004**, *114* (1–4), 113–231.
<https://doi.org/10.1007/s11214-004-1408-2>.
- (34) Glein, C. R.; Waite, J. H. The Carbonate Geochemistry of Enceladus’ Ocean. *Geophys. Res. Lett.* **2020**, *47* (3). <https://doi.org/10.1029/2019GL085885>.
- (35) Neveu, M.; Aspin, A.; Naseem, M.; Yang, Z. Effect of the Liquid-Vacuum Transition on the Relative Abundances of Amino and Fatty Acids Sought as Biosignatures on Icy Ocean Worlds. *Earth Planet. Sci. Lett.* **2024**, *630*, 118622.
<https://doi.org/10.1016/j.epsl.2024.118622>.
- (36) Fox-Powell, M.; Moral, Á. del; Stephens, B.; Dazley, C.; Slade, D.; Richards, G.; Cousins, C.; Olsson-Francis, K. *Entombment of Microbial Biomass within Rapidly Frozen Fluid Droplets Relevant to the Plumes of Enceladus*; EPSC2022-1230; Copernicus Meetings, 2022. <https://doi.org/10.5194/epsc2022-1230>.
- (37) Dannenmann, M.; Klenner, F.; Bönigk, J.; Pavlista, M.; Napoleoni, M.; Hillier, J.; Khawaja, N.; Olsson-Francis, K.; Cable, M. L.; Malaska, M. J.; Abel, B.; Postberg, F. Toward Detecting Biosignatures of DNA, Lipids, and Metabolic Intermediates from Bacteria in Ice Grains Emitted by Enceladus and Europa. *Astrobiology* **2023**, *23* (1), 60–75. <https://doi.org/10.1089/ast.2022.0063>.
- (38) Napoleoni, M.; Klenner, F.; Khawaja, N.; Hillier, J. K.; Postberg, F. Mass Spectrometric Fingerprints of Organic Compounds in NaCl-Rich Ice Grains from Europa and Enceladus. *ACS Earth Space Chem.* **2023**, *7* (4), 735–752.
<https://doi.org/10.1021/acsearthspacechem.2c00342>.

- (39) Chelius, D.; Bondarenko, P. V. Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry. *J. Proteome Res.* **2002**, *1* (4), 317–323. <https://doi.org/10.1021/pr025517j>.
- (40) Venable, J. D.; Dong, M. Q.; Wohlschlegel, J.; Dillin, A.; Yates, J. R. Automated Approach for Quantitative Analysis of Complex Peptide Mixtures from Tandem Mass Spectra. *Nat. Methods* **2004**, *1* (1), 39–45. <https://doi.org/10.1038/nmeth705>.
- (41) Stahl, D. C.; Swiderek, K. M.; Davis, M. T.; Lee, T. D. *Data-Controlled Automation of Liquid Chromatography/Tandem Mass Spectrometry Analysis of Peptide Mixtures*.
- (42) Granger, R. M.; Yochum, H. M.; Granger, J. N.; Sienerth, K. D. *Instrumental Analysis*; Oxford University Press, 2017.
- (43) Wang, L.; Morita, A.; North, N. M.; Baumler, S. M.; Springfield, E. W.; Allen, H. C. Identification of Ion Pairs in Aqueous NaCl and KCl Solutions in Combination with Raman Spectroscopy, Molecular Dynamics, and Quantum Chemical Calculations. *J. Phys. Chem. B* **2023**, *127* (7), 1618–1627. <https://doi.org/10.1021/acs.jpcc.2c07923>.
- (44) Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media, Inc., 2019.
- (45) Fu, C.; Yang, J. Granular Classification for Imbalanced Datasets: A Minkowski Distance-Based Method. *Algorithms* **2021**, *14* (2), 54. <https://doi.org/10.3390/a14020054>.
- (46) *What is the k-nearest neighbors algorithm? | IBM*. <https://www.ibm.com/topics/knn> (accessed 2024-02-27).
- (47) *sklearn.gaussian_process.kernels.RBF*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.gaussian_process.kernels.RBF.html (accessed 2024-02-27).
- (48) Llop, E.; Pinho, P.; Matos, P.; Pereira, M. J.; Branquinho, C. The Use of Lichen Functional Groups as Indicators of Air Quality in a Mediterranean Urban Environment. *Ecol. Indic.* **2012**, *13* (1), 215–221. <https://doi.org/10.1016/j.ecolind.2011.06.005>.
- (49) Infantes, L.; Chisholm, J.; Motherwell, S. Extended Motifs from Water and Chemical Functional Groups in Organic Molecular Crystals. *CrystEngComm* **2003**, *5* (85), 480–486. <https://doi.org/10.1039/B312846F>.
- (50) Tsou, C. L. Relation between Modification of Functional Groups of Proteins and Their Biological Activity. I. A Graphical Method for the Determination of the Number and Type of Essential Groups. *Sci. Sin.* **1962**, *11*, 1535–1558.
- (51) Coe, J. V.; Chen, Z.; Li, R.; Nystrom, S. V.; Butke, R.; Miller, B.; Hitchcock, C. L.; Allen, H. C.; Povoski, S. P.; Jr, E. W. M. Molecular Constituents of Colorectal Cancer Metastatic to the Liver by Imaging Infrared Spectroscopy. In *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XIII*; SPIE, 2015; Vol. 9328, pp 98–104. <https://doi.org/10.1117/12.2079884>.

- (52) Kind, T.; Fiehn, O. Advances in Structure Elucidation of Small Molecules Using Mass Spectrometry. *Bioanal. Rev.* **2010**, *2* (1), 23–60. <https://doi.org/10.1007/s12566-010-0015-9>.
- (53) Levsen, K.; Schiebel, H.-M.; Behnke, B.; Dötzer, R.; Dreher, W.; Elend, M.; Thiele, H. Structure Elucidation of Phase II Metabolites by Tandem Mass Spectrometry: An Overview. *J. Chromatogr. A* **2005**, *1067* (1–2), 55–72. <https://doi.org/10.1016/j.chroma.2004.08.165>.
- (54) Winston, R. L.; Fitzgerald, M. C. Mass Spectrometry as a Readout of Protein Structure and Function. *Mass Spectrom. Rev.* **1997**, *16* (4), 165–179. [https://doi.org/10.1002/\(SICI\)1098-2787\(1997\)16:4<165::AID-MAS1>3.0.CO;2-F](https://doi.org/10.1002/(SICI)1098-2787(1997)16:4<165::AID-MAS1>3.0.CO;2-F).
- (55) Prabhudesai, V. S.; Kelkar, A. H.; Nandi, D.; Krishnakumar, E. Functional Group Dependent Site Specific Fragmentation of Molecules by Low Energy Electrons. *Phys. Rev. Lett.* **2005**, *95* (14), 143202. <https://doi.org/10.1103/PhysRevLett.95.143202>.
- (56) Ghojogh, B.; Ghodsi, A.; Karray, F.; Crowley, M. Generative Adversarial Networks and Adversarial Autoencoders: Tutorial and Survey. arXiv November 25, 2021. <http://arxiv.org/abs/2111.13282> (accessed 2022-09-20).
- (57) Weiss, S. M.; Kapouleas, I. An Empirical Comparison of Pattern Recognition, Neural Nets, and Machine Learning Classification Methods. In *Proceedings of the 11th international joint conference on Artificial intelligence - Volume 1*; IJCAI'89; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1989; pp 781–787.
- (58) Zhou, C.; Bowler, L. D.; Feng, J. A Machine Learning Approach to Explore the Spectra Intensity Pattern of Peptides Using Tandem Mass Spectrometry Data. *BMC Bioinformatics* **2008**, *9* (1), 325. <https://doi.org/10.1186/1471-2105-9-325>.
- (59) Ulintz, P. J.; Zhu, J.; Qin, Z. S.; Andrews, P. C. Improved Classification of Mass Spectrometry Database Search Results Using Newer Machine Learning Approaches. *Mol. Cell. Proteomics MCP* **2006**, *5* (3), 497–509. <https://doi.org/10.1074/mcp.M500233-MCP200>.
- (60) Rivera, S. M.; Christou, P.; Canela-Garayoa, R. Identification of Carotenoids Using Mass Spectrometry. *Mass Spectrom. Rev.* **2014**, *33* (5), 353–372. <https://doi.org/10.1002/mas.21390>.
- (61) Le Lacheur, R. M.; Sonnenberg, L. B.; Singer, P. C.; Christman, R. F.; Charles, M. J. Identification of Carbonyl Compounds in Environmental Samples. *Environ. Sci. Technol.* **1993**, *27* (13), 2745–2753. <https://doi.org/10.1021/es00049a013>.
- (62) Shimbo, K.; Kubo, S.; Harada, Y.; Oonuki, T.; Yokokura, T.; Yoshida, H.; Amao, M.; Nakamura, M.; Kageyama, N.; Yamazaki, J.; Ozawa, S.; Hirayama, K.; Ando, T.; Miura, J.; Miyano, H. Automated Precolumn Derivatization System for Analyzing Physiological Amino Acids by Liquid Chromatography/Mass Spectrometry. *Biomed. Chromatogr.* **2010**, *24* (7), 683–691. <https://doi.org/10.1002/bmc.1346>.
- (63) Bidlingmeyer, B. A.; Cohen, S. A.; Tarvin, T. L. Rapid Analysis of Amino Acids Using Pre-Column Derivatization. *J. Chromatogr. B. Biomed. Sci. App.* **1984**, *336* (1), 93–104. [https://doi.org/10.1016/S0378-4347\(00\)85133-6](https://doi.org/10.1016/S0378-4347(00)85133-6).

- (64) Dron, J.; Abidi, E.; Haddad, I. E.; Marchand, N.; Wortham, H. Precursor Ion Scanning–Mass Spectrometry for the Determination of Nitro Functional Groups in Atmospheric Particulate Organic Matter. *Anal. Chim. Acta* **2008**, *618* (2), 184–195. <https://doi.org/10.1016/j.aca.2008.04.057>.
- (65) Dührkop, K.; Nothias, L.-F.; Fleischauer, M.; Reher, R.; Ludwig, M.; Hoffmann, M. A.; Petras, D.; Gerwick, W. H.; Rousu, J.; Dorrestein, P. C.; Böcker, S. Systematic Classification of Unknown Metabolites Using High-Resolution Fragmentation Mass Spectra. *Nat. Biotechnol.* **2021**, *39* (4), 462–471. <https://doi.org/10.1038/s41587-020-0740-8>.
- (66) Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. MSNovelist: De Novo Structure Generation from Mass Spectra. *Nat. Methods* **2022**, *19* (7), 865–870. <https://doi.org/10.1038/s41592-022-01486-3>.
- (67) Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Searching Molecular Structure Databases with Tandem Mass Spectra Using CSI:FingerID. *Proc. Natl. Acad. Sci.* **2015**, *112* (41), 12580–12585. <https://doi.org/10.1073/pnas.1509788112>.
- (68) Heinonen, M.; Shen, H.; Zamboni, N.; Rousu, J. Metabolite Identification and Molecular Fingerprint Prediction through Machine Learning. *Bioinformatics* **2012**, *28* (18), 2333–2341. <https://doi.org/10.1093/bioinformatics/bts437>.
- (69) Asef, C. K.; Rainey, M. A.; Garcia, B. M.; Gouveia, G. J.; Shaver, A. O.; Leach, F. E. I.; Morse, A. M.; Edison, A. S.; McIntyre, L. M.; Fernández, F. M. Unknown Metabolite Identification Using Machine Learning Collision Cross-Section Prediction and Tandem Mass Spectrometry. *Anal. Chem.* **2023**, *95* (2), 1047–1056. <https://doi.org/10.1021/acs.analchem.2c03749>.
- (70) Feucherolles, M.; Nennig, M.; Becker, S. L.; Martiny, D.; Losch, S.; Penny, C.; Cauchie, H.-M.; Ragimbeau, C. Combination of MALDI-TOF Mass Spectrometry and Machine Learning for Rapid Antimicrobial Resistance Screening: The Case of *Campylobacter* Spp. *Front. Microbiol.* **2022**, *12*.
- (71) Hao, Y.; Lynch, K.; Fan, P.; Jurtschenko, C.; Cid, M.; Zhao, Z.; Yang, H. S. Development of a Machine Learning Algorithm for Drug Screening Analysis on High-Resolution UPLC-MSE/QTOF Mass Spectrometry. *J. Appl. Lab. Med.* **2023**, *8* (1), 53–66. <https://doi.org/10.1093/jalm/jfac100>.
- (72) Enders, A. A.; North, N. M.; Fensore, C. M.; Velez-Alvarez, J.; Allen, H. C. Functional Group Identification for FTIR Spectra Using Image-Based Machine Learning Models. *Anal. Chem.* **2021**, *93* (28), 9711–9718. <https://doi.org/10.1021/acs.analchem.1c00867>.
- (73) Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision; 2016; pp 2818–2826.
- (74) Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*; 2009; pp 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>.

- (75) Recht, B.; Roelofs, R.; Schmidt, L.; Shankar, V. Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*; PMLR, 2019; pp 5389–5400.
- (76) You, Y.; Zhang, Z.; Hsieh, C.-J.; Demmel, J.; Keutzer, K. ImageNet Training in Minutes. In *Proceedings of the 47th International Conference on Parallel Processing*; ICPP '18; Association for Computing Machinery: New York, NY, USA, 2018; pp 1–10. <https://doi.org/10.1145/3225058.3225069>.
- (77) Ahmed, E.; Jones, M.; Marks, T. K. An Improved Deep Learning Architecture for Person Re-Identification; 2015; pp 3908–3916.
- (78) Bantupalli, K.; Xie, Y. American Sign Language Recognition Using Deep Learning and Computer Vision. In *2018 IEEE International Conference on Big Data (Big Data)*; 2018; pp 4896–4899. <https://doi.org/10.1109/BigData.2018.8622141>.
- (79) Albatayneh, O.; Forslöf, L.; Ksaibati, K. Image Retraining Using TensorFlow Implementation of the Pretrained Inception-v3 Model for Evaluating Gravel Road Dust. *J. Infrastruct. Syst.* **2020**, *26* (2), 04020014. [https://doi.org/10.1061/\(ASCE\)IS.1943-555X.0000545](https://doi.org/10.1061/(ASCE)IS.1943-555X.0000545).
- (80) Zhang, X.; Lin, T.; Xu, J.; Luo, X.; Ying, Y. DeepSpectra: An End-to-End Deep Learning Approach for Quantitative Spectral Analysis. *Anal. Chim. Acta* **2019**, *1058*, 48–57. <https://doi.org/10.1016/j.aca.2019.01.002>.
- (81) Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Liu, C.; Zhang, X.; Shan, B.; Ghodsi, A.; Li, M. Deep Learning Enables de Novo Peptide Sequencing from Data-Independent-Acquisition Mass Spectrometry. *Nat. Methods* **2019**, *16* (1), 63–66. <https://doi.org/10.1038/s41592-018-0260-3>.
- (82) Kirasich, K.; Smith, T.; Sadler, B. Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Sci. Rev.* **2018**, *1* (3).
- (83) Feng, J.; Xu, H.; Mannor, S.; Yan, S. Robust Logistic Regression and Classification. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2014; Vol. 27.
- (84) Rácz, A.; Bajusz, D.; Héberger, K. Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecules* **2021**, *26* (4), 1111. <https://doi.org/10.3390/molecules26041111>.
- (85) Khaire, U. M.; Dhanalakshmi, R. Stability of Feature Selection Algorithm: A Review. *J. King Saud Univ. - Comput. Inf. Sci.* **2022**, *34* (4), 1060–1073. <https://doi.org/10.1016/j.jksuci.2019.06.012>.
- (86) Kumar, V. Feature Selection: A Literature Review. *Smart Comput. Rev.* **2014**, *4* (3). <https://doi.org/10.6029/smarterc.2014.03.007>.
- (87) Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; Liu, H. Feature Selection: A Data Perspective. *ACM Comput. Surv.* **2017**, *50* (6), 94:1-94:45. <https://doi.org/10.1145/3136625>.
- (88) Venkatesh, B.; Anuradha, J. A Review of Feature Selection and Its Methods. *Cybern. Inf. Technol.* **2019**, *19* (1), 3–26. <https://doi.org/10.2478/cait-2019-0001>.
- (89) Lu, Y.; Cohen, I.; Zhou, X. S.; Tian, Q. Feature Selection Using Principal Feature Analysis. In *Proceedings of the 15th ACM international conference on*

- Multimedia*; MM '07; Association for Computing Machinery: New York, NY, USA, 2007; pp 301–304. <https://doi.org/10.1145/1291233.1291297>.
- (90) El-Amir, H.; Hamdy, M. *Deep Learning Pipeline: Building a Deep Learning Model with TensorFlow*; Apress, 2019.
- (91) Sultana, N.; Chilamkurti, N.; Peng, W.; Alhadad, R. Survey on SDN Based Network Intrusion Detection System Using Machine Learning Approaches. *Peer--Peer Netw. Appl.* **2019**, *12* (2), 493–501. <https://doi.org/10.1007/s12083-017-0630-0>.
- (92) Drgoňa, J.; Picard, D.; Kvasnica, M.; Helsen, L. Approximate Model Predictive Building Control via Machine Learning. *Appl. Energy* **2018**, *218*, 199–216. <https://doi.org/10.1016/j.apenergy.2018.02.156>.
- (93) R.m., S. P.; Maddikunta, P. K. R.; M., P.; Koppu, S.; Gadekallu, T. R.; Chowdhary, C. L.; Alazab, M. An Effective Feature Engineering for DNN Using Hybrid PCA-GWO for Intrusion Detection in IoMT Architecture. *Comput. Commun.* **2020**, *160*, 139–149. <https://doi.org/10.1016/j.comcom.2020.05.048>.
- (94) Orellana, M. V.; Matrai, P. A.; Leck, C.; Rauschenberg, C. D.; Lee, A. M.; Coz, E. Marine Microgels as a Source of Cloud Condensation Nuclei in the High Arctic. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (33), 13612–13617. <https://doi.org/10.1073/pnas.1102457108>.
- (95) Ogunro, O. O.; Burrows, S. M.; Elliott, S.; Frossard, A. A.; Hoffman, F.; Letscher, R. T.; Moore, J. K.; Russell, L. M.; Wang, S.; Wingenter, O. W. Global Distribution and Surface Activity of Macromolecules in Offline Simulations of Marine Organic Chemistry. *Biogeochemistry* **2015**, *126* (1–2), 25–56. <https://doi.org/10.1007/s10533-015-0136-x>.
- (96) Burrows, S. M.; Easter, R.; Liu, X.; Ma, P.-L.; Wang, H.; Elliott, S. M.; Singh, B.; Zhang, K.; Rasch, P. J. OCEANFILMS Sea-Spray Organic Aerosol Emissions – Part 1: Implementation and Impacts on Clouds. *Atmospheric Chem. Phys. Discuss.* **2018**, 1–27. <https://doi.org/10.5194/acp-2018-70>.
- (97) Elliott, S.; Menzo, Z.; Jayasinghe, A.; Allen, H. C.; Ogunro, O.; Gibson, G.; Hoffman, F.; Wingenter, O. Biogeochemical Equation of State for the Sea-Air Interface. *Atmosphere* **2019**, *10* (5), 1–17. <https://doi.org/10.3390/atmos10050230>.
- (98) Elliott, S.; Burrows, S.; Cameron-Smith, P.; Hoffman, F.; Hunke, E.; Jeffery, N.; Liu, Y.; Maltrud, M.; Menzo, Z.; Ogunro, O.; Van Roekel, L.; Wang, S.; Brunke, M.; Jin, M.; Letscher, R.; Meskhidze, N.; Russell, L.; Simpson, I.; Stokes, D.; Wingenter, O. Does Marine Surface Tension Have Global Biogeography? Addition for the OCEANFILMS Package. *Atmosphere* **2018**, *9* (6), 216. <https://doi.org/10.3390/atmos9060216>.
- (99) Hasenecz, E. S.; Kaluarachchi, C. P.; Lee, H. D.; Tivanski, A. V.; Stone, E. A. Saccharide Transfer to Sea Spray Aerosol Enhanced by Surface Activity, Calcium, and Protein Interactions. *ACS Earth Space Chem.* **2019**, *3* (11), 2539–2548. <https://doi.org/10.1021/acsearthspacechem.9b00197>.
- (100) Liu, N.; Yang, Y.; Li, F.; Ge, F.; Kuang, Y. Importance of Controlling pH-Depended Dissolved Inorganic Carbon to Prevent Algal Bloom Outbreaks.

- Bioresour. Technol.* **2016**, *220*, 246–252.
<https://doi.org/10.1016/j.biortech.2016.08.059>.
- (101) Kisand, V.; Tammert, H. Bacterioplankton Strategies for Leucine and Glucose Uptake after a Cyanobacterial Bloom in an Eutrophic Shallow Lake. *Soil Biol. Biochem.* **2000**, *32* (13), 1965–1972. [https://doi.org/10.1016/S0038-0717\(00\)00171-1](https://doi.org/10.1016/S0038-0717(00)00171-1).
- (102) Jayarathne, T.; Sultana, C. M.; Lee, C.; Malfatti, F.; Cox, J. L.; Pendergraft, M. A.; Moore, K. A.; Azam, F.; Tivanski, A. V.; Cappa, C. D.; Bertram, T. H.; Grassian, V. H.; Prather, K. A.; Stone, E. A. Enrichment of Saccharides and Divalent Cations in Sea Spray Aerosol During Two Phytoplankton Blooms. *Environ. Sci. Technol.* **2016**, *50* (21), 11511–11520. <https://doi.org/10.1021/acs.est.6b02988>.
- (103) Jayarathne, T.; Sultana, C. M.; Lee, C.; Malfatti, F.; Cox, J. L.; Pendergraft, M. A.; Moore, K. A.; Azam, F.; Tivanski, A. V.; Cappa, C. D.; Bertram, T. H.; Grassian, V. H.; Prather, K. A.; Stone, E. A. Enrichment of Saccharides and Divalent Cations in Sea Spray Aerosol During Two Phytoplankton Blooms. *Environ. Sci. Technol.* **2016**, *50* (21), 11511–11520. <https://doi.org/10.1021/acs.est.6b02988>.
- (104) Cochran, R. E.; Laskina, O.; Jayarathne, T.; Laskin, A.; Laskin, J.; Lin, P.; Sultana, C.; Lee, C.; Moore, K. A.; Cappa, C. D.; Bertram, T. H.; Prather, K. A.; Grassian, V. H.; Stone, E. A. Analysis of Organic Anionic Surfactants in Fine and Coarse Fractions of Freshly Emitted Sea Spray Aerosol. *Environ. Sci. Technol.* **2016**, *50* (5), 2477–2486. <https://doi.org/10.1021/acs.est.5b04053>.
- (105) Enders, A. A.; North, N. M.; Fensore, C. M.; Velez-Alvarez, J.; Allen, H. C. Functional Group Identification for FTIR Spectra Using Image-Based Machine Learning Models. *Anal. Chem.* **2021**.
<https://doi.org/10.1021/acs.analchem.1c00867>.
- (106) Schleder, G. R.; Acosta, C. M.; Fazzio, A. Exploring Two-Dimensional Materials Thermodynamic Stability via Machine Learning. *ACS Appl. Mater. Interfaces* **2020**, *12* (18), 20149–20157. <https://doi.org/10.1021/acsami.9b14530>.
- (107) Moosavi, S. M.; Jablonka, K. M.; Smit, B. The Role of Machine Learning in the Understanding and Design of Materials. *J. Am. Chem. Soc.* **2020**, *142* (48), 20273–20287. <https://doi.org/10.1021/jacs.0c09105>.
- (108) Batra, K.; Zorn, K. M.; Foil, D. H.; Minerali, E.; Gawriljuk, V. O.; Lane, T. R.; Ekins, S. Quantum Machine Learning Algorithms for Drug Discovery Applications. *J. Chem. Inf. Model.* **2021**, *61* (6), 2641–2647.
<https://doi.org/10.1021/acs.jcim.1c00166>.
- (109) Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskiy, V.; Mamoshina, P.; Bozdaganyan, M.; Aliper, A.; Zhavoronkov, A.; Kadurin, A. Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery. *Mol. Pharm.* **2018**, *15* (10), 4398–4405.
<https://doi.org/10.1021/acs.molpharmaceut.8b00839>.
- (110) Zhang, J.; Hu, P.; Wang, H. Amorphous Catalysis: Machine Learning Driven High-Throughput Screening of Superior Active Site for Hydrogen Evolution Reaction. *J. Phys. Chem. C* **2020**, *124* (19), 10483–10494.
<https://doi.org/10.1021/acs.jpcc.0c00406>.

- (111) Ting, K. W.; Kamakura, H.; Poly, S. S.; Takao, M.; Siddiki, S. M. A. H.; Maeno, Z.; Matsushita, K.; Shimizu, K.; Toyao, T. Catalytic Methylation of *m*-Xylene, Toluene, and Benzene Using CO₂ and H₂ over TiO₂-Supported Re and Zeolite Catalysts: Machine-Learning-Assisted Catalyst Optimization. *ACS Catal.* **2021**, *11* (9), 5829–5838. <https://doi.org/10.1021/acscatal.0c05661>.
- (112) Miyake, Y.; Saeki, A. Machine Learning-Assisted Development of Organic Solar Cell Materials: Issues, Analyses, and Outlooks. *J. Phys. Chem. Lett.* **2021**, *12* (51), 12391–12401. <https://doi.org/10.1021/acs.jpcelett.1c03526>.
- (113) Masood, H.; Toe, C. Y.; Teoh, W. Y.; Sethu, V.; Amal, R. Machine Learning for Accelerated Discovery of Solar Photocatalysts. *ACS Catal.* **2019**, *9* (12), 11774–11787. <https://doi.org/10.1021/acscatal.9b02531>.
- (114) Al Ibrahim, E.; Farooq, A. Transfer Learning Approach to Multitarget Temperature-Dependent Reaction Rate Prediction. *J. Phys. Chem. A* **2022**, *126* (28), 4617–4629. <https://doi.org/10.1021/acs.jpca.2c00713>.
- (115) Freitas, R. S. M.; Lima, Á. P. F.; Chen, C.; Rochinha, F. A.; Mira, D.; Jiang, X. Towards Predicting Liquid Fuel Physicochemical Properties Using Molecular Dynamics Guided Machine Learning Models. *Fuel* **2022**, *329*, 125415. <https://doi.org/10.1016/j.fuel.2022.125415>.
- (116) Brandt, J.; Mattsson, K.; Hassellöv, M. Deep Learning for Reconstructing Low-Quality FTIR and Raman Spectra—A Case Study in Microplastic Analyses. *Anal. Chem.* **2021**, *93* (49), 16360–16368. <https://doi.org/10.1021/acs.analchem.1c02618>.
- (117) Fan, X.; Wang, Y.; Yu, C.; Lv, Y.; Zhang, H.; Yang, Q.; Wen, M.; Lu, H.; Zhang, Z. A Universal and Accurate Method for Easily Identifying Components in Raman Spectroscopy Based on Deep Learning. *Anal. Chem.* **2023**. <https://doi.org/10.1021/acs.analchem.2c03853>.
- (118) Takamura, A.; Halamkova, L.; Ozawa, T.; Lednev, I. K. Phenotype Profiling for Forensic Purposes: Determining Donor Sex Based on Fourier Transform Infrared Spectroscopy of Urine Traces. *Anal. Chem.* **2019**, *91* (9), 6288–6295. <https://doi.org/10.1021/acs.analchem.9b01058>.
- (119) Butler, H. J.; Brennan, P. M.; Cameron, J. M.; Finlayson, D.; Hegarty, M. G.; Jenkinson, M. D.; Palmer, D. S.; Smith, B. R.; Baker, M. J. Development of High-Throughput ATR-FTIR Technology for Rapid Triage of Brain Cancer. *Nat. Commun.* **2019**, *10* (1), 1–9. <https://doi.org/10.1038/s41467-019-12527-5>.
- (120) Lei, B.; Bissonnette, J. R.; Hogan, Ú. E.; Bec, A. E.; Feng, X.; Smith, R. D. L. Customizable Machine-Learning Models for Rapid Microplastic Identification Using Raman Microscopy. *Anal. Chem.* **2022**. <https://doi.org/10.1021/acs.analchem.2c02451>.
- (121) Richardson, P. I. C.; Muhamadali, H.; Ellis, D. I.; Goodacre, R. Rapid Quantification of the Adulteration of Fresh Coconut Water by Dilution and Sugars Using Raman Spectroscopy and Chemometrics. *Food Chem.* **2019**, *272* (January 2018), 157–164. <https://doi.org/10.1016/j.foodchem.2018.08.038>.
- (122) Gillio Meina, E.; Niyogi, S.; Liber, K. Multiple Linear Regression Modeling Predicts the Effects of Surface Water Chemistry on Acute Vanadium Toxicity to

- Model Freshwater Organisms. *Environ. Toxicol. Chem.* **2020**, *39* (9), 1737–1745. <https://doi.org/10.1002/etc.4798>.
- (123) Esbaugh, A. J.; Brix, K. V.; Mager, E. M.; De Schampelaere, K.; Grosell, M. Multi-Linear Regression Analysis, Preliminary Biotic Ligand Modeling, and Cross Species Comparison of the Effects of Water Chemistry on Chronic Lead Toxicity in Invertebrates. *Comp. Biochem. Physiol. Part C Toxicol. Pharmacol.* **2012**, *155* (2), 423–431. <https://doi.org/10.1016/j.cbpc.2011.11.005>.
- (124) *sklearn.linear_model.LinearRegression*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.LinearRegression.html (accessed 2023-11-02).
- (125) Mohammadi, M.; Khanmohammadi Khorrami, M.; Vatani, A.; Ghasemzadeh, H.; Vatanparast, H.; Bahramian, A.; Fallah, A. Genetic Algorithm Based Support Vector Machine Regression for Prediction of SARA Analysis in Crude Oil Samples Using ATR-FTIR Spectroscopy. *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* **2021**, *245*, 118945. <https://doi.org/10.1016/j.saa.2020.118945>.
- (126) Chen, C.; Liang, R.; Ge, Y.; Li, J.; Yan, B.; Cheng, Z.; Tao, J.; Wang, Z.; Li, M.; Chen, G. Fast Characterization of Biomass Pyrolysis Oil via Combination of ATR-FTIR and Machine Learning Models. *Renew. Energy* **2022**, *194*, 220–231. <https://doi.org/10.1016/j.renene.2022.05.097>.
- (127) Schill, S. R.; Burrows, S. M.; Hasenecz, E. S.; Stone, E. A.; Bertram, T. H. The Impact of Divalent Cations on the Enrichment of Soluble Saccharides in Primary Sea Spray Aerosol. *Atmosphere* **2018**, *9* (12), 13–17. <https://doi.org/10.3390/atmos9120476>.
- (128) Roy, S. Distributions of Phytoplankton Carbohydrate, Protein and Lipid in the World Oceans from Satellite Ocean Colour. *ISME J.* **2018**, *12* (6), 1457–1472. <https://doi.org/10.1038/s41396-018-0054-8>.
- (129) Cheng, Y. C.; Bianco, C. L.; Sandler, S. I.; Lenhoff, A. M. Salting-out of Lysozyme and Ovalbumin from Mixtures: Predicting Precipitation Performance from Protein-Protein Interactions. *Ind. Eng. Chem. Res.* **2008**, *47* (15), 5203–5213. <https://doi.org/10.1021/ie071462p>.
- (130) Kudryashova, E. V.; Meinders, M. B. J.; Visser, A. J. W. G.; Van Hoek, A.; De Jongh, H. H. J. Structure and Dynamics of Egg White Ovalbumin Adsorbed at the Air/Water Interface. *Eur. Biophys. J.* **2003**, *32* (6), 553–562. <https://doi.org/10.1007/s00249-003-0301-3>.
- (131) Langmuir, I.; Waugh, D. F. The Adsorption of Proteins at Oil-Water Interfaces and Artificial Protein-Lipoid Membranes. *J. Gen. Physiol.* **1938**, 745–755. <https://doi.org/10.1085/jgp.21.6.745>.
- (132) Angle, K. J.; Nowak, C. M.; Davasam, A.; Dommer, A. C.; Wauer, N. A.; Amaro, R. E.; Grassian, V. H. Amino Acids Are Driven to the Interface by Salts and Acidic Environments. *J. Phys. Chem. Lett.* **2022**, *13* (12), 2824–2829. <https://doi.org/10.1021/acs.jpcclett.2c00231>.

- (133) Benner, R.; Kaiser, K. Abundance of Amino Sugars and Peptidoglycan in Marine Particulate and Dissolved Organic Matter. *Limnol. Oceanogr.* **2003**, *48* (1), 118–128. <https://doi.org/10.4319/lo.2003.48.1.0118>.
- (134) Borkowski, M.; Orvalho, S.; Warszyński, P.; Demchuk, O. M.; Jarek, E.; Zawala, J. Experimental and Theoretical Study of Adsorption of Synthesized Amino Acid Core Derived Surfactants at an Air/Water Interface. *Phys. Chem. Chem. Phys.* **2022**, *24* (6), 3854–3864. <https://doi.org/10.1039/D1CP05322A>.
- (135) Lukita, A. *The Dreaded Antagonist: Data Leakage in Machine Learning*. Medium. <https://towardsdatascience.com/the-dreaded-antagonist-data-leakage-in-machine-learning-5f08679852cc> (accessed 2024-02-06).
- (136) *sklearn.linear_model.Ridge*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.Ridge.html (accessed 2023-12-21).
- (137) *sklearn.neighbors.KNeighborsRegressor*. scikit-learn. <https://scikit-learn/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html> (accessed 2023-12-21).
- (138) *sklearn.tree.DecisionTreeRegressor*. scikit-learn. <https://scikit-learn/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html> (accessed 2023-12-21).
- (139) *sklearn.ensemble.GradientBoostingRegressor*. scikit-learn. <https://scikit-learn/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html> (accessed 2023-12-21).
- (140) *sklearn.neural_network.MLPRegressor*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.neural_network.MLPRegressor.html (accessed 2023-12-21).
- (141) *sklearn.svm.SVR*. scikit-learn. <https://scikit-learn/stable/modules/generated/sklearn.svm.SVR.html> (accessed 2023-12-21).
- (142) Frossard, A. A.; Gérard, V.; Duplessis, P.; Kinsey, J. D.; Lu, X.; Zhu, Y.; Bisgrove, J.; Maben, J. R.; Long, M. S.; Chang, R. Y.-W.; Beaupré, S. R.; Kieber, D. J.; Keene, W. C.; Nozière, B.; Cohen, R. C. Properties of Seawater Surfactants Associated with Primary Marine Aerosol Particles Produced by Bursting Bubbles at a Model Air–Sea Interface. *Environ. Sci. Technol.* **2019**, *53* (16), 9407–9417. <https://doi.org/10.1021/acs.est.9b02637>.
- (143) Quinn, P. K.; Collins, D. B.; Grassian, V. H.; Prather, K. A.; Bates, T. S. Chemistry and Related Properties of Freshly Emitted Sea Spray Aerosol. *Chem. Rev.* **2015**, *115* (10), 4383–4399. <https://doi.org/10.1021/cr500713g>.
- (144) Sauer, J. S.; Mayer, K. J.; Lee, C.; Alves, M. R.; Amiri, S.; Bahaveolos, C. J.; Franklin, E. B.; Crocker, D. R.; Dang, D.; Dinasquet, J.; Garofalo, L. A.; Kaluarachchi, C. P.; Kilgour, D. B.; Mael, L. E.; Mitts, B. A.; Moon, D. R.; Moore, A. N.; Morris, C. K.; Mullenmeister, C. A.; Ni, C.-M.; Pendergraft, M. A.; Petras, D.; Simpson, R. M. C.; Smith, S.; Tumminello, P. R.; Walker, J. L.; DeMott, P. J.; Farmer, D. K.; Goldstein, A. H.; Grassian, V. H.; Jaffe, J. S.; Malfatti, F.; Martz, T. R.; Slade, J. H.; Tivanski, A. V.; Bertram, T. H.; Cappa, C. D.; Prather, K. A. The Sea Spray Chemistry and Particle Evolution Study

- (SeaSCAPE): Overview and Experimental Methods. *Environ. Sci. Process. Impacts* **2022**, *24* (2), 290–315. <https://doi.org/10.1039/D1EM00260K>.
- (145) Schiffer, J. M.; Mael, L. E.; Prather, K. A.; Amaro, R. E.; Grassian, V. H. Sea Spray Aerosol: Where Marine Biology Meets Atmospheric Chemistry. *ACS Cent. Sci.* **2018**, *4* (12), 1617–1623. <https://doi.org/10.1021/acscentsci.8b00674>.
- (146) Frossard, A. A.; Long, M. S.; Keene, W. C.; Duplessis, P.; Kinsey, J. D.; Maben, J. R.; Kieber, D. J.; Chang, R. Y.-W.; Beaupré, S. R.; Cohen, R. C.; Lu, X.; Bisgrove, J.; Zhu, Y. Marine Aerosol Production via Detrainment of Bubble Plumes Generated in Natural Seawater With a Forced-Air Venturi. *J. Geophys. Res. Atmospheres* **2019**, *124* (20), 10931–10950. <https://doi.org/10.1029/2019JD030299>.
- (147) Russell, L. M.; Hawkins, L. N.; Frossard, A. A.; Quinn, P. K.; Bates, T. S. Carbohydrate-like Composition of Submicron Atmospheric Particles and Their Production from Ocean Bubble Bursting. *Proc. Natl. Acad. Sci.* **2010**, *107* (15), 6652–6657. <https://doi.org/10.1073/pnas.0908905107>.
- (148) Cochran, R. E.; Laskina, O.; Trueblood, J. V.; Estillore, A. D.; Morris, H. S.; Jayarathne, T.; Sultana, C. M.; Lee, C.; Lin, P.; Laskin, J.; Laskin, A.; Dowling, J. A.; Qin, Z.; Cappa, C. D.; Bertram, T. H.; Tivanski, A. V.; Stone, E. A.; Prather, K. A.; Grassian, V. H. Molecular Diversity of Sea Spray Aerosol Particles: Impact of Ocean Biology on Particle Composition and Hygroscopicity. *Chem* **2017**, *2* (5), 655–667. <https://doi.org/10.1016/j.chempr.2017.03.007>.
- (149) Dommer, A. C.; Wauer, N. A.; Angle, K. J.; Davasam, A.; Rubio, P.; Luo, M.; Morris, C. K.; Prather, K. A.; Grassian, V. H.; Amaro, R. E. Revealing the Impacts of Chemical Complexity on Submicrometer Sea Spray Aerosol Morphology. *ACS Cent. Sci.* **2023**, *9* (6), 1088–1103. <https://doi.org/10.1021/acscentsci.3c00184>.
- (150) Wu, H.; Liang, C.; Zhang, C.; Chang, H.; Zhang, X.; Zhang, Y.; Zhong, N.; Xu, Y.; Zhong, D.; He, X.; Zhang, L.; Ho, S.-H. Mechanisms and Enhancements on Harmful Algal Blooms Conversion to Bioenergy Mediated with Dual-Functional Chitosan. *Appl. Energy* **2022**, *327*, 120142. <https://doi.org/10.1016/j.apenergy.2022.120142>.
- (151) Chang, H.; Wu, H.; Zhang, L.; Wu, W.; Zhang, C.; Zhong, N.; Zhong, D.; Xu, Y.; He, X.; Yang, J.; Zhang, Y.; Zhang, T.; Liao, Q.; Ho, S.-H. Gradient Electro-Processing Strategy for Efficient Conversion of Harmful Algal Blooms to Biohythane with Mechanisms Insight. *Water Res.* **2022**, *222*, 118929. <https://doi.org/10.1016/j.watres.2022.118929>.
- (152) Barile, P. J. Widespread Sewage Pollution of the Indian River Lagoon System, Florida (USA) Resolved by Spatial Analyses of Macroalgal Biogeochemistry. *Mar. Pollut. Bull.* **2018**, *128*, 557–574. <https://doi.org/10.1016/j.marpolbul.2018.01.046>.
- (153) Passow, U.; Lee, K. Future Oil Spill Response Plans Require Integrated Analysis of Factors That Influence the Fate of Oil in the Ocean. *Curr. Opin. Chem. Eng.* **2022**, *36*, 100769. <https://doi.org/10.1016/j.coche.2021.100769>.
- (154) Zapelini de Melo, A. P.; Hoff, R. B.; Molognoni, L.; de Oliveira, T.; Daguer, H.; Manique Barreto, P. L. Disasters with Oil Spills in the Oceans: Impacts on Food

- Safety and Analytical Control Methods. *Food Res. Int.* **2022**, *157*, 111366. <https://doi.org/10.1016/j.foodres.2022.111366>.
- (155) Jiang, M.; Chen, S.; Li, J.; Liu, L. The Biological and Chemical Diversity of Tetramic Acid Compounds from Marine-Derived Microorganisms. *Mar. Drugs* **2020**, *18* (2), 114. <https://doi.org/10.3390/md18020114>.
- (156) Xu, K.; Li, X.-Q.; Zhao, D.-L.; Zhang, P. Antifungal Secondary Metabolites Produced by the Fungal Endophytes: Chemical Diversity and Potential Use in the Development of Biopesticides. *Front. Microbiol.* **2021**, *12*.
- (157) Yan, X.; Liu, J.; Leng, X.; Ouyang, H. Chemical Diversity and Biological Activity of Secondary Metabolites from Soft Coral Genus *Sinularia* since 2013. *Mar. Drugs* **2021**, *19* (6), 335. <https://doi.org/10.3390/md19060335>.
- (158) Puzzarini, C.; Barone, V. Diving for Accurate Structures in the Ocean of Molecular Systems with the Help of Spectroscopy and Quantum Chemistry. *Acc. Chem. Res.* **2018**, *51* (2), 548–556. <https://doi.org/10.1021/acs.accounts.7b00603>.
- (159) Back, H. de M.; Vargas Junior, E. C.; Alarcon, O. E.; Pottmaier, D. Training and Evaluating Machine Learning Algorithms for Ocean Microplastics Classification through Vibrational Spectroscopy. *Chemosphere* **2022**, *287*, 131903. <https://doi.org/10.1016/j.chemosphere.2021.131903>.
- (160) Tripathy, B.; Dash, A.; Das, A. P. Detection of Environmental Microfiber Pollutants through Vibrational Spectroscopic Techniques: Recent Advances of Environmental Monitoring and Future Prospects. *Crit. Rev. Anal. Chem.* **2022**, *0* (0), 1–11. <https://doi.org/10.1080/10408347.2022.2144994>.
- (161) Zhang, X.; Kirkwood, W. J.; Walz, P. M.; Peltzer, E. T.; Brewer, P. G. A Review of Advances in Deep-Ocean Raman Spectroscopy. *Appl. Spectrosc.* **2012**, *66* (3), 237–249. <https://doi.org/10.1366/11-06539>.
- (162) *Raman Spectroscopy in the Deep Ocean: Successes and Challenges*. <https://doi.org/10.1366/0003702041389319>.
- (163) Pirutin, S. K.; Jia, S.; Yusipovich, A. I.; Shank, M. A.; Parshina, E. Y.; Rubin, A. B. Vibrational Spectroscopy as a Tool for Bioanalytical and Biomonitoring Studies. *Int. J. Mol. Sci.* **2023**, *24* (8), 6947. <https://doi.org/10.3390/ijms24086947>.
- (164) Qi, Y.; Hu, D.; Jiang, Y.; Wu, Z.; Zheng, M.; Chen, E. X.; Liang, Y.; Sadi, M. A.; Zhang, K.; Chen, Y. P. Recent Progresses in Machine Learning Assisted Raman Spectroscopy. *Adv. Opt. Mater.* **2023**, *11* (14), 2203104. <https://doi.org/10.1002/adom.202203104>.
- (165) Guo, S.; Popp, J.; Bocklitz, T. Chemometric Analysis in Raman Spectroscopy from Experimental Design to Machine Learning-Based Modeling. *Nat. Protoc.* **2021**, *16* (12), 5426–5459. <https://doi.org/10.1038/s41596-021-00620-3>.
- (166) Ke, J.; Gao, C.; Folgueiras-Amador, A. A.; Jolley, K. E.; de Frutos, O.; Mateos, C.; Rincón, J. A.; Brown, R. C. D.; Poliakov, M.; George, M. W. Self-Optimization of Continuous Flow Electrochemical Synthesis Using Fourier Transform Infrared Spectroscopy and Gas Chromatography. *Appl. Spectrosc.* **2022**, *76* (1), 38–50. <https://doi.org/10.1177/00037028211059848>.

- (167) Ralbovsky, N. M.; Lednev, I. K. Towards Development of a Novel Universal Medical Diagnostic Method: Raman Spectroscopy and Machine Learning. *Chem. Soc. Rev.* **2020**, *49* (20), 7428–7453. <https://doi.org/10.1039/D0CS01019G>.
- (168) Ryzhikova, E.; Ralbovsky, N. M.; Sikirzhyski, V.; Kazakov, O.; Halamkova, L.; Quinn, J.; Zimmerman, E. A.; Lednev, I. K. Raman Spectroscopy and Machine Learning for Biomedical Applications: Alzheimer’s Disease Diagnosis Based on the Analysis of Cerebrospinal Fluid. *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* **2021**, *248*, 119188. <https://doi.org/10.1016/j.saa.2020.119188>.
- (169) Zhang, L.; Li, C.; Peng, D.; Yi, X.; He, S.; Liu, F.; Zheng, X.; Huang, W. E.; Zhao, L.; Huang, X. Raman Spectroscopy and Machine Learning for the Classification of Breast Cancers. *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* **2022**, *264*, 120300. <https://doi.org/10.1016/j.saa.2021.120300>.
- (170) North, N.; Enders, A.; Clark, J.; Allen, H.; Duah, K. Saccharide Concentration Prediction from Proxy Sea Surface Microlayer Samples Analyzed via Infrared Spectroscopy and Quantitative Machine Learning. ChemRxiv January 4, 2024. <https://doi.org/10.26434/chemrxiv-2023-d2ztk-v3>.
- (171) Grooms, A. J.; Burris, B. J.; Badu-Tawiah, A. K. Mass Spectrometry for Metabolomics Analysis: Applications in Neonatal and Cancer Screening. *Mass Spectrom. Rev.* *n/a* (n/a), e21826. <https://doi.org/10.1002/mas.21826>.
- (172) Wang, H.; Liu, J.; Cooks, R. G.; Ouyang, Z. Paper Spray for Direct Analysis of Complex Mixtures Using Mass Spectrometry. *Angew. Chem. Int. Ed.* **2010**, *49* (5), 877–880. <https://doi.org/10.1002/anie.200906314>.
- (173) S. Kulyk, D.; V. Baryshnikov, G.; S. Damale, P.; Maher, S.; K. Badu-Tawiah, A. Charge Inversion under Plasma-Nanodroplet Reaction Conditions Excludes Fischer Esterification for Unsaturated Fatty Acids: A Chemical Approach for Type II Isobaric Overlap. *Chem. Sci.* **2024**, *15* (3), 914–922. <https://doi.org/10.1039/D3SC05369E>.
- (174) Amoah, E.; Kulyk, D. S.; Callam, C. S.; Hadad, C. M.; Badu-Tawiah, A. K. Mass Spectrometry Approach for Differentiation of Positional Isomers of Saccharides: Toward Direct Analysis of Rare Sugars. *Anal. Chem.* **2023**, *95* (13), 5635–5642. <https://doi.org/10.1021/acs.analchem.2c05375>.
- (175) Harvey, G. W.; Burzell, L. A. A Simple Microlayer Method for Small Samples. *Limnol. Oceanogr.* **1972**, *17* (1), 156–157. <https://doi.org/10.4319/lo.1972.17.1.0156>.
- (176) *sklearn.ensemble.RandomForestClassifier*. scikit-learn. <https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> (accessed 2024-01-13).
- (177) *sklearn.ensemble.HistGradientBoostingRegressor*. scikit-learn. <https://scikit-learn/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html> (accessed 2024-01-13).
- (178) Cochran, R. E.; Laskina, O.; Jayarathne, T.; Laskin, A.; Laskin, J.; Lin, P.; Sultana, C.; Lee, C.; Moore, K. A.; Cappa, C. D.; Bertram, T. H.; Prather, K. A.; Grassian, V. H.; Stone, E. A. Analysis of Organic Anionic Surfactants in Fine and Coarse

- Fractions of Freshly Emitted Sea Spray Aerosol. *Environ. Sci. Technol.* **2016**, *50* (5), 2477–2486. <https://doi.org/10.1021/acs.est.5b04053>.
- (179) Volpe, A. M.; Esser, B. K. Real-Time Ocean Chemistry for Improved Biogeochemical Observation in Dynamic Coastal Environments. *J. Mar. Syst.* **2002**, *36* (1), 51–74. [https://doi.org/10.1016/S0924-7963\(02\)00125-2](https://doi.org/10.1016/S0924-7963(02)00125-2).
- (180) Bates, N. R.; Astor, Y. M.; Church, M. J.; Currie, K.; Dore, J. E.; González-Dávila, M.; Lorenzoni, L.; Muller-Karger, F.; Olafsson, J.; Santana-Casiano, J. M. A Time-Series View of Changing Surface Ocean Chemistry Due to Ocean Uptake of Anthropogenic CO₂ and Ocean Acidification. *Oceanography* **2014**, *27* (1), 126–141.
- (181) Takahashi, T.; Sutherland, S. C.; Chipman, D. W.; Goddard, J. G.; Ho, C.; Newberger, T.; Sweeney, C.; Munro, D. R. Climatological Distributions of pH, pCO₂, Total CO₂, Alkalinity, and CaCO₃ Saturation in the Global Surface Ocean, and Temporal Changes at Selected Locations. *Mar. Chem.* **2014**, *164*, 95–125. <https://doi.org/10.1016/j.marchem.2014.06.004>.
- (182) Hagens, M.; Middelburg, J. J. Attributing Seasonal pH Variability in Surface Ocean Waters to Governing Factors. *Geophys. Res. Lett.* **2016**, *43* (24), 12,528–12,537. <https://doi.org/10.1002/2016GL071719>.
- (183) Matoo, O. B.; Lannig, G.; Bock, C.; Sokolova, I. M. Temperature but Not Ocean Acidification Affects Energy Metabolism and Enzyme Activities in the Blue Mussel, *Mytilus Edulis*. *Ecol. Evol.* **2021**, *11* (7), 3366–3379. <https://doi.org/10.1002/ece3.7289>.
- (184) Xue, J.; Lee, C.; Wakeham, S. G.; Armstrong, R. A. Using Principal Components Analysis (PCA) with Cluster Analysis to Study the Organic Geochemistry of Sinking Particles in the Ocean. *Org. Geochem.* **2011**, *42* (4), 356–367. <https://doi.org/10.1016/j.orggeochem.2011.01.012>.
- (185) Alonso-González, I. J.; Aristegui, J.; Lee, C.; Calafat, A. Regional and Temporal Variability of Sinking Organic Matter in the Subtropical Northeast Atlantic Ocean: A Biomarker Diagnosis. *Biogeosciences* **2010**, *7* (7), 2101–2115. <https://doi.org/10.5194/bg-7-2101-2010>.
- (186) Wu, C.; Zhao, X.; Wu, X.; Wen, C.; Li, H.; Chen, X.; Peng, X. Exogenous Glycine and Serine Promote Growth and Antifungal Activity of *Penicillium Citrinum* W1 from the South-West Indian Ocean. *FEMS Microbiol. Lett.* **2015**, *362* (8), fnv040. <https://doi.org/10.1093/femsle/fnv040>.
- (187) Triesch, N.; van Pinxteren, M.; Engel, A.; Herrmann, H. Concerted Measurements of Free Amino Acids at the Cabo Verde Islands: High Enrichments in Submicron Sea Spray Aerosol Particles and Cloud Droplets. *Atmospheric Chem. Phys.* **2021**, *21* (1), 163–181. <https://doi.org/10.5194/acp-21-163-2021>.
- (188) Triesch, N.; van Pinxteren, M.; Salter, M.; Stolle, C.; Pereira, R.; Zieger, P.; Herrmann, H. Sea Spray Aerosol Chamber Study on Selective Transfer and Enrichment of Free and Combined Amino Acids. *ACS Earth Space Chem.* **2021**, *5* (6), 1564–1574. <https://doi.org/10.1021/acsearthspacechem.1c00080>.
- (189) Wu, H.; Liang, C.; Zhang, C.; Chang, H.; Zhang, X.; Zhang, Y.; Zhong, N.; Xu, Y.; Zhong, D.; He, X.; Zhang, L.; Ho, S.-H. Mechanisms and Enhancements on

- Harmful Algal Blooms Conversion to Bioenergy Mediated with Dual-Functional Chitosan. *Appl. Energy* **2022**, 327, 120142.
<https://doi.org/10.1016/j.apenergy.2022.120142>.
- (190) Cai, J.; Chen, M.; Wang, G.; Pan, G.; Yu, P. Fermentative Hydrogen and Polyhydroxybutyrate Production from Pretreated Cyanobacterial Blooms. *Algal Res.* **2015**, 12, 295–299. <https://doi.org/10.1016/j.algal.2015.09.014>.
- (191) Srain, B. M.; Sobarzo, M.; Daneri, G.; González, H. E.; Testa, G.; Farías, L.; Schwarz, A.; Pérez, N.; Pantoja-Gutiérrez, S. Fermentation and Anaerobic Oxidation of Organic Carbon in the Oxygen Minimum Zone of the Upwelling Ecosystem Off Concepción, in Central Chile. *Front. Mar. Sci.* **2020**, 7.
- (192) US EPA, O. *The Effects: Dead Zones and Harmful Algal Blooms*. <https://www.epa.gov/nutrientpollution/effects-dead-zones-and-harmful-algal-blooms> (accessed 2024-01-16).
- (193) *Eutrophication: Causes, Consequences, and Controls in Aquatic Ecosystems | Learn Science at Scitable*. <https://www.nature.com/scitable/knowledge/library/eutrophication-causes-consequences-and-controls-in-aquatic-102364466/> (accessed 2024-01-16).
- (194) Melzner, F.; Thomsen, J.; Koeve, W.; Oeschies, A.; Gutowska, M. A.; Bange, H. W.; Hansen, H. P.; Körtzinger, A. Future Ocean Acidification Will Be Amplified by Hypoxia in Coastal Habitats. *Mar. Biol.* **2013**, 160 (8), 1875–1888.
<https://doi.org/10.1007/s00227-012-1954-1>.
- (195) Zeppenfeld, S.; van Pinxteren, M.; Hartmann, M.; Bracher, A.; Stratmann, F.; Herrmann, H. Glucose as a Potential Chemical Marker for Ice Nucleating Activity in Arctic Seawater and Melt Pond Samples. *Environ. Sci. Technol.* **2019**, 53 (15), 8747–8756. <https://doi.org/10.1021/acs.est.9b01469>.

Appendix A. Supplemental Information for Chapter 3

A.1 Python Scripts

Jupyter notebooks describing data preprocessing and model training are available on GitHub.

https://github.com/Ohio-State-Allen-Lab/Mass_Spec_Functional_Group_ML

A.2 Python Scripts

The mass spectra were web scraped from the NIST webbook using a web scraping implementation, details of which are described in our previous publication.⁷² All the spectra were electron ionization. We obtained a total of 21,166 mass spectra.

A.3 Definition of Functional Groups and Functional Group Classifications

A total of 16 different functional groups were chosen to generate machine learning models to determine the functional group's presence or absence. Single functional groups were chosen in this case to explore how well a machine learning model does at specificity; the goal of this test was to determine if a model can pick up on specific fragments to make an assessment. These functional groups were chosen due to the ease of identifying the functional groups using the InChiKeys that were available to us. **Table S1** shows all these functional groups and the structure of each.

Acyl Halide	Alcohol	Aldehyde	Alkane
	$R-OH$		
Alkene	Alkyl Halide	Alkyne	Amide
	$R-X$		
Amine	Carboxylic Acid	Ester	Ether
Ketone	Methyl	Nitrile	Nitro
	$R-CH_3$	$R-C\equiv N$	

Table S1. Structure of the 16 different functional groups explored during the specific functional group portion of our modeling experiments. In each of these structures the R groups stand for an undefined organic structure attaching the functional group to the rest of the molecule.

After training all of the functional group specific models three functional group classifications were then defined. These classifications were defined to look at multiple possible functional groups at once. The goal of this test was to explore how a machine learning model works with generalizability. Can the models identify a class when all of the constituents of a given class may not be present? **Table S2** defines these classifications by the functional groups within them.

Generalized Functional Group Classifications		
N Containing	O Containing	A Containing
Nitrile Amine Nitro Amide	Aldehyde Carboxylic Acid Ketone Nitro Amide Ether Ester Alcohol	Contains aromatic bonds, alternating single and double bonds

Table S2. Functional groups that make up each of the different functional group classifications.

A.4 Separation of Training and Testing Data

Data Preprocessing

To create the images and the array, the mass spectrometry data files needed to be converted from jcamp-dx files to csv files. This conversion was necessary to complete the preprocessing required to analyze the data both as an array and as images. The jcamp-dx data files only had mass values associated with non-zero intensity. Missing data needed to be added back in with an associated intensity of zero. This was done to ensure that the images being plotted were not incorrectly plotting peaks (**Figure S1**); for the array-based approach, each column needed to be associated with the correct mass value so their column placement would not change with the number of peaks present in a spectrum. Mass values missing between 1 and 500 m/z were added in so that data sets utilizing 250 and 500 mass numbers could be created. After adding in the missing mass values, all the spectra were normalized to ensure that the most intense peak of each spectrum was equal to 1. Jupyter notebooks are available showing the calculation and sorting processes for the mass spec data (https://github.com/Ohio-State-Allen-Lab/MS_Machine_Learning).

Prior to training the functional group specific and the functional group classification models the data needed to be split training and testing data. The test data set is removed first. The test set included 10 of the positive and negative case of each of the model's classes. For example, in the case of the alcohol specific functional group model the test set would include 10 molecules that contained alcohol functional groups, and 10 that did not. Once those had been removed the remaining data was for training into 80% for adjusting the weights and 20% for internal validation during the training process (**Figure S1**).

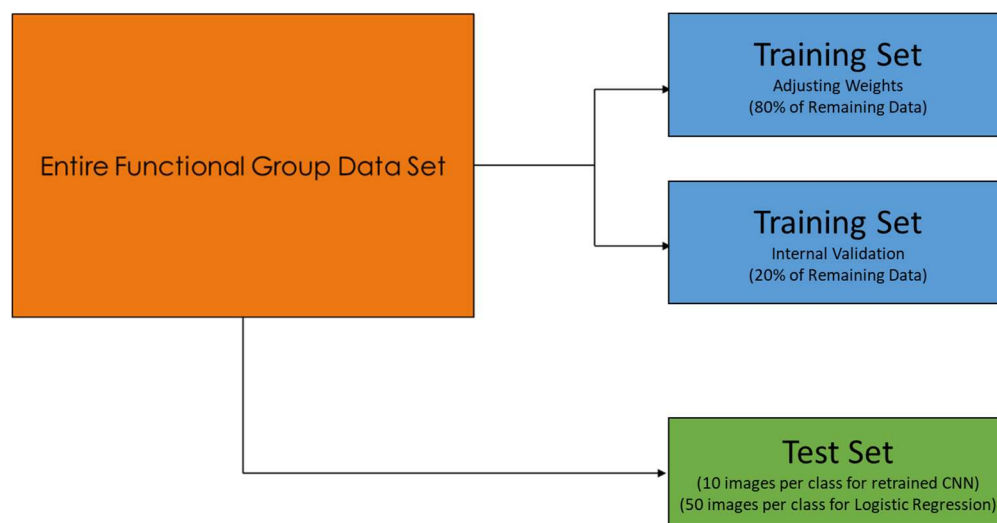


Figure S1. Flowchart depicting the splitting of the functional group dataset into training and testing data sets.

A.5 Trouble Shooting Mass Spectra Plotting

When converting the downloaded files from .jcamp to csv to do both the plotting for the ICNN and for creating the large array for the ALR it was found that the .jcamp files do not have mass values for intensities that are equal to 0. This needed to be corrected before we could move forward with the creation of the datasets. This was necessary for the

ALR so that the placement in the array could be correlated to the same mass value for each spectrum. This also led to problems with the ICNN as well. It is not a requirement for plotting in matplotlib to define every x value however in the case of the sharp discrete peaks in mass spectrometry the plotting program assumed that peaks that were placed close together needed to be connected which led to the plotting of broad triangles. Figure 1 shows an example of this misplotted data.

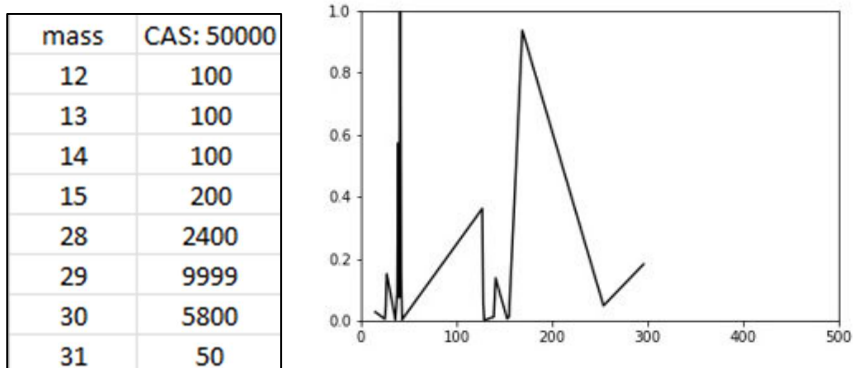


Figure S2. Left: After converting the raw data from .jcamp to .csv from the NIST Webbook only the masses with an associated intensity are reported. Right an example of the 2D representation of the data when plotting in matplotlib without defining all of the mass values with associated intensities of 0.


After training each of the models for the specific functional groups there were two metrics that were then compared for each of the models to determine how well the model had been trained and how well it could identify new data. The first of these metrics was the final training accuracy of the models. This metric defined how well the final model was able to fit the training data. The final testing accuracies of each model is shown in

A6. Accuracies for Transfer Learning CNN Models

Training steps are defined by how many times the model can adjust its weights, biases, or thresholds before needing to present a final model. All models were trained multiple times with different numbers of training steps, the logistic regression models always converged to the same final weights and biases leading to the same estimate for each test spectra regardless of how many steps were made available to them. The convolutional neural networks coupled with the ICNN had different results corresponding to the number of training steps so for those columns the highest accuracy is reported, and the average accuracy is reported within the parentheses.

Table S3. Final training accuracies for all models. These tests seek to show how these different architectures succeed at building specific models for a given functional group. The final training accuracy is defined as the accuracy of the identification of the training data set after the final step in training for the case of the image-based models, and after model convergence for the array-based models.

Final Training Accuracy of CNN Models				
Functional Group	Image 250 MU (%)	Image 500 MU (%)	Array 250 MU (%)	Array 500 MU (%)
Acyl Halide	84 (77)	86 (74.5)	85.5	86.4
Alkyne	92 (89.5)	98 (96.3)	85	85.5
Amide	98(95.3)	96(93.3)	82.9	84.8
Aldehyde	100 (93.3)	100 (97.6)	77.9	79.4
Nitrile	99 (97.3)	100 (96.8)	81.4	82.4
Nitro	66 (56.7)	64 (59.7)	85.2	86.0
Carboxylic Acid	97 (94.2)	99 (95.7)	73.3	74.5
Ketone	89 (81.6)	92 (80.2)	73.3	74.8
Amine	80 (70.3)	74 (61.3)	82.6	83.6
Ester	88 (76.8)	86 (80.7)	74.1	74.5
Alkene	80 (72)	81 (75.2)	76.3	77.5
Alkyl Halide	74 (66.7)	74 (62.8)	78.7	79.5
Alcohol	75 (61.5)	74 (65)	70.5	71.3
Ether	73 (61.1)	84 (69)	68.6	69.4
Alkane	77 (63.6)	67 (57)	81.4	81.8
Methyl	61(56.7)	68 (56.5)	79.9	80.2



The next metric used to describe the model's fitness was the final testing accuracy. While the final training accuracy showed how well the model fit the training data, the final testing accuracy describes how well the model does at identifying new previously unseen data. This is measured by testing the models with the reserved testing data that it had never seen before. The models predicted if each spectra contained or did not contain the given functional group. Then the accuracy of the correct identifications out of the 20 reserved test spectra was used to calculate the final test accuracy. These final test accuracies are shown in

Table S4.

Table S4. Final testing accuracies for each of the functional group models are shown here. These tests seek to show how these different architectures succeed at building specific models for a given functional group. Final testing accuracies are defined as the final accuracy of identification of the test data, data that was withheld from the entire training and testing process.

Final Testing Accuracy of CNN Models				
Functional Group	Image 250 MU (%)	Image 500 MU (%)	Array 250 MU (%)	Array 500 (MU) (%)
Acyl Halide	55 (50.8)	65 (53.3)	80	75
Alkyne	50 (50)	80 (75)	80	80
Amide	(50) 50	60 (43.3)	70	70
Aldehyde	50 (50)	65 (56.6)	75	75
Nitrile	50 (50)	65 (51.7)	85	85
Nitro	60 (51.7)	60 (50.8)	80	80
Carboxylic Acid	50 (50)	50 (41.6)	65	65
Ketone	50 (50)	55 (50.8)	70	70
Amine	60 (51.6)	50 (50)	60	45
Ester	50 (50)	60 (51.6)	70	70
Alkene	50 (49.2)	55 (50.8)	90	85
Alkyl Halide	55 (54.1)	65 (58.3)	80	80
Alcohol	50 (50)	55 (51.7)	60	55
Ether	50 (50)	55 (50.83)	65	70
Alkane	70 (58.3)	60 (54.2)	60	60
Methyl	55 (52.5)	80 (58.3)	90	90



This process was then repeated for the generalized functional group classification models. The final training accuracies are shown in **Table S5**.

Table S5. Final training accuracy for the models for the functional group classifications is presented here. The goal of these tests is to show how these architectures can create models to fit a more generalized classification of functional groups. The final training accuracy is defined as the accuracy of identification of the training dataset after the last training step in the case of the image- based approach, and after model convergence in the array-based approach. The highest accuracy is highlighted and in the case of the image-based approach multiple models were trained with different parameters and the average of all the different models for each functional group is shown in the parenthetical.

Final Training Accuracy of CNN Models				
Functional Group	Image 250 MU (%)	Image 500 MU (%)	Array 250 MU (%)	Array 500(MU) (%)
A Containing	87(77.7)	86(81)	90.4	90.6
O Containing	68(56.3)	65(57.0)	90.6	68.9
N Containing	87 (72.5)	84(74.8)	79.2	79.2

Table S6 contains the final testing accuracies for the generalized functional group classification models.

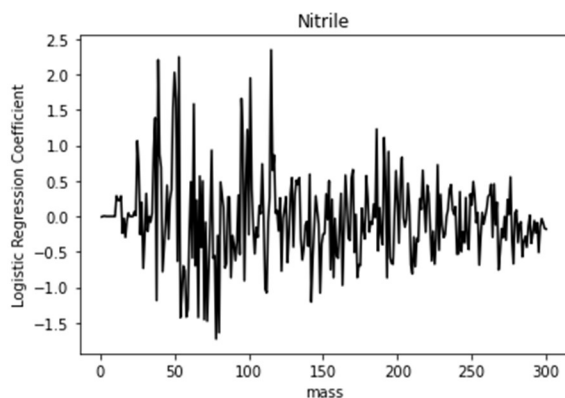
Table S6. Final testing accuracies for each of the functional group models. These models seek to show how these different architectures succeed at building generalized models for a given functional group classification. Final testing accuracies are defined as the final accuracy of identification of the test data, data that was withheld from the entire training and testing process. A higher accuracy at identifying the test data suggests an increased ability to generalize from the training and testing data and thus suggests an increased ability to identify novel data in the future. The highest final testing accuracy for a given functional group is highlighted. The image-based models were trained under multiple different training parameters, because of this the highest accuracy is reported and the average accuracy is within the parenthetical.

Final Testing Accuracy of CNN Models				
Functional Group	Image 250 MU (%)	Image 500 MU (%)	Array 250 MU (%)	Array 500 MU (%)
A Containing	70(57.5)	100(90.0)	95	95
O Containing	55(47.5)	50(45.8)	75	75
N Containing	50(48.3)	55(50)	80	80

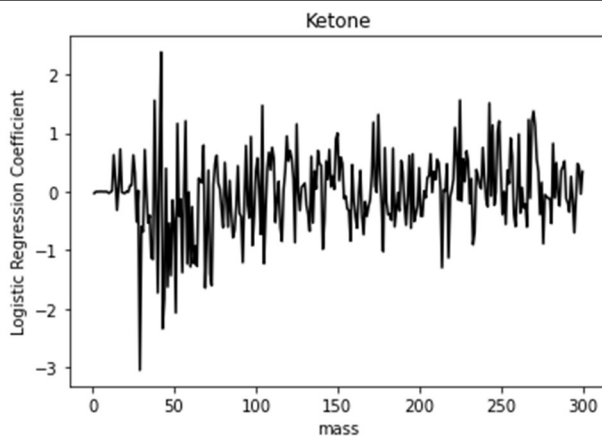
A.7 Model Coefficients for Each of the Models

Once all of the models had been trained coefficients were assigned to each of the features (in our case those features were the individual masses). These coefficients when plotted show the effect of each mass on the model's assignment of the spectra. A positive value signifies that that peak is associated with the presence of that functional group, a

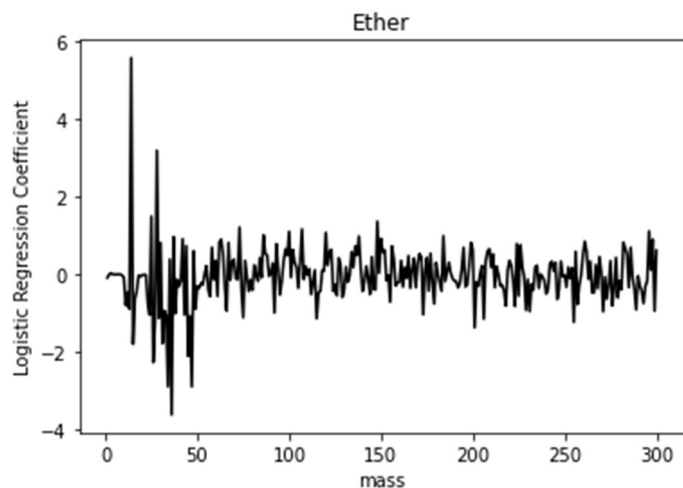
negative value suggests that that peak is associated with the absence of that functional group. The intensity in either direction suggests how heavily that mass effects the model's assignments.



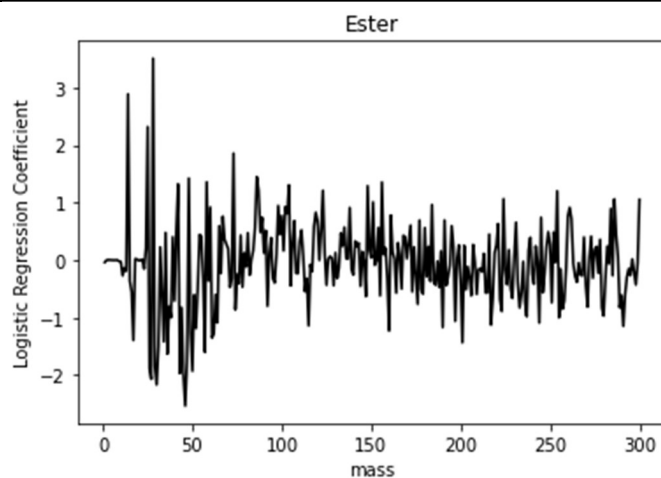
Nitrile			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
115	2.349312	54	-1.42746
53	2.248544	70	-1.45842
39	2.208312	72	-1.47921
50	2.029428	80	-1.64241
101	1.95127	78	-1.73067



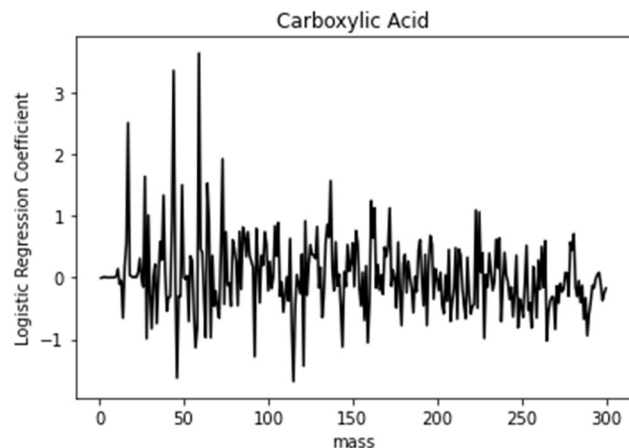
Ketone			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
42	2.381197	40	-1.72529
225	1.561286	44	-1.82377
38	1.557563	51	-2.06585
243	1.51114	43	-2.33556
104	1.470557	29	-3.04368



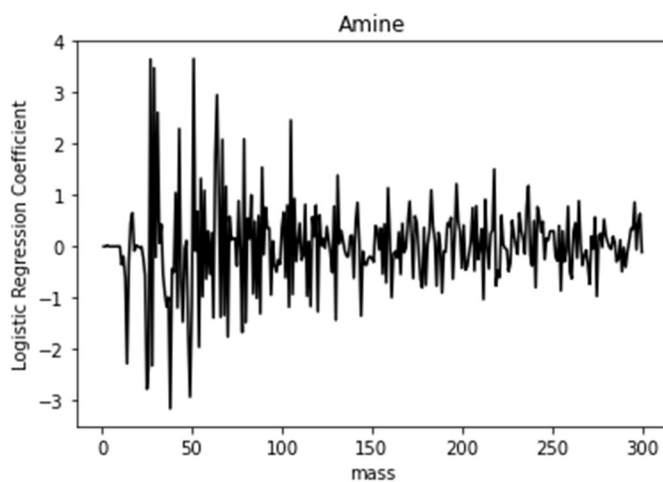
Ether			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
14	5.562795	45	-2.09552
28	3.1815	26	-2.26069
25	1.492238	34	-2.88094
148	1.360295	47	-2.88284
73	1.209785	36	-3.60599



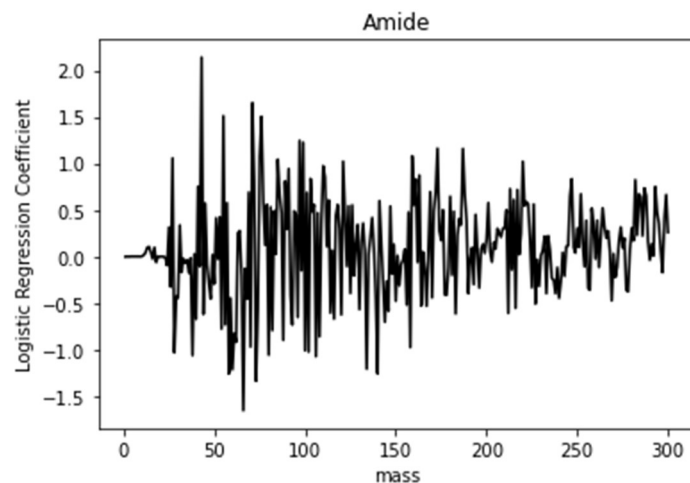
Ester			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
28	3.512014	43	-1.96859
14	2.89041	45	-2.01347
25	2.320361	27	-2.07046
73	1.858746	30	-2.16861
86	1.448884	46	-2.53712



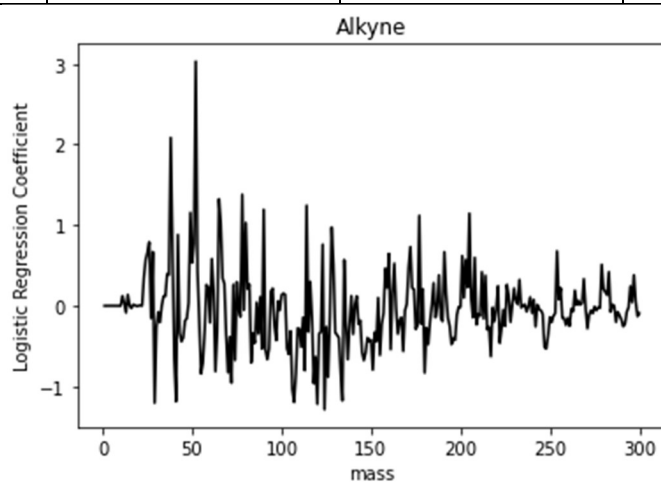
Carboxylic Acid			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
59	3.636344	57	-1.13647
44	3.354331	92	-1.28263
17	2.505814	121	-1.43412
73	1.922039	46	-1.62928
27	1.636504	115	-1.68373



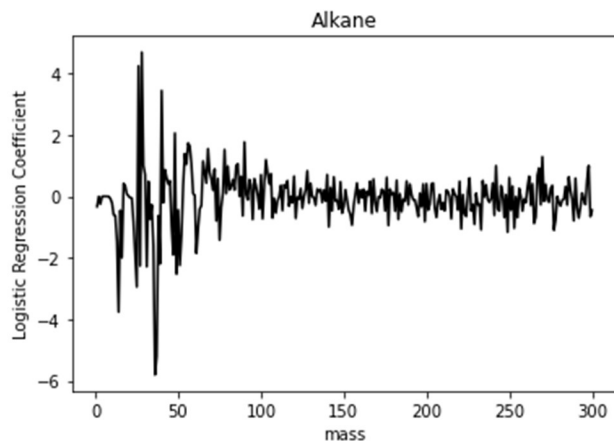
Amine			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
51	3.661088	26	-2.31438
27	3.649376	28	-2.3205
29	3.480256	25	-2.77679
64	2.954008	49	-2.92806
31	2.614077	38	-3.15894



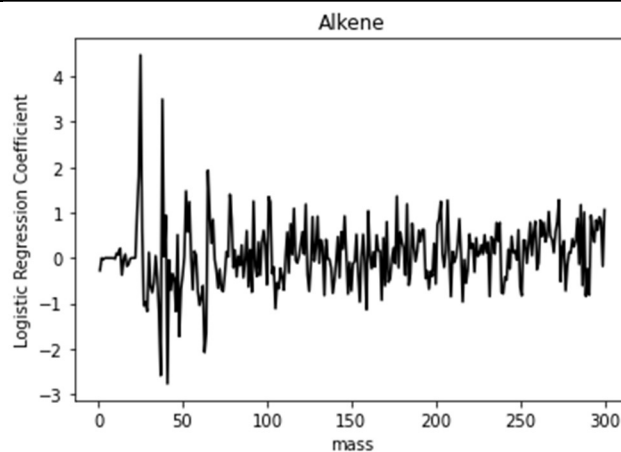
Amide			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
43	2.140079	60	-1.21471
71	1.6496	140	-1.25772
55	1.510346	58	-1.2585
76	1.503447	73	-1.33673
97	1.245007	66	-1.65425



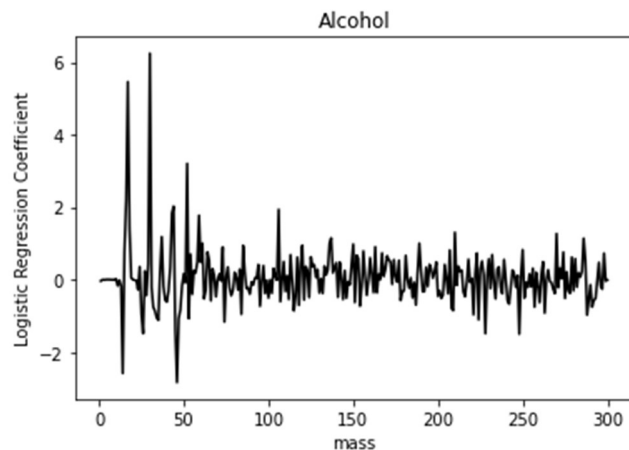
Alkyne			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
52	3.031475	41	-1.18556
38	2.085062	107	-1.19687
78	1.379646	29	-1.20538
65	1.323678	120	-1.21927
114	1.244326	124	-1.28709



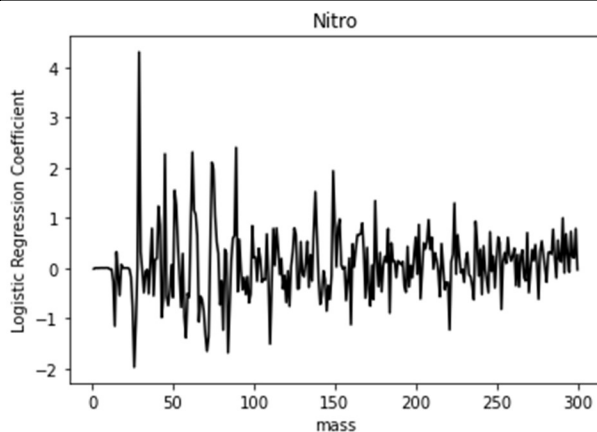
Alkane			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
28	4.694609	49	-2.52253
26	4.245265	25	-2.94078
40	3.444903	14	-3.75927
48	2.060636	37	-5.20325
90	1.7719	36	-5.80462



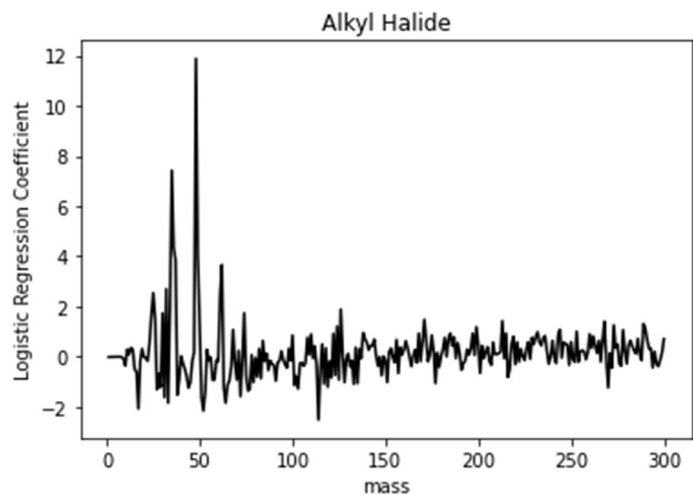
Alkene			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
25	4.472632	64	-1.64609
38	3.49691	48	-1.72455
65	1.930545	63	-2.07894
24	1.801369	37	-2.59039
52	1.475922	41	-2.77066



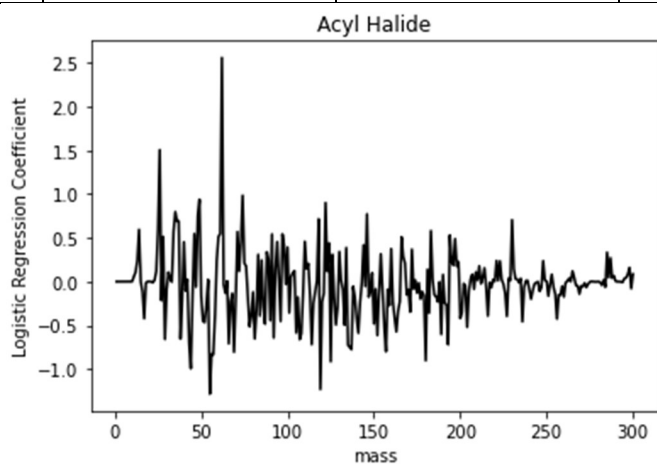
Alcohol			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
30	6.248085	26	-1.48686
17	5.462362	228	-1.48941
52	3.206055	248	-1.50221
16	2.197711	14	-2.58611
44	2.025667	46	-2.83862



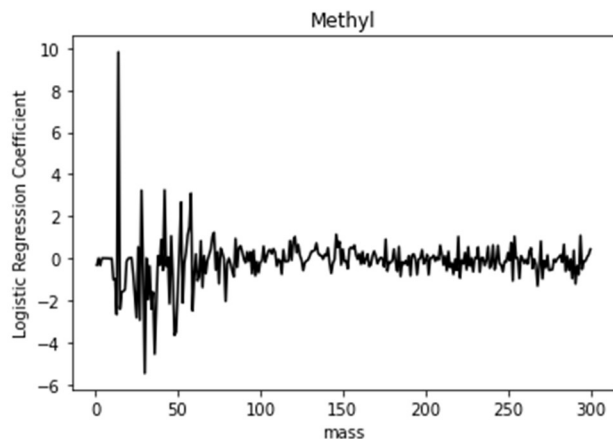
Nitro			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
29	4.299858	58	-1.38806
89	2.397963	110	-1.51309
62	2.305777	71	-1.6563
45	2.27186	84	-1.68743
74	2.10265	26	-1.97449



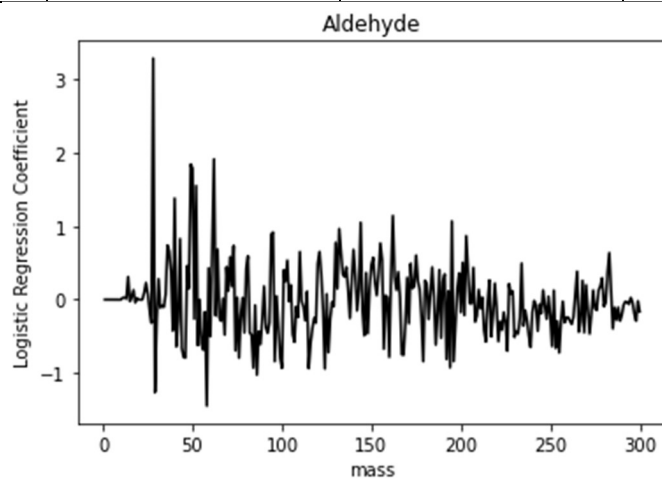
Alkyl Halide			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
48	11.87642	64	-1.83414
35	7.423916	33	-1.83426
36	4.362391	17	-2.06371
49	4.032308	52	-2.1476
37	3.83969	114	-2.50603



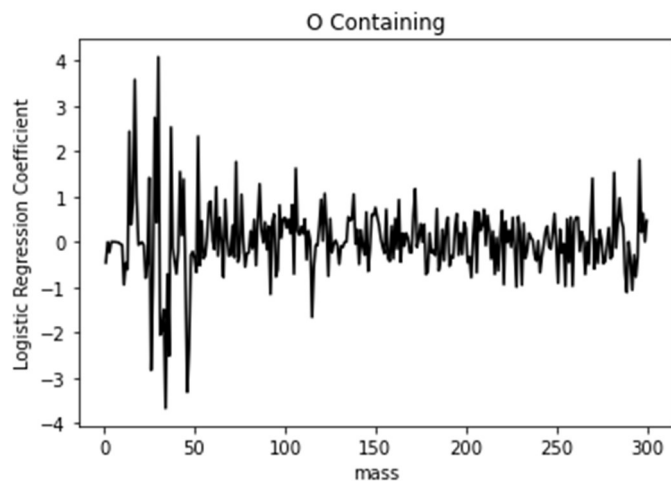
Acyl Halide			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
62	2.557588	180	-0.90233
26	1.504645	125	-0.91391
74	0.984313	44	-0.99313
49	0.937561	119	-1.23031
122	0.900851	55	-1.28368



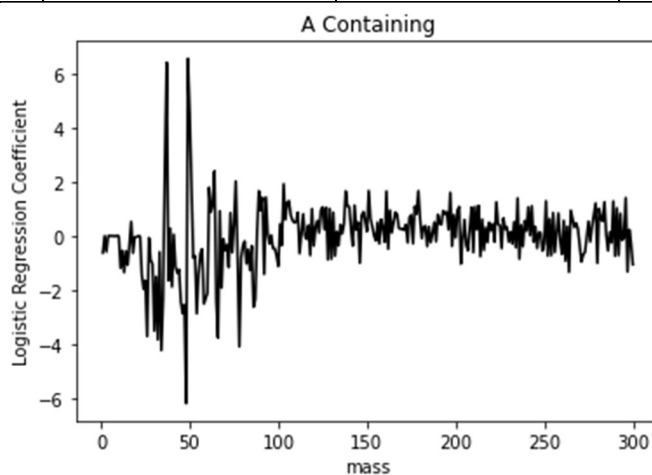
Methyl			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
14	9.811189	27	-2.94057
42	3.24801	49	-3.51
28	3.22984	48	-3.66074
58	3.090506	36	-4.55675
52	2.661909	30	-5.47522



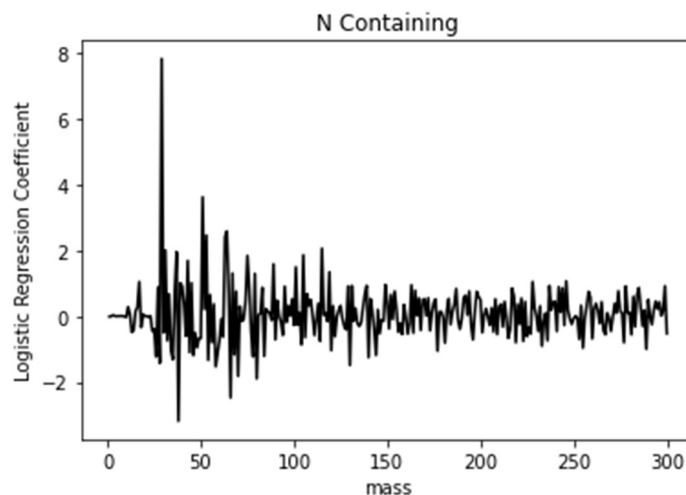
Aldehyde			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
28	3.290154	115	-0.93832
62	1.915359	124	-0.94284
49	1.843428	86	-1.02607
50	1.789735	29	-1.2638
52	1.551895	58	-1.44946



O Containing			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
30	4.081637	47	-2.3676
17	3.580336	36	-2.51273
28	2.744514	26	-2.82851
37	2.531534	46	-3.31254
14	2.439157	34	-3.66808



A Containing			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
30	4.081637	47	-2.3676
17	3.580336	36	-2.51273
28	2.744514	26	-2.82851
37	2.531534	46	-3.31254
14	2.439157	34	-3.66808



N Containing			
Largest Coefficients		Smallest Coefficients	
Mass	Coefficient	Mass	Coefficient
29	7.838848	58	-1.53219
51	3.624683	70	-1.82283
64	2.585419	80	-1.90367
53	2.460293	66	-2.48063
63	2.416839	38	-3.19508

A.8 Identified Impactful Fragments for Mass Spectral Analysis

Once we switched to logistic regression our accuracies both for training and testing increased. Reported in **Table S7**. Top five most beneficial peaks for analyzing each of the different functional groups and functional group classifications. The beneficial peaks were defined as the peaks that caused the largest decrease in final training accuracy when removed from the feature set., are those accuracies.

Once the final accuracies for each of the models had been determined the focus shifted to understanding how the models made their assignments. Reported below are the

peaks that were most positively impactful to the models correctly identifying the functional groups. The difference in training accuracy was defined as in **Equation S1**.

Equation S1: Calculating delta accuracy.

$$\Delta Accuracy_j = Accuracy_{All\ Masses\ Present} - Accuracy_{model\ missing\ mass\ j}$$

In the above equation j is defined as the mass that has been removed for that iteration of the feature analysis. The accuracies are final training accuracies. These were chosen to understand how the model is making its fit of the training data.

Table S7. Top five most beneficial peaks for analyzing each of the different functional groups and functional group classifications. The beneficial peaks were defined as the peaks that caused the largest decrease in final training accuracy when removed from the feature set.

Most Helpful Peaks (Led to a Decrease in Accuracy When Removed)		
Alcohol	Training Accuracy	Difference in Training Accuracy
106	70.06	-0.47
52	70.05	-0.48
43	69.91	-0.62
44	69.71	-0.82
30	68.15	-2.38
Aldehyde	Training Accuracy	Difference in Training Accuracy
73	78.17	-0.90
110	78.17	-0.90
80	77.94	-1.13
95	77.94	-1.13
28	76.47	-2.60
Alkane	Training Accuracy	Difference in Training Accuracy

57	81.19	-0.23
27	81.17	-0.25
40	80.93	-0.49
28	80.92	-0.50
26	80.77	-0.65
Alkene	Training Accuracy	Difference in Training Accuracy
101	76.34	-0.29
65	76.22	-0.41
54	76.07	-0.56
41	76.04	-0.59
38	75.82	-0.82
Alkyl Halide	Training Accuracy	Difference in Training Accuracy
100	78.81	-0.30
126	78.80	-0.31
35	78.49	-0.63
62	78.46	-0.65
48	77.73	-1.39
Alkyne	Training Accuracy	Difference in Training Accuracy
66	84.51	-0.78
40	84.38	-0.91
78	84.38	-0.91
114	84.24	-1.04
52	83.85	-1.43
Amide	Training Accuracy	Difference in Training Accuracy
43	83.16	-0.52
71	83.16	-0.52
118	83.03	-0.65
119	82.90	-0.78
76	82.64	-1.04
Amine	Training Accuracy	Difference in Training Accuracy
104	82.66	-0.30
27	82.51	-0.45
105	82.49	-0.48
51	82.39	-0.58
29	82.31	-0.65

Carboxylic Acid	Training Accuracy	Difference in Training Accuracy
29	73.54	-0.28
106	73.54	-0.28
163	73.50	-0.32
59	73.40	-0.42
44	72.66	-1.16
Ester	Training Accuracy	Difference in Training Accuracy
27	73.74	-0.31
104	73.66	-0.40
14	73.50	-0.56
42	73.50	-0.56
28	72.93	-1.13
Ether	Training Accuracy	Difference in Training Accuracy
36	68.57	-0.22
42	68.49	-0.31
56	68.49	-0.31
14	67.70	-1.09
28	67.10	-1.69
Ketone	Training Accuracy	Difference in Training Accuracy
44	73.60	-0.68
40	73.57	-0.71
43	73.54	-0.74
29	73.43	-0.85
42	71.27	-3.01
Methyl	Training Accuracy	Difference in Training Accuracy
27	79.26	-0.71
28	79.02	-0.96
42	78.85	-1.12
30	78.73	-1.25
14	78.40	-1.58
Nitrile	Training Accuracy	Difference in Training Accuracy
67	81.34	-0.59
95	81.34	-0.59
66	81.26	-0.67

78	81.26	-0.67
72	81.09	-0.84
Nitro	Training Accuracy	Difference in Training Accuracy
120	85.20	-0.24
149	85.14	-0.29
42	85.08	-0.35
62	85.08	-0.35
29	82.49	-2.95
A Containing	Training Accuracy	Difference in Training Accuracy
68	90.25	-0.22
50	90.25	-0.22
66	90.20	-0.27
42	90.09	-0.38
78	89.96	-0.51
N Containing	Training Accuracy	Difference in Training Accuracy
53	78.39	-0.42
38	78.32	-0.49
43	78.21	-0.60
105	78.10	-0.71
29	76.26	-2.54
O Containing	Training Accuracy	Difference in Training Accuracy
46	68.74	-0.35
26	68.73	-0.36
30	68.46	-0.64
42	68.32	-0.78
28	68.14	-0.95

The same was done with the most negatively correlated peaks with accuracy. Removing these masses led to an increase in accuracy suggesting that they are confusing the model and generating a worse fit of the data by being present.

Table S8. Top five most hindbersome peaks for analyzing each of the different functional groups and functional group classifications. The hindbersome peaks were defined as the peaks that caused the largest increase in final training accuracy when removed from the feature set.

Most Hindbersome Peaks (Led to an Increase in Accuracy When Removed)		
Alcohol	Training Accuracy	Difference in Training Accuracy
106	70.69	+0.16
52	70.63	+0.10
43	70.62	+0.09
44	70.62	+0.09
30	70.62	+0.09
Aldehyde	Training Accuracy	Difference in Training Accuracy
73	79.41	+0.34
110	79.19	+0.11
80	79.19	+0.11
95	79.19	+0.11
28	79.19	+0.11
Alkane	Training Accuracy	Difference in Training Accuracy
57	81.52	+0.11
27	81.52	+0.11
40	81.52	+0.10
28	81.51	+0.09
26	81.51	+0.09
Alkene	Training Accuracy	Difference in Training Accuracy
101	76.82	+0.19
65	76.74	+0.11
54	76.70	+0.07
41	76.70	+0.07
38	76.70	+0.07
Alkyl Halide	Training Accuracy	Difference in Training Accuracy

100	79.22	+0.11
126	79.22	+0.11
35	79.20	+0.09
62	79.20	+0.09
48	79.19	+0.08
Alkyne	Training Accuracy	Difference in Training Accuracy
66	85.94	+0.65
40	85.55	+0.26
78	85.55	+0.26
114	85.55	+0.26
52	85.42	+0.13
Amide	Training Accuracy	Difference in Training Accuracy
43	84.97	+1.30
71	84.59	+0.91
118	84.33	+0.65
119	84.33	+0.65
76	84.20	+0.52
Amine	Training Accuracy	Difference in Training Accuracy
104	83.14	+0.18
27	83.09	+0.13
105	83.09	+0.13
51	83.07	+0.10
29	83.07	+0.10
Carboxylic Acid	Training Accuracy	Difference in Training Accuracy
29	74.33	+0.51
106	74.28	+0.46
163	74.19	+0.37
59	74.19	+0.37
44	74.19	+0.37
Ester	Training Accuracy	Difference in Training Accuracy
27	74.23	+0.17
104	74.20	+0.14
14	74.18	+0.12
42	74.18	+0.12
28	74.16	+0.10

Ether	Training Accuracy	Difference in Training Accuracy
36	69.04	+0.25
42	69.03	+0.23
56	69.01	+0.21
14	69.00	+0.20
28	68.98	+0.18
Ketone	Training Accuracy	Difference in Training Accuracy
44	74.45	+0.17
40	74.42	+0.14
43	74.42	+0.14
29	74.39	+0.11
42	74.39	+0.11
Methyl	Training Accuracy	Difference in Training Accuracy
27	80.04	+0.07
28	80.03	+0.06
42	80.02	+0.04
30	80.02	+0.04
14	80.02	+0.04
Nitrile	Training Accuracy	Difference in Training Accuracy
67	82.35	+0.42
95	82.18	+0.25
66	82.18	+0.25
78	82.18	+0.25
72	82.18	+0.25
Nitro	Training Accuracy	Difference in Training Accuracy
120	85.91	+0.47
149	85.85	+0.41
42	85.79	+0.35
62	85.79	+0.35
29	85.73	+0.29
A Containing	Training Accuracy	Difference in Training Accuracy
68	90.53	+0.06
50	90.51	+0.03
66	90.49	+0.02

42	90.49	+0.02
78	90.49	+0.01
N Containing	Training Accuracy	Difference in Training Accuracy
53	78.99	+0.18
38	78.89	+0.08
43	78.88	+0.07
105	78.88	+0.07
29	78.86	+0.06
O Containing	Training Accuracy	Difference in Training Accuracy
46	69.21	+0.12
26	69.19	+0.10
30	69.19	+0.09
42	69.18	+0.09
28	69.18	+0.09

A9. Analysis of Specific NIST Compounds

To show how our models may be used to analyze mass spectra quickly, we took some specific examples from the NIST database that would be of particular interest for planetary missions, amino acids. We ran these spectra through our 20 models and report those results in table S9. The accuracy of the models' description of the molecule is about 75% accurate with the more chemically complex amino acid generating more positive assignments from the models.

Table S9. Model results for looking at two different amino acids, tryptophan and histidine.

The overall description of the molecules was approximately 75% accurate.

	Tryptophan	Histidine
Models Identified Present	Nitrile Carboxylic Acid Amine Acyl Halide Nitrogen Containing Aromatic Containing	Nitrile Ester Amine Aromatic containing
Models Identified Absent	Ketone Ether Ester Amide Alkyne Alkane Alkene Alcohol Nitro Alkyl Halide Methyl Aldehyde Oxygen Containing	Ketone Ether Carboxylic Acid Amide Alkyne Alkane Alkene Alcohol Nitro Alkyl Halide Acyl Halide Methyl Aldehyde Nitrogen Containing Oxygen Containing
Accuracy	16/19 Correct Results	14/19 Correct Results

We also analyzed our models on their ability to identify experimental data outside of the NIST database. We took GC MS data of 2-furan methanol, limonene, and pyridine. When analyzed with all 20 models the final accuracy in describing the molecules was ~80%.

Table S10. Model results for experimental data outside of the NIST dataset. The description of the functional groups for all three was approximately 80%.

	2 Furan Methanol	Limonene	Pyridine
Models Identified Present	Alcohol Alkene Alkane O Containing Ketone Alkyne Aldehyde	Alkane Alkene Methyl Alkyne Alcohol O Containing	Aromatic Containing Nitrogen Containing Nitrile Amide Alkyl Halide
Models Identified Absent	Nitrile Ketone Ester Ether Carboxylic Acid Amine Amide Alkyne Nitro Alkyl Halide Acyl Halide Aldehyde Methyl Nitrogen Containing Aromatic Containing	Nitrile Ketone Ether Ester Carboxylic Acid Amine Amide Alkyne Alcohol Nitro Alkyl Halide Acyl Halide Aldehyde Nitrogen Containing Aromatic Containing	Ketone Ether Ester Carboxylic Acid Amine Alkyne Alkane Alkene Alcohol Nitro Acyl Halide Methyl Aldehyde Nitrogen Containing
Accuracy	15/19 Correct Results	16/19 Correct Results	16/19 Correct Results

Appendix B. Supplemental Information for Chapter 4

B.1 ATR-FTIR Spectra of All Training Samples

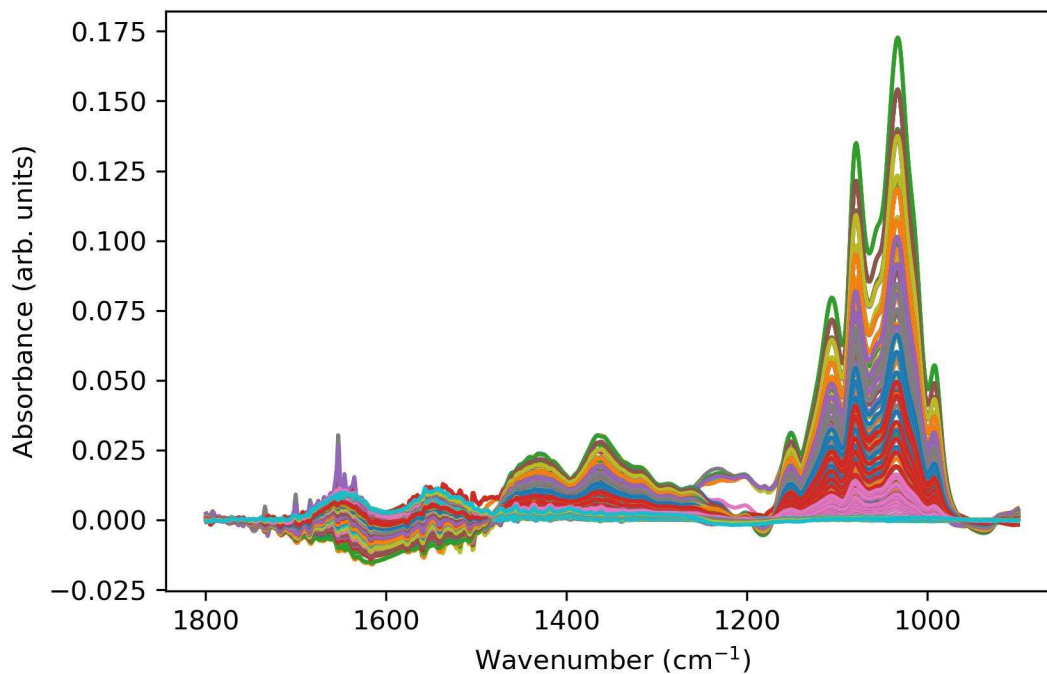


Figure S3. Composite spectra of all 100 samples used for training in each machine learning model. Negative intensities between 1700 and 1500 cm⁻¹ arise from the subtraction of water from the samples as a preprocessing step.

B.2 Vibrational Analysis of Glucose and ESA

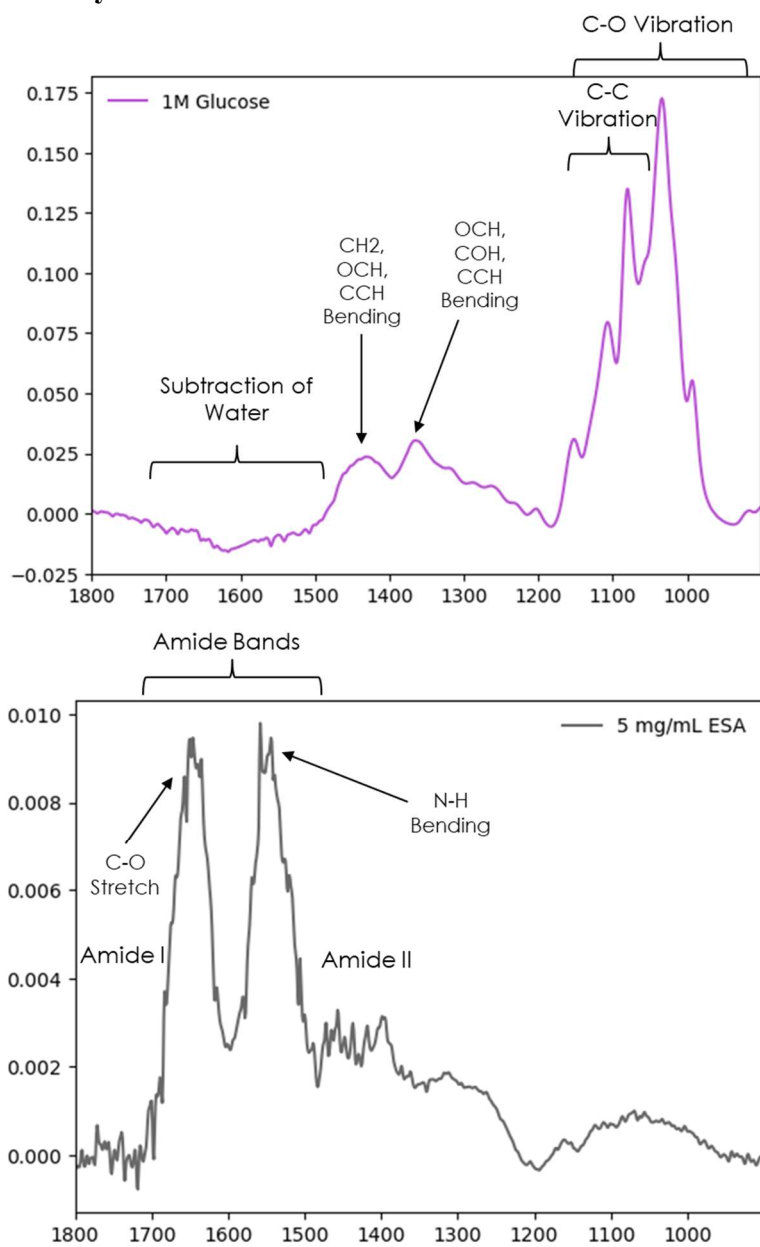


Figure S4. Vibrational analysis of glucose (top) and ESA (bottom).

B.3 Highest Concentration of ESA and Glucose

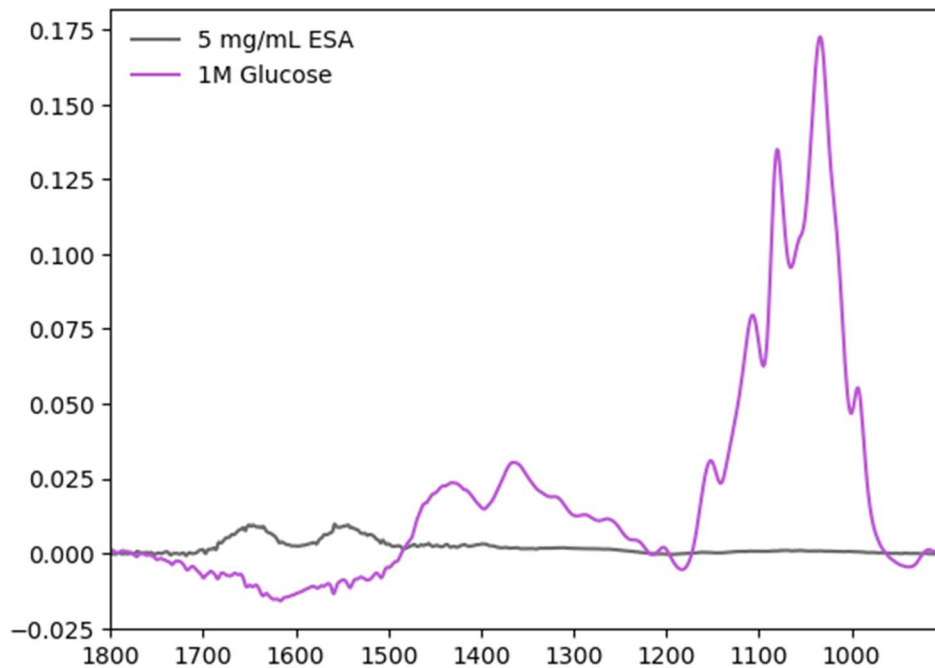


Figure S5. ATR-FTIR spectra of 1 M glucose and 5 mg/mL egg serum albumin isolated in aqueous solution. Glucose's negative band occurring at the same location of the amide bands of the ESA is likely the cause of the lack of any contribution from these bands in the contour plot in **Figure 26**.

B.4 Selected Single Peak Beer's Law Analysis

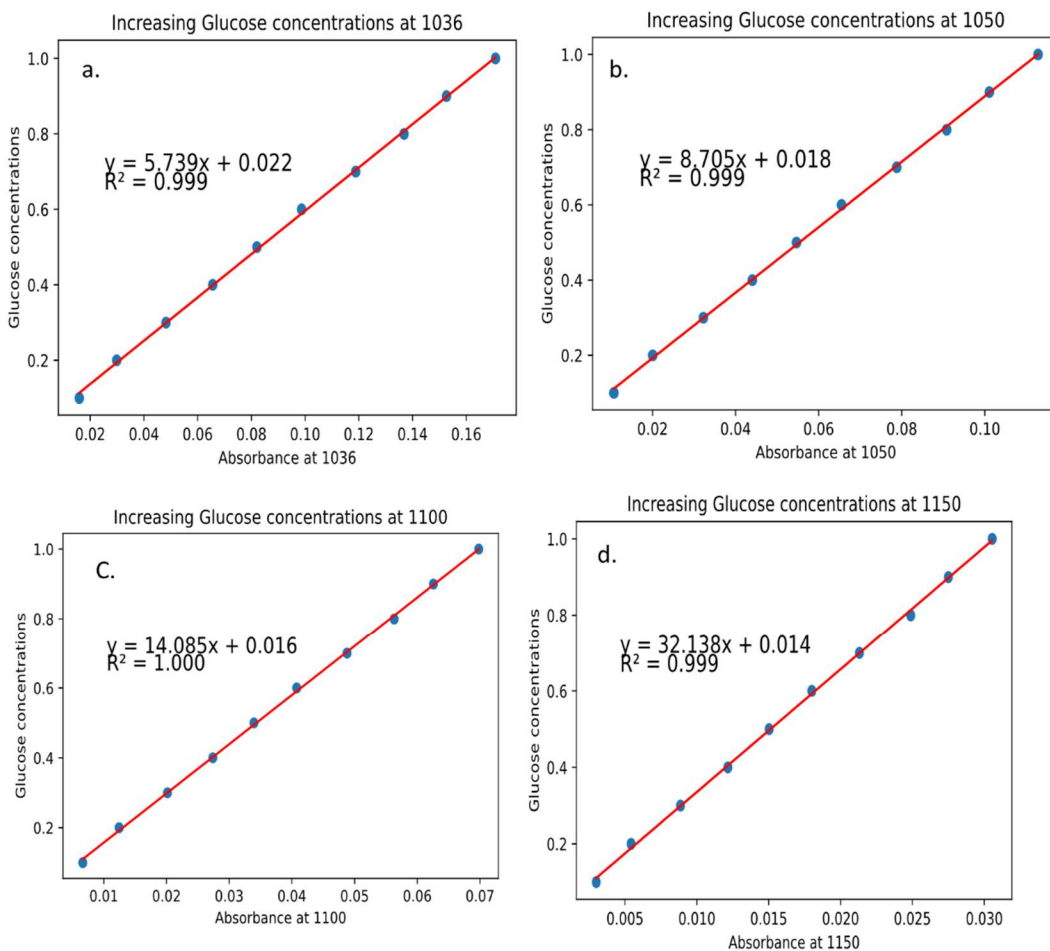


Figure S6. Linear fits of 1036 cm^{-1} (a), 1050 cm^{-1} (b), 1100 cm^{-1} (c), and 1150 cm^{-1} (d).

All linear fits exhibit high R^2 values.

Table S11. Results of the linear fits from Figure S3 on identifying the concentration of saccharide (glucose and sucrose) from the ocean proxy samples. These results show that an individual linear fit is insufficient to identifying the generalized concentration of saccharide in aqueous solution.

Absorbance (cm ⁻¹)	Known Concentration (M)	Predicted concentration (M)
1036	0.200	0.101
1036	0.100	0.101
1150	0.200	0.108
1150	0.100	0.109

B.5 Analysis of MLR and SVR Weights

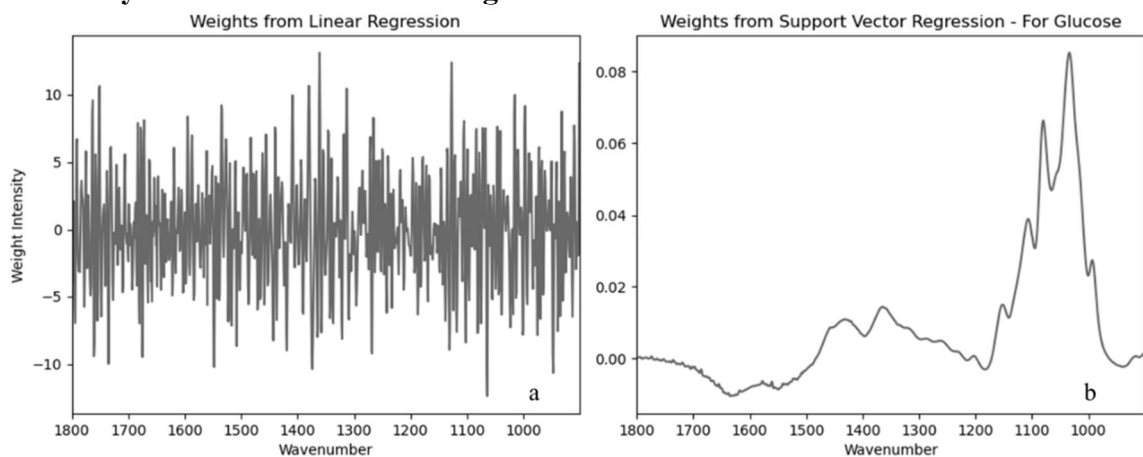


Figure S7. Final model weights of linear regression (a) and support vector regression (b).

The weights for the LR show no correlation with vibrational spectra, or anything with chemical significance. In contrast, the SVR weights show correspondence with the vibrational modes of glucose. These results suggest that the LR is over fitting and does not have the same “understanding” of the chemical system. Note SVR models were trained with a linear kernel and a Radial Bias Function (RBF) kernel. The RBF kernel was used to

produce our final assessments of the lab proxy samples in Table 3. (The weights for an RBF kernel cannot be visualized in the same way that we can for a linear kernel even though the model accuracies are comparable. The linear kernel weights are evaluated in this figure.)

B.6 Tabulated Values for Accuracy and Fit for Each ML Model

Table S12. Numerical results from the error and fit analysis for each ML method.

	<i>MLR</i>	<i>KNN</i>	<i>DT</i>	<i>GBR</i>	<i>MLP</i>	<i>SVR</i>
<i>Training Error (M)</i>	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000
<i>Testing Error (M)</i>	0.0060	0.020089	0.003333	0.025101	0.028260	0.021941
<i>Validation Error (M)</i>	0.0039	0.016000	0.003111	0.022554	0.023631	0.016585
<i>R² Value</i>	0.999295	0.945941	0.996691	0.971263	0.932400	0.955016

B.7 Concentration Predictions for Proxy Solutions for Each ML Model

Table S13. Numerical results from the estimates of the models for the lab proxy samples.

	<i>MLR</i>	<i>KNN</i>	<i>DT</i>	<i>GBR</i>	<i>MLP</i>	<i>SVR</i>
<i>Sample A</i> (0.2002 M)	0.1130791	0.078	0.2	0.14260666	0.18047648	0.19378671
<i>Sample B</i> (0.1502 M)	0.08558797	0.068	0.12	0.09309408	0.13607501	0.13939388
<i>Sample B</i> (0.1001 M)	0.05907225	0.046	0.05	0.06630661	0.10100766	0.09224692

Appendix C. Supplemental Information for Chapter 5

C.1 Raman Spectra Before and After Preprocessing

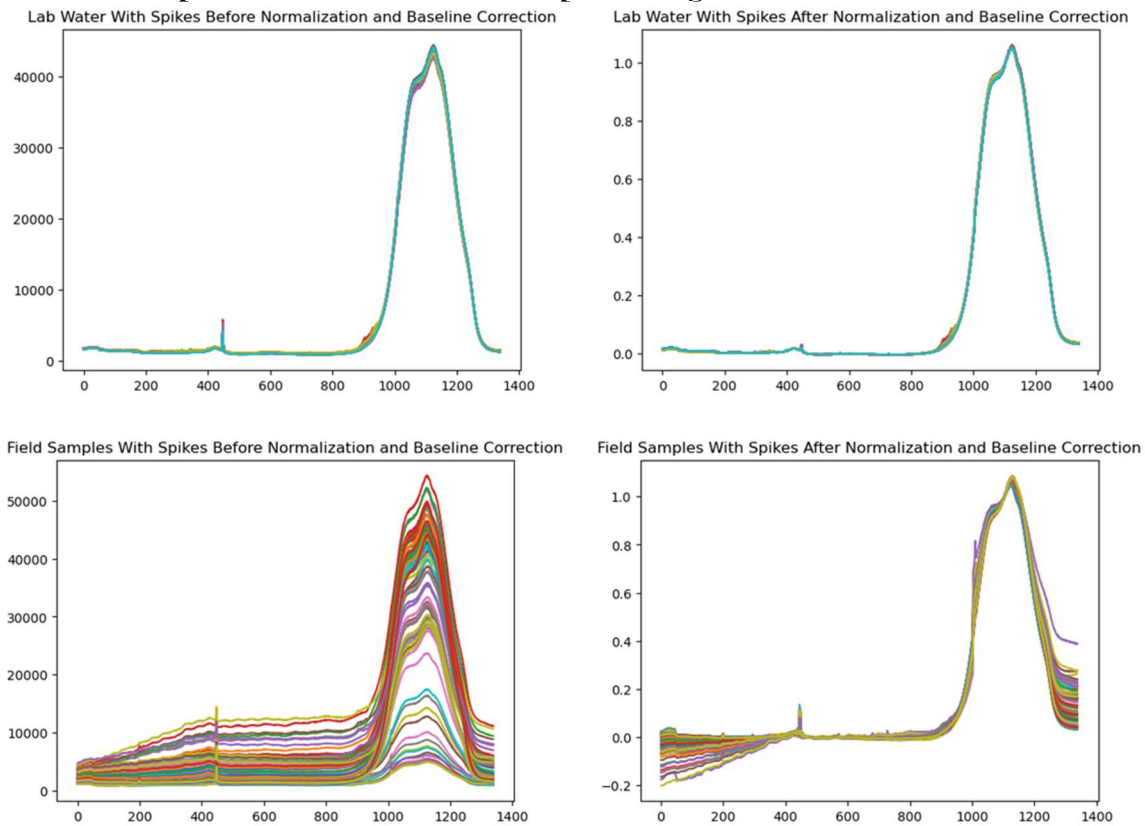


Figure S8. Spiked lab dataset (top) and spiked ocean dataset (bottom) before (left) and after data preprocessing (right).

C.2 Calibration Curves for Mass Spectral Analysis

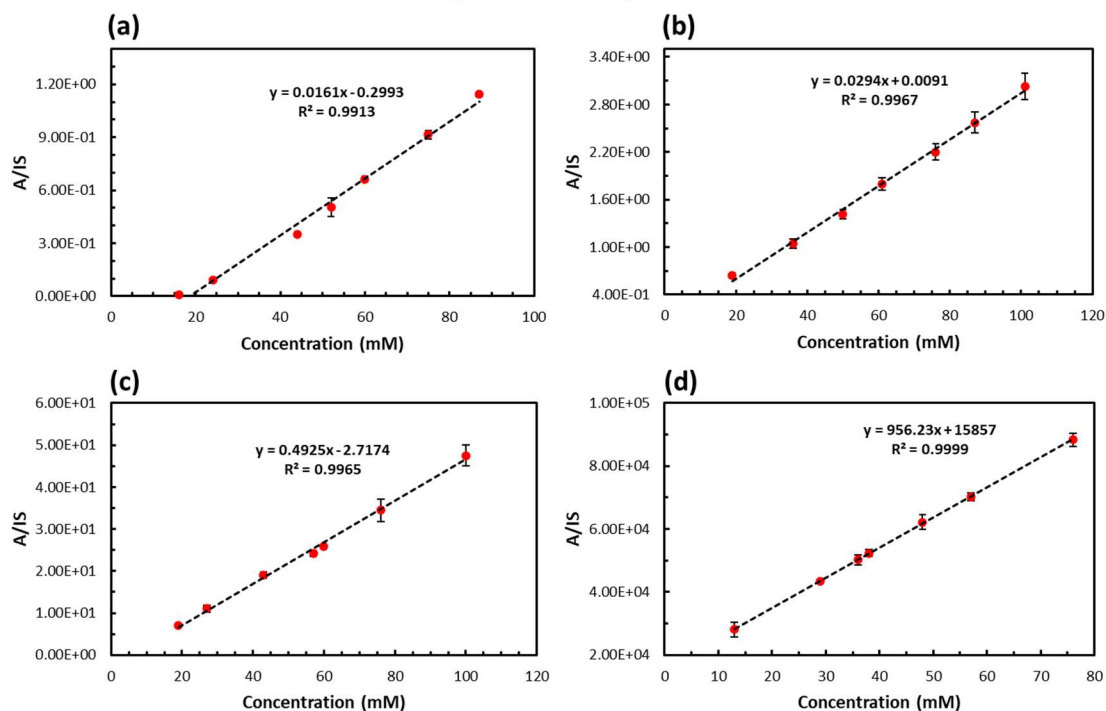


Figure S9. Calibration of analytes in standard neat water solutions. (a). Calibration of butyric acid standard solutions (16-87 mM), using the MS/MS transition m/z 89 \rightarrow 71. The internal standard used was succinic acid 2,2,4,4-d₄ with MS/MS transitions m/z 123 \rightarrow 95. (b). Calibration of glucose standard solutions (19-101 mM) using MS/MS transition m/z 215 \rightarrow 179. The internal standard used was D-Glucose-¹³C₆ with MS/MS transition m/z 221 \rightarrow 185. (c). Calibration of glycine standard solutions (19-100 mM) using MS/MS transition m/z 76 \rightarrow 48. The internal standard used was glycine-d₅ with MS/MS transition m/z 81 \rightarrow 53. (d). Calibration of histidine standard solutions (13-76 mM) using MS/MS transition m/z 156 \rightarrow 110. The internal standard used was glycine-d₅ with MS/MS transition m/z 81 \rightarrow 53.

C.3 Spiked Lab (SL) Models Test Accuracy

Spiked Lab (SL) Models – Test Accuracy

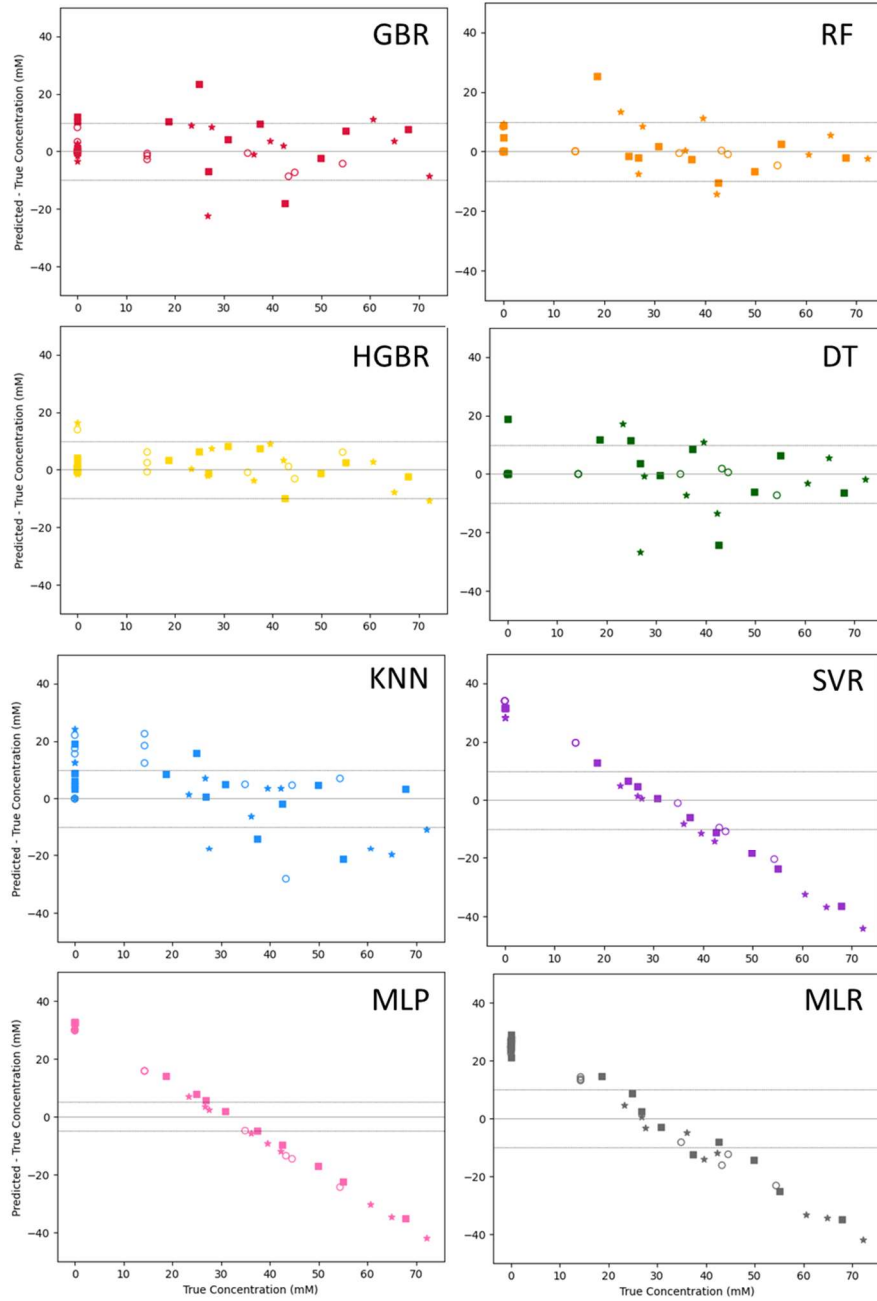


Figure S10. Model test accuracy for the spiked lab models. The test accuracy evaluates the internal accuracy of the model.

C.4 Spiked Lab (SL) Models – Unspiked Marine (UM) sample accuracy per model

Spiked Lab (SL) Models – Unspiked Marine UM Sample Accuracy

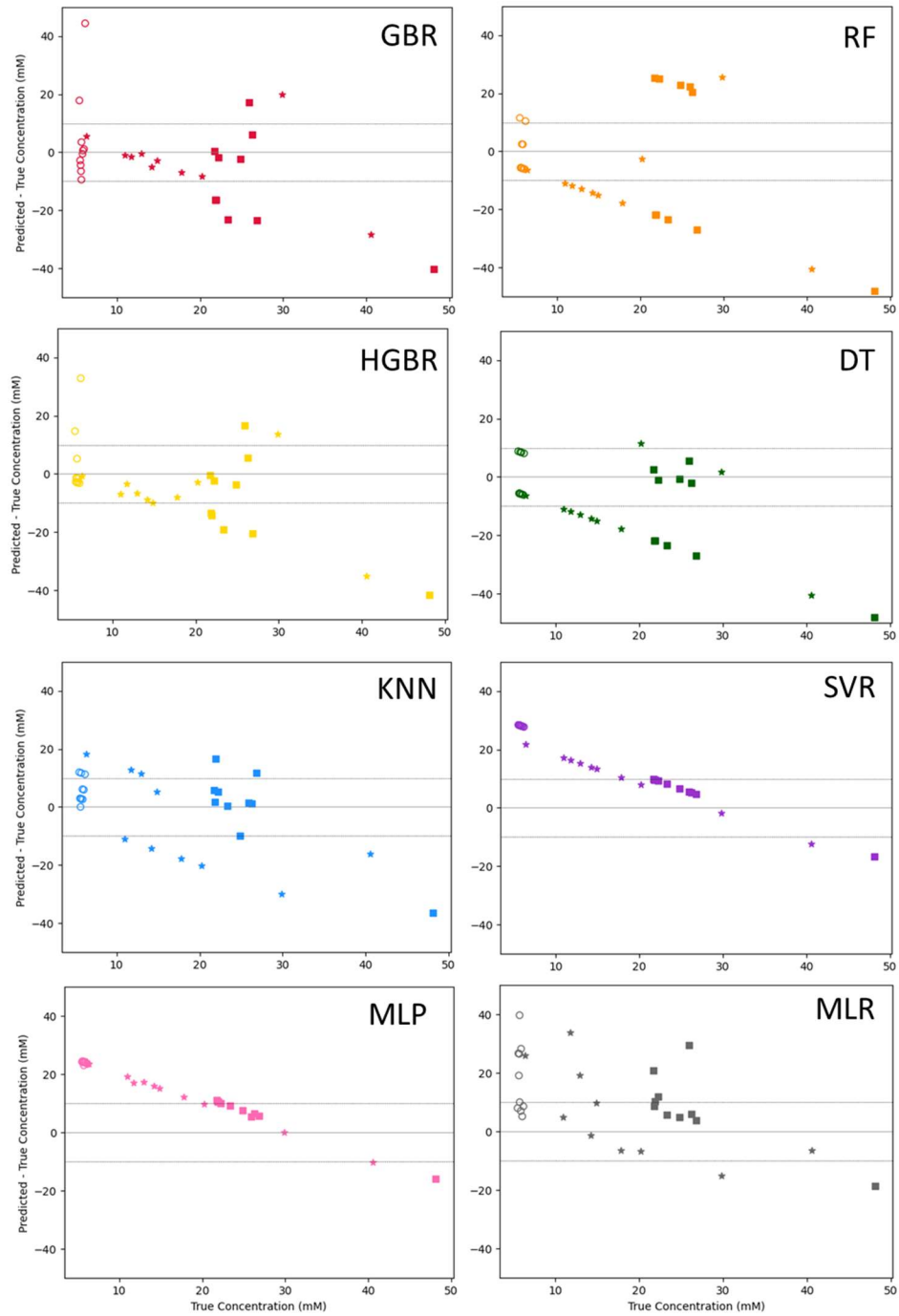


Figure S11. Field sample test accuracy for the spiked lab models.

C.5 Spiked Marine (SM) Models Test Accuracy

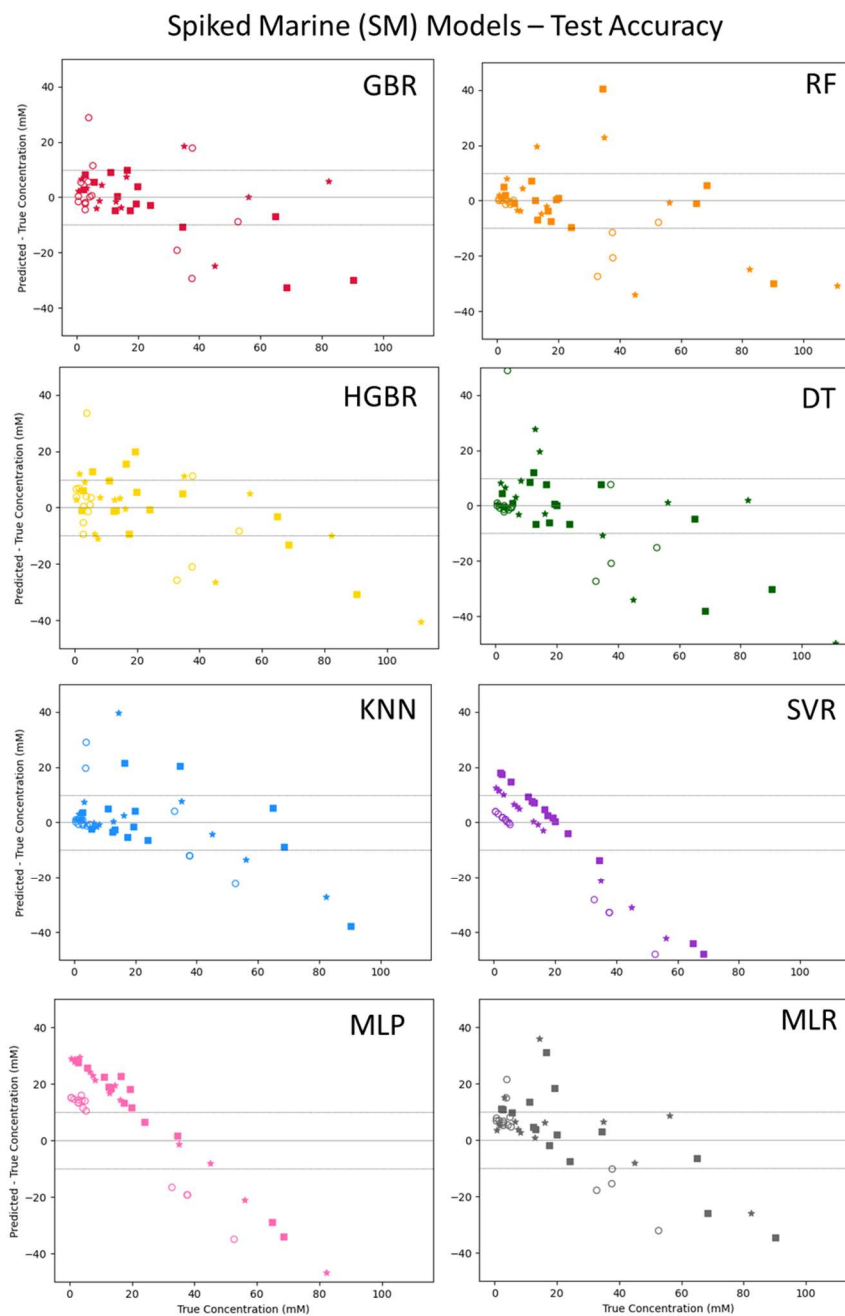


Figure S12. Model test accuracy for the spiked ocean models. The test accuracy evaluates the internal accuracy of the model.

C.6 Spiked Marine (SM) Models - Unspiked Marine (UM) sample accuracy per model

Spiked Marine (SM) Models - Unspiked Marine UM Sample Accuracy

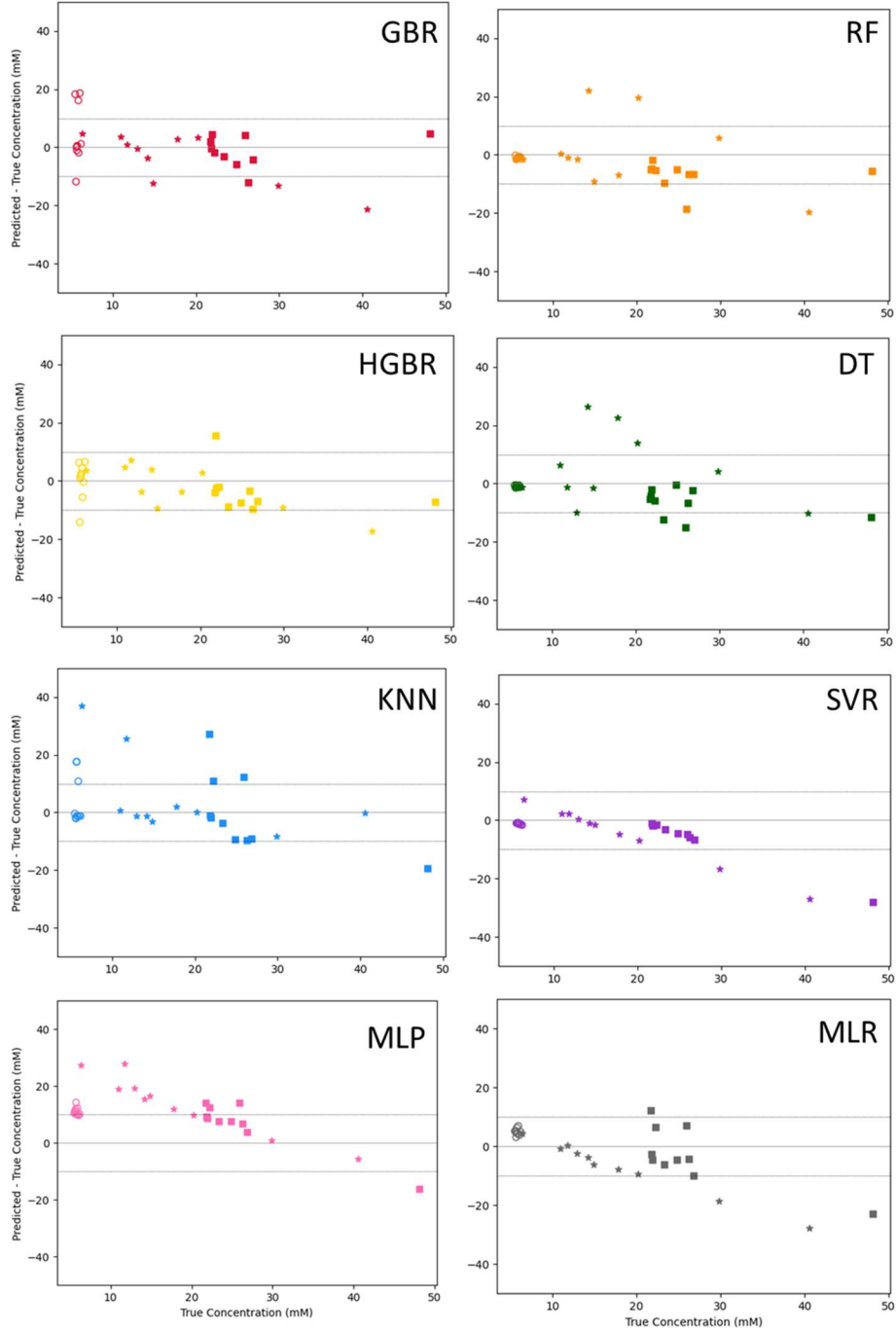


Figure S13. Field sample test accuracy for the spiked ocean models.

Appendix D. Supplemental Information for 532 nm Polarized Raman

D.1 Reading a Micrometer for Adjusting Slit Width

The micrometer that controls the entrance slit width of the IsoPlane spectrograph (see chapter 2.2.1 Raman Spectroscopy) is controlled by twisting the top knob clockwise to close and counterclockwise to open. Reading the slit width is done by first reading the number on the lowest part of the micrometer (**Figure S14 A**). The gradations here indicate millimeters (mm) for the integer values. Each of the ticks represent 0.25 mm. Next, the higher dial indicates smaller gradations within each 0.25 mm (**Figure S14 B**). The smallest ticks on this portion represent 0.010 mm and each full turn is a change of 0.25 mm. The value from A is added to B to determine the total slit width. Note, the micrometer and slit can be damaged if the micrometer is opened beyond 3 mm or closed beneath 0.010 mm.

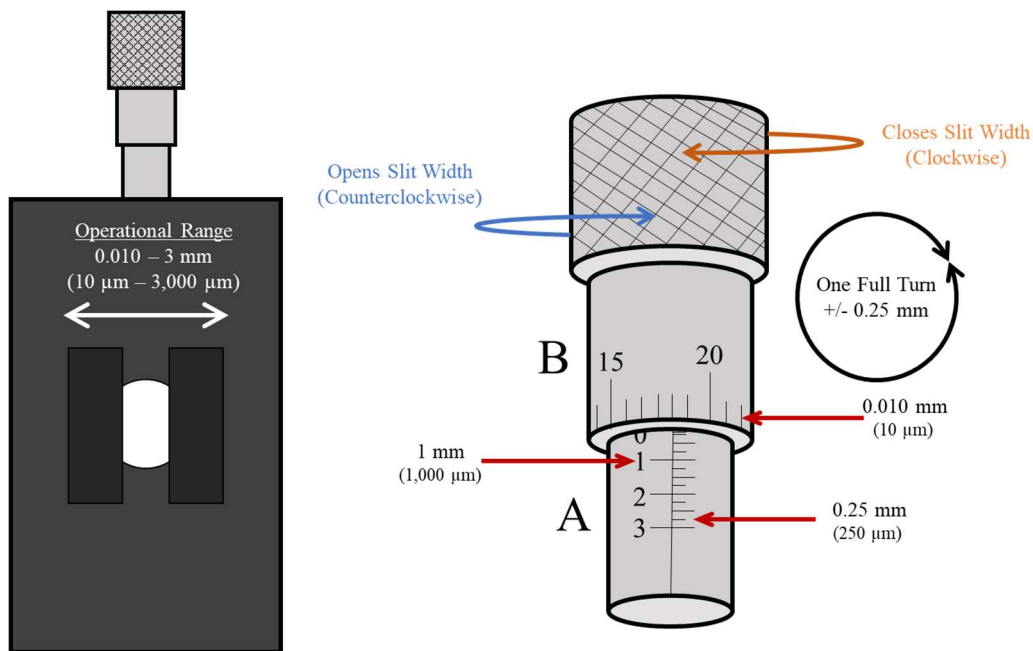


Figure S14. Diagram depicting the use and reading of the micrometer attached to the Princeton IsoPlane spectrograph.