

Unraveling the Complexity of the Ocean Surface Through Applied Computational  
Methods

Dissertation

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

By

Abigail Ann Enders

Graduate Program in Chemistry

The Ohio State University

2023

Dissertation Committee

Heather C. Allen, Advisor

Zachary Schultz

James Coe

Scott Elliott (LANL)

Copyrighted by  
Abigail Ann Enders  
2023

## Abstract

In totality, Earth's oceans comprise 70% of the planet's surface. The bulk water itself is teeming with inorganic ions, carbonaceous matter, and over a million *unknown* aquatic species. The surface, where all compounds must pass through when adsorbing from the atmosphere or releasing from sea spray, is chemically enriched and exerts control over several global processes, including micrometeorology, wave dampening, wave breaking, cloud condensation, ice nucleation, and sea spray aerosol generation. Like much of the ocean, the surface remains incompletely understood. Improved comprehension of the chemical and physical properties of the unique interfacial region is necessary for modern climate science.

The ocean's surface and interfacial region can be investigated in numerous ways to garner further details; each methodology provides insightful information and has drawbacks that must be acknowledged. A complete understanding of the surface requires a holistic scientific approach. In the first part of this work, gas-phase infrared spectra are utilized to train image-based machine learning (ML) models to identify functional groups. Using transfer learning, a small dataset (in respect to the large datasets necessary for complex neural networks) of spectra was sufficient to create models that correctly identified present and absent functional groups (>80% testing accuracy). Functional groups with weak infrared responses in crowded regions were the most difficult to identify. The

development of ML models that predict relevant spectral information improves upon the variability present when analysis is done manually. ML models make consistent predictions long after training, which gives them longevity. These models are examined further in the final work to evaluate their effectiveness for spectra of aqueous mixtures.

In the second study, methodology is developed to utilize Energy Exascale Earth Systems model output data that corresponds with relevant satellite data to map carbon on the ocean surface throughout an entire calendar year. In contrast to the first approach, which is more fundamental, the surface carbon is modeled to give global estimates of the total carbon mass and mapped to show global variability. At the sea surface *nanolayer* there is approximately  $10^{-4}$  gigatons of carbon globally. Interestingly, the *total* mass does not change month to month or seasonally, indicating a global equilibrium of carbon distribution. Yet, seasonality creates striking maps where winter months (for either hemisphere) have decreased surface carbon and summer brings the opposite effect. The dissolved organic carbon, an important chemical measurement in the field of study, is also modeled in addition to the carbon maps to provide reference values for future work.

Using established experimental methodology, the third approach presented is an investigation into the effect of temperature and bulk-phase on surface adsorption of the protein bovine serum albumin. Infrared reflection absorbance spectroscopy is employed to determine how the surface chemistry changes when lipids, proteins, temperature, and ionic strength are changed. Temperature has a less significant effect on bovine serum albumin adsorbing to the surface than ionic strength. Artificial sea water at 20°C has the most intense amide bands, which indicates that the solution conditions promote the most protein

adsorption compared to 0.45 M sodium chloride and water. The results are indicative of an ocean surface that is affected not only by cationic enrichment, but also by temperature, which has not been previously documented. The propensity for protein to adsorb to the surface is also debated in the literature; these results indicate that protein spontaneously adsorbs to the surface under all conditions and there is variability depending on the temperature, solution, and presence of a lipid.

The culmination of the first three studies is explored in the final study where mixture solutions are measured using attenuated total reflectance Fourier transform infrared spectroscopy and machine learning techniques are applied to identify analyte concentration. A matrix of sugar and protein solutions at varying concentrations was used to create a complex training dataset to create linear regression, principal component analysis, and support vector regression models. Using the support vector regression model, sugar concentrations of more complex samples, including additional sugars, were accurately predicted. Unknown ocean samples were tested on the same model and predicted sugar concentrations were in a reasonable range (<100 mM) compared to literature values of measured sugar concentration. These model results on a simple system provide an avenue for measuring the ocean surface chemistry more readily with less sample preparation required.

While ocean surface chemistry remains mostly undefined, the multifaceted approaches addressed by the work presented herein provide invaluable information towards understanding the vital interface. The machine learning development of the first work and capstone study expands the options for analyzing the astounding amount of data

collected from field missions and long-term studies. Because of the laboratory surface studies, it is known that the ocean's surface is affected by the chemical and physical properties of the region. Inevitably, the vastness of the world's oceans means that local adsorption properties are ever-changing from minute to minute and certainly season to season. Global models from calculated carbon mass based on literature values provide further input data for modeling, and we understand from the models that the surface of the ocean is variable globally, as is supported by the surface studies. The overarching cumulative effect of each study is an improved understanding of the ocean's dynamic surface and its impact on Earth's climate.

## **Dedication**

*To my friends who are like family, and family who are like friends.*

*And to my cats, Olive and Pepper, the reason for much of my laughter.*

## Acknowledgments

Dr. Heather Allen, my advisor and mentor, has been an invaluable source of knowledge and guidance beyond what I could have hoped for in pursuing my doctorate degree in Chemistry. She makes a difference for women in STEM. In addition, my wonderful Los Alamos National Laboratory Research Advisor, Dr. Scott Elliott, has guided me through challenging science and provided indispensable mentorship. My committee members have stimulated interesting avenues of scientific exploration and supported me at each milestone. Dr. Nicole Karn has provided me access to instrumentation that we do not have in the Allen Lab, allowing me to analyze ocean samples. Current Allen Group members Nicole North and Jessica Clark have provided scientific advice and been my daily sounding board through the entire process. Kathy Hayes kindly copy-edited, which polished this document.

Beyond science, I'd be remiss to not mention the people I've always had in my corner. To my family, Mom, Dad, Ma, Emily, Jordan, and kids: I'm thankful that everything we have been through has brought us together now more than ever. I share every bit of this journey with all of you. My parents have supported my wildest dream that has now become a reality. I know with every word, my grandfather, Joe, is with me. My friends, Marrison and Abby: your friendship and love has not gone unnoticed. My fiancé, Will, thank you for listening to every presentation and always being there for my sticky-note



idea scribbles, you're nothing short of amazing. I'm beyond grateful for your partnership; you've supported me in every aspect.

This dissertation was written in several trusty chairs, mostly listening to Taylor Swift, sometimes in 5-minute bouts and often hour-long stints. There is nothing so personal I have done in my life as much as writing this document. I am indebted to everyone who has helped me throughout my entire academic journey. It is what makes a uniquely and intensely independent process possible. Lastly, I must acknowledge what I've held close to me all these years, the eight-year-old me who wanted to become a scientist.

## Vita

B.S., St. Lawrence University.....2019  
Graduate Teaching Assistant, The Ohio State University.....2019-2021  
Graduate Research Intern, Los Alamos National Lab .....2021-2022  
Graduate Research Assistant, The Ohio State University.....2021-present

## Publications

- Samuel S. Tartakoff, Abigail A. Enders, Wenyao Zhang, Adam D. Hill. Spectroscopic and computational evidence for the concerted mechanism of the Wagner-Jauregg reaction, *Journal of Physical Organic Chemistry*, **2020**, 34 (2), e4140.
- Abigail A. Enders, Nicole M. North, Chase M. Fensore, Juan Velez-Alvarez, and Heather C. Allen. Functional group identification for FTIR spectra using image-based machine learning models, *Analytical Chemistry*, **2021**, 93 (28), 9711-9718.
- Abigail A. Enders, Scott M. Elliott, Heather C. Allen. Carbon on the ocean surface: Temporal and geographical investigation, *ACS Earth and Space Chemistry*, **2023**, 7 (2), 360–369.
- Abigail A. Enders, Jessica B. Clark, Scott M. Elliott, Heather C. Allen. New insights into cation-driven protein adsorption to the air-water interface through

infrared reflection studies of bovine serum albumin, *Langmuir*, **2023**, published April 7, 2023.

- Nicole M. North, Abigail A. Enders, Morgan Cable, Heather C. Allen. Array based machine learning for functional group detection in electron ionization mass spectrometry, *ACS Omega*, **2023**, resubmitted.
- Abigail A. Enders, Nicole M. North, Jessica B. Clark, Heather C. Allen. Saccharide concentration prediction from proxy-sea surface microlayer samples analyzed via ATR-FTIR spectroscopy and quantitative machine learning, **2023**, in submission.

#### Fields of Study

Major Field: Chemistry

## Table of Contents

Abstract.....	i
Dedication.....	v
Acknowledgments.....	vi
Vita.....	viii
List of Tables .....	xv
List of Figures.....	xvi
List of Equations.....	xxi
List of Common Abbreviations .....	xxiii
Chapter 1. Introduction.....	1
1.1. Motivation.....	1
1.2. Approach.....	2
1.3. Dissertation Highlights .....	3
Chapter 2. Theoretical Background, Instrumentation, and Computational Methods .....	5
2.1. Surface Tension Theory.....	5
2.2. Surface Pressure Theory .....	6
2.3. Air-Water Interface.....	7
2.3.1. Air-Sea Interface/Sea Surface Microlayer.....	7
2.3.2. Enrichment Factor, Surface Activity, Monolayer Formation .....	8
2.4. Infrared Spectroscopy .....	9
2.5. Absorption of Light.....	12
2.6. Fourier Transform Infrared Spectroscopy (FTIR).....	12
2.6.1. Attenuated Total Reflectance.....	16
2.6.2. Infrared Reflection Absorbance Spectroscopy .....	17
2.7. Computational Method Theory.....	18
2.7.1. Linear Regression .....	19
2.7.2. Principal Component Analysis .....	19
2.7.3. Support Vector Regression .....	21
2.7.4. Convolutional Neural Networks .....	22

Chapter 3. Functional Group Identification for FTIR Spectra Using Image-based Machine Learning Models .....	25
3.1. Introduction .....	25
3.2. Methods.....	30
3.2.1. Python Scripts .....	30
3.2.2. Spectra Collection.....	30
3.2.3. Data Pre-Processing .....	30
3.2.4. Labeling .....	31
3.2.5. Machine Learning .....	31
3.2.6. Accuracy and Loss .....	33
3.2.7. Classification of Validation Data.....	33
3.3. Results and Discussion .....	34
3.4. Conclusion .....	38
Chapter 4. Carbon on the Ocean Surface: Temporal and Geographical Investigation.....	47
4.1. Introduction.....	47
4.2. Methods.....	50
4.2.1. Proteins .....	51
4.2.2. Lipids .....	52
4.2.3. Carbohydrates .....	52
4.2.4. Chlorophyll Data and Plankton Concentration.....	53
4.2.5. Carbon Calculations.....	53
4.3. Results and Discussion .....	56
4.4. Conclusions.....	61
Chapter 5. New insights into cation and temperature driven protein adsorption to the air-water interface through infrared reflection studies of bovine serum albumin .....	71
5.1. Introduction.....	71
5.2. Methods.....	76
5.2.1. Materials and Sample Preparation .....	76
5.2.2. Infrared Spectroscopy .....	77
5.3. Results and Discussion .....	78
5.3.1. Solution Effect .....	79
5.3.2. Temperature Effect .....	80
5.3.3. Monolayer Effect .....	82

5.3.4. Application to Climate Models.....	85
5.4. Conclusions.....	86
Chapter 6. Saccharide Concentration Prediction from Proxy-Sea Surface Microlayer Samples Analyzed via ATR-FTIR Spectroscopy and Quantitative Machine Learning ...	96
6.1. Introduction.....	96
6.2. Methods.....	100
6.2.1. Training Solution Preparation, Data Collection, and Data Preprocessing....	100
6.2.2. Proxy-Sample and Real Sea Surface Water Preparation and Sampling .....	101
6.2.3. Field Sampling.....	102
6.3. Results and Discussion .....	104
6.4. Conclusions.....	111
Chapter 7. Conclusions .....	124
References.....	127
Appendix A. Computational Methods Utilized in Chapter 3 for FTIR Functional Group Analysis.....	148
A.1. Obtaining Spectra.....	148
A.2. Spectral Processing .....	148
A.2.1 Creating Directories on Computer .....	148
A.2.2. Removing Spectra not in Absorbance or Wavenumbers .....	149
A.2.3. Convert from ‘jcamp-dx’ to ‘csv’ .....	149
A.2.4. Move ‘csv’ Files.....	149
A.2.5. Normalize ‘csv’ Files.....	149
A.2.6. Convert ‘csv’ to ‘jpg’ and move Spectra Images.....	149
A.2.7. Copy Spectra Images to Functional Group Directories .....	150
A.2.7. Separation of Test Images and Setting Equivalent Examples per Class.....	150
A.3. Model Training .....	150
A.4. Analysis of Models .....	151
A.4.1. Classify.....	151
A.4.2. Pearson’s Correlation Coefficient.....	151
A.4.3. Plotting.....	151
Appendix B. Inception V3 Architecture and Optimization Functions Utilized in Chapter 3 for FTIR Analysis .....	152
Appendix C. Global carbon maps and additional figures from Chapter 4.....	155

C.1. Maps of Surface Concentrations for Proteins and Lipids, Fractional Surface Coverage, and Non-normalized SSnL Carbon.....	155
C.2. Longhurst Region Carbon Results .....	163
C.3. Normalized SSnL Carbon Maps for the Year of 2005.....	165
Appendix D. Attenuated Total Reflectance FTIR Absorbance Analysis and Details Regarding Pathlength Variability Discussed in Chapter 5 .....	166
Appendix E. Bovine Serum Albumin Concentration Dependent IRRAS from Chapter 5 .....	170
Appendix F. Explanation of Machine Learning Specifications, Plots of all Matrix Samples, Plots of Real Field Samples, and Optimization Results for SVR used in Chapter 6.....	178
Appendix G. Chemometric Investigation via Factor Analysis of Phosphate Raman Spectra to Elucidate Phosphate Monomer and Oligomer Spectral Components.....	188
Appendix H. General Python Codes and GitHub Resources.....	192
H.1. Support Vector Regression .....	192
H.2. Work-Up of IRRAS Data.....	194
H.3. Calculating SSnL Carbon Using Chlorophyll and Zooplankton data from E3SM Model .....	197
H.4. Principal Component Analysis.....	200
H.5. Support Vector Machine .....	201
H.6. FTIR Spectrum Calculations Based on Angle of Incident Light.....	202
H.7. Linear Regression Model .....	204
H.8. Subprocess Script for Preprocessing NIST FTIR Spectra .....	206
H.9. Check that Spectrum is in Units of Absorbance .....	208
H.10. Copy a Given File to a New Directory .....	210
H.11. Convert File From ‘csv’ to ‘jpeg’ Format.....	211
H.12. Convert Specified File From ‘jcampdx’ to ‘csv’ File.....	212
H.13. Make Directory Given Keywords for Naming .....	213
H.14. Move File to Different Directory .....	213
H.15. Normalize Data .....	214
H.16. Plot Confusion Matrix.....	215
H.17. Remove Random Files Until Directories are Equal in Files .....	217
H.18. Train Model to Predict Functional Group Subprocess Code .....	218
H.19. Classify Unknown Images in Batches .....	219

H.20. GitHub Resources .....	221
Appendix I. Sea Surface and Bulk Sampling of Atlantic Ocean and Banana River in Florida in January 2023 .....	222
I.1. Precleaning the Glass Sample Vessels .....	222
I.2. SSML/ Surface Sampling .....	222
I.3. Sea Foam Sampling .....	222
I.4. Bulk Ocean Sampling .....	222
Appendix J. Permissions .....	227



## List of Tables

Table 1. Functional groups for which successful models were trained. ....	39
Table 2. Functional groups are presented with the total number of positive spectra examples used in training as well as the total number of functional group examples used. ....	40
Table 3. Final accuracy and cross entropy for train and validation of each functional group model is presented in order of increasing number of training images.....	41
Table 4. Aldehyde and carboxylic acid IR stretching and bending mode frequencies. ....	42
Table 5. Summary of variables used in calculation of total surface carbon mass including relevant references for literature values. ....	65
Table 6. Three different ionic strengths were used to evaluate the adsorption of BSA to the surface with and without a competing stearic acid monolayer at both 10° and 20°C. Experiments outlined here are for IRRAS measurements. ....	89
Table 7. Predicted sugar concentration (M) in more complex samples containing glucose, sucrose, ESA, bovine serum albumin (BSA), and 1-butanol are predicted by the SVR and LR model. Values are the average predicted concentration (M). The SVR model predicts reasonable concentration values in the range of the true concentration, while the LR model predictions do not provide any reasonable estimates of concentration.....	117
Table 8. R <sup>2</sup> and mean squared error of linear regression (LR) and support vector regression (SVR) models after training. ....	118
Table 9. Functional group analysis of proxy sample and unknown SSML sample. Red text indicates that the model incorrectly predicted (e.g., nitrile is predicted present for the proxy sample, yet it is not present). An asterisk (for Banana River sample only) indicates that the GC-MS of the sample has characteristic m/z values for that functional group identification. ....	122
Table 10. Sensitivity, specificity, positive predictive value, and negative predictive value for model results on proxy sample prediction of functional groups. These metrics provide a more thorough analysis of how the model performs and detail the model’s performance more holistically.....	123
Table 11. Longhurst regional codes and the sum total carbon in the region for the months of May and November 2005. ....	163
Table 12. Ion/element concentrations for Instant Ocean from manufacturer. ....	170
Table 13. Glucose concentration (M) of training samples.....	179
Table 14. Egg serum albumin concentration (mg/mL) for training data solutions.....	179

## List of Figures

Figure 1. IRRAS configuration used in the Allen lab. $\theta$ is angle relative to surface normal that the incident beam (green) is directed onto the surface (blue). Reflected light (yellow) is directed to the detector.....	18
Figure 2. Principal component analysis general schema. ....	21
Figure 3. Nodes $a^l$ and $a^{l'}$ are connected to one another. The variable $l$ denotes location regarding each node. ....	23
Figure 4. Number of spectra used to train each functional group model, (a) carbon-containing, (b) nitrogen-containing, (c) halide-containing, and (d) oxygen-containing. The number of images is equivalent for the positive and negative cases used in training and testing. ....	43
Figure 5. General summary of Inception V3 architecture. Additional details are provided in the Supporting Information, including a summary of the model preprocessing parameters. ....	44
Figure 6. Confusion matrix is identical for carboxylic acid, aromatic, methyl, and ether functional group models. ....	45
Figure 7. Final train and validation accuracy and cross entropy as a function of the number of spectra used to train each functional group. Insets (a) and (b) show validation results; (c) and (d) show training results. Pearson’s correlation coefficients (PCC) are inset in the plots for final accuracy and cross entropy of training and validation as a function of the number of spectra. The coefficients closer to $\pm 1$ indicate that the train accuracy and cross entropy are linearly correlated (negative is inversely correlated and positive is directly correlated) to the number of spectra used in training. Validation accuracy and cross entropy are not linearly correlated to the spectral examples used. ....	46
Figure 8. Simplified schematics illustrating (a) relationships between the marine boundary layer (white), sea surface nanolayer (light blue), and sea surface microlayer (dark blue) and (b) highlighting some of the major oceanic processes that occur including vertical transport from the bulk, enrichment of organics at the surface nanolayer, adsorption of atmospheric aerosols and gases, and release of sea spray aerosols from the ocean to the atmosphere.....	64
Figure 9. Comparison of four field study DOC concentrations <sup>109,149–151</sup> and the average mass of carbon each month in the region of 0-10 east longitudes and 78-80 north latitudes, which aligns with the regions studied in the Rossel et al., 2020 study. ....	66
Figure 10. Maps of normalized SSnL carbon for the months of March (3), June (6), September (9), and December (12) from E3SM output for the year 2005 are presented. ....	67
Figure 11. Normalized SSnL carbon for May 2005 (top, ‘5’) and November 2005 (bottom, ‘11’) calculated from E3SM chlorophyll-a and zooplankton output. ....	68

Figure 12. Normalized SSnL carbon across (a) 0° and (b) 25 north latitudes for May and November 2021. Only locations where estimates are greater than zero are included (plots exclude land). Less seasonal variability is observed at the equator in (a), and coastal regions in (b) are well emphasized by the uptick in calculated surface carbon. As we approach land, carbon increases and then decreases in more open ocean regions. .... 69

Figure 13. Subset of Longhurst regions with SSnL carbon mass (g) between May and November 2021. Regions are ordered alphabetically by their four-letter standard codes. All values are presented in Appendix C and a subset is discussed in text to underscore key observations. Province acronyms are defined as follows: Northwest Arabian Sea Upwelling (ARAB), Archipelagic Deep Basins (ARCH), East Australian Coastal (AUSE), Western Australian and Indonesian Coast (AUSW), California Current (CCAL), Chile Current Coastal (CHIL), China Sea Coastal (CHIN), South Subtropical Convergence (SSTC), Sunda-Arafura Shelves (SUND), Tasman Sea (TASM), Western Pacific Warm Pool (WARM), and Western Tropical Atlantic (WTRA)..... 70

Figure 14. Experimental design of IRRAS assembly combined with a Langmuir trough and temperature control via a recirculating chiller. .... 88

Figure 15. FTIR of BSA in water at increasing concentrations measuring solution-phase concentrations starting from 1 μM to 1000 μM shown in the amide region, presented with error of one standard deviation. The value of the standard deviation is so small it is approximately the thickness of the line of each spectrum. The sharp features likely result from gas phase water present in ambient conditions. .... 90

Figure 16a-c. Surface-IR responses in amide region (1800-1450 cm<sup>-1</sup>) after bovine serum albumin (BSA) injection into the H<sub>2</sub>O, 0.45 M NaCl, and the artificial seawater solutions at **a) 20°C, b) 10°C, c) difference spectra of 20°C – 10°C**. Here, reflectance-absorbance bands are observed as negative peaks. .... 91

Figure 17. Integrated peak area for amide-I (1660 cm<sup>-1</sup>) and amide-II (1540 cm<sup>-1</sup>) bands at 10°C (dark, solid) and 20°C (light, diagonal lines). As noted, the peak area and intensity is a result of surface adsorbed protein as confirmed through bulk measurements via ATR-FTIR..... 92

Figure 18a-b. IR response in amide region (1800-1500 cm<sup>-1</sup>) after bovine serum albumin (BSA) injection into each solution with a stearic acid monolayer (~45 Å<sup>2</sup>/molecule) on the surface at a) 20°C and b) 10°C. .... 93

Figure 19. Spectra of ultrapure water surface after injection of bovine serum albumin (BSA) at variable temperature with and without stearic acid monolayer to emphasize the minimal intensity change in the water system. Spectra are offset horizontally for clarity. .... 94

Figure 20. Observed maximum intensity of amide I band (1660 cm<sup>-1</sup>) for each solution at varying temperatures. IR responses from each solution with stearic acid monolayer at 10°C are not included due to the variability in the amide I and II region. .... 95

Figure 21. Schematic flow chart of data collection process to the ML pipeline. .... 113

Figure 22. ATR-FTIR spectrum of 0.6 M glucose and 2 mg/mL egg serum albumin. The labels are provided to emphasize that the components do not compound on one another and are well resolved, despite being in a similar wavenumber region. .... 114

Figure 23. Principal components (PCs) one and two from the data dimensionality reduction performed using principal component analysis (PCA). The relative ratio is respective to glucose. Solutions with a relative ratio of ‘1’ have no ESA. PC1 mainly captures glucose response and PC2 mainly captures ESA response. ....	115
Figure 24. Linear regression (LR) (a) fits the experimental training data well with a 100 % R <sup>2</sup> and no mean squared error. ‘True’ indicates the known concentration of saccharide while ‘predicted’ is the model’s estimate. Proxy sample saccharide concentrations are not correctly predicted, as shown with the teal ‘X’ demarcating the known saccharide concentration. Support vector regression (SVR) (b) results show that the test data accurately follows the training data. Predicted concentrations for the known complex samples are much closer to the true concentration. The training results in an R <sup>2</sup> of 97.1%. ....	116
Figure 25. Support vector regression (SVR) model predictions on unknown field samples are closely aligned with the training and test data although are in the low absorbance range.....	119
Figure 26. Feature extraction for SVR model. Positive values indicate strong influence on model prediction and negative values indicate negative impacts on model success. ....	120
Figure 27. Feature extraction for SVR model with reduced features. The model uses similar features for prediction and the negatively impacting features have been removed. ....	121
Figure 28. Overview of preprocessing steps included in model training. Dimensions of the output jpeg image file after each step. ....	152
Figure 29. Summary of convolutional layer; gamma and beta are weights for the neuron nodes. Additionally, circles denote constant parameters (may be adjusted before or after but remain constant during the step with the arrow pointed at it.) ....	152
Figure 30. Modeled SSML lipid concentrations (μM) for May 2005. ....	155
Figure 31. Modeled SSML lipid concentrations (μM) for November 2005. ....	156
Figure 32. Modeled SSML protein concentrations (x 10 μM) for May 2005. ....	157
Figure 33. Modeled SSML protein concentrations (x 10 μM) for November 2005. ....	158
Figure 34. Modeled fractional surface coverage for May 2005.....	159
Figure 35. Modeled fractional surface coverage for November 2005.....	160
Figure 36. Modeled SSnL carbon mass (x10 <sup>7</sup> g) for May 2005. ....	161
Figure 37. Modeled SSnL carbon mass (x10 <sup>7</sup> g) for November 2005.....	162
Figure 38. Modeled SSnL carbon normalized to the highest observed mass over all months for January (‘1’) through December (‘12’) in 2005 calculated from E3SM output. ....	165
Figure 39. Maximum observed absorbance in amide I region (1653 cm <sup>-1</sup> ) as a function of concentration. The dashed black line is a linear fit with an R <sup>2</sup> value of 0.99 (inset). ....	166
Figure 40. Integrated peak area for amide region is linear with increasing concentration. ....	167
Figure 41. O-H stretching region for water, 0.45 M NaCl, 1 M NaCl, and artificial sea water (ASW). Standard deviation is shown and it is approximately the thickness of the line of the peak.....	169

Figure 42. O-H bend of each solution. Standard deviation is shown and is approximately the thickness of the line.....	169
Figure 43. Maximum observed absorbance in amide I region (1653 cm <sup>-1</sup> ) as a function of concentration. The dashed black line is a linear fit with an R <sup>2</sup> value of 0.99 (inset). ....	172
Figure 44. Integrated peak area for amide region is linear with increasing concentration. ....	173
Figure 45. Background-corrected IRRAS showing O-H stretching region changes at variable BSA concentrations (given in μM). Injections of 1 and 50 μM solutions do not have a significant IR response as evidenced by the low intensity. ....	175
Figure 46. Maximum absolute value of reflectance-absorbance for O-H stretch at 3585 cm <sup>-1</sup> . While data for 1 and 50 μM are presented, the conclusions that can be drawn from the observed IR response are limited because of the limit of detection.....	175
Figure 47. Background-corrected IRRAS showing amide I region changes at variable BSA concentrations (given in μM). Injections of 1 and 50 μM solutions do not have a significant IR response as evidenced by the low intensity. ....	176
Figure 48. Maximum absolute value of reflectance-absorbance for amide I (νC=O) at 1640 cm <sup>-1</sup> . While data for 1 and 50 μM are presented, the conclusions that can be drawn from the observed IR response are limited because of the limit of detection. ....	176
Figure 49. Standard deviation of RA for each solution at minimum peak intensity for amide I mode.....	177
Figure 50. Composite spectra of all 100 samples used for training in each machine learning model. ....	180
Figure 51. Average spectra of real ocean samples from Cocoa Beach, Florida. Standard deviation is shown but is approximately the thickness of the line.....	181
Figure 52. Average spectra of ocean and river samples from Cocoa Beach, Florida for comparison of sampling sites. Standard deviation is shown but is approximately the thickness of the line. ....	182
Figure 53. MS of GC retention for January 11, 2023, ocean surface sample from Cocoa Beach, Florida. ....	183
Figure 54. MS of GC retention from January 11, 2023, river surface sample from the Banana River in Cocoa Beach, Florida. ....	184
Figure 55. MS of bulk surface water sample from Banana River in Cocoa Beach, Florida on January 10, 2023. ....	185
Figure 56. Optimization of regularization parameter C for the support vector regression (SVR). Variability shown is that of changing ε, the tolerance limit, which varies little compared to the optimization of C.....	186
Figure 57. Optimization of train-test size split for the SVR. Minimization of MSE is prioritized for model performance. An 80/20 split minimizes MSE and has literature precedence.....	187
Figure 58. Select range of concentrations of phosphate ion Raman spectra. ....	190
Figure 59. Resultant factor spectra after dimensionality reduction via factor analysis..	190
Figure 60. Reconstructed spectrum for 1M phosphate from factors and original 1M spectrum. The factors are reasonably similar. ....	191

Figure 61. Photo of Abbie rinsing glass vessel in accordance with protocol for collecting bulk samples..... 223

Figure 62. Picture of Abbie (light blue) with glass slide on Banana River, assisted by Nicole (gray long sleeve) holds the kayak steady and Jess (gray short sleeve) operates the squeegee and glass storage vessel..... 224

Figure 63. Jess squeegees the glass slide, held by Abbie, after it was dipped in the Banana River during surface sampling..... 225

Figure 64. Nicole (left) and Jess (right) collect sea foam/surface in Atlantic ocean by placing slide on surface squeegeeing off the water into glass storage vessel..... 226

## List of Equations

Equation 1 .....	5
Equation 2 .....	6
Equation 3 .....	6
Equation 4 .....	7
Equation 5 .....	9
Equation 6 .....	10
Equation 7 .....	10
Equation 8 .....	10
Equation 9 .....	10
Equation 10 .....	11
Equation 11 .....	11
Equation 12 .....	11
Equation 13 .....	11
Equation 14 .....	12
Equation 15 .....	12
Equation 16 .....	13
Equation 17 .....	13
Equation 18 .....	14
Equation 19 .....	14
Equation 20 .....	15
Equation 21 .....	15
Equation 22 .....	15
Equation 23 .....	15
Equation 24 .....	15
Equation 25 .....	16
Equation 26 .....	16
Equation 27 .....	16
Equation 28 .....	17
Equation 29 .....	17
Equation 30 .....	19
Equation 31 .....	20
Equation 32 .....	20
Equation 33 .....	20
Equation 34 .....	21
Equation 35 .....	22
Equation 36 .....	22

Equation 37 .....	23
Equation 38 .....	23
Equation 39 .....	23
Equation 40 .....	24
Equation 41 .....	24
Equation 42 .....	24
Equation 43 .....	24
Equation 44 .....	54
Equation 45 .....	55
Equation 46 .....	56
Equation 47 .....	153
Equation 48 .....	153
Equation 49 .....	153
Equation 50 .....	153
Equation 51 .....	154
Equation 52 .....	154



## List of Common Abbreviations

<b>ASW</b>	Artificial Sea Water
<b>ATR</b>	Attenuated Total Reflectance
<b>CCN</b>	Cloud Condensation Nuclei
<b>CNN</b>	Convolutional Neural Network
<b>DOC</b>	Dissolved Organic Carbon
<b>E3SM</b>	Energy Exascale Earth Systems Model
<b>FTIR</b>	Fourier Transform Infrared Spectroscopy
<b>INP</b>	Ice Nucleating Particle
<b>IRRAS</b>	Infrared Reflection Absorbance Spectroscopy
<b>LR</b>	Linear Regression
<b>ML</b>	Machine Learning
<b>PCA</b>	Principal Component Analysis
<b>SSA</b>	Sea Spray Aerosol
<b>SSML</b>	Sea Surface Microlayer
<b>SSnL</b>	Sea Surface Nanolayer
<b>SVR</b>	Support Vector Regression

## Chapter 1. Introduction

### 1.1. Motivation

The research presented in this dissertation was conducted to explore ocean surface chemistry through computation methods and provide insight into the complexity of the ocean. The work is motivated by both the complexity of current monitoring methods and the need for more consistent sampling or predictions of the sea surface chemistry.

The sea surface microlayer (**SSML**) is enriched with organics that form a monolayer at the interface. Organic monolayers likely impact micrometeorological phenomena because monolayers are known to dampen wave formation. Enriched organics are also transported to the atmosphere as sea spray aerosols (**SSAs**) through wave breaking or bubble bursting. SSAs are a mechanism for release of ice nucleating particles and cloud condensation nuclei, both of which affect climate and atmosphere.

Machine learning (**ML**) provides an avenue to understand chemical space beyond what is achievable with human analysis of spectroscopic data. Through ML, a higher throughput of data is achievable, while also enabling exploration of connections between data that previously could not be studied. Utilizing existing ML methods, the work in this dissertation emphasizes the improved analytical opportunities for understanding the SSML.

## 1.2. Approach

Image-based convolutional neural networks (**CNN**) were utilized to establish a consistent method for analyzing Fourier Transform Infrared (**FTIR**) spectra to determine the functional groups in a spectrum. The qualitative identification achieved with ML enabled exploration of more complex FTIR spectra that were created as proxies for SSML samples. Proxy samples were used to evaluate the qualitative ML models built on thousands of spectra and quantitative ML methods were introduced to achieve accurate concentration of carbohydrates in new samples.

The complexity and intricacy of studying the SSML require a holistic approach to consistent monitoring of organics. Satellites and models, such as NASA's MODIS and Energy Exascale Earth Systems Model (**E3SM**), respectively, provide global data on chlorophyll-a (**chl-a**). Chl-a is utilized herein to model carbon at the ocean surface using phytoplankton, the ocean's primary producer, because the detritus from these microorganisms becomes dissolved organic carbon (**DOC**) and generates the organic monolayer or sea surface nanolayer (**SSnL**).

Laboratory experiments, including attenuated total reflectance (**ATR**) and infrared reflection-absorbance spectroscopy (**IRRAS**) are used to provide necessary insight into complex SSML samples. IRRAS provides surface-sensitive analysis of the behavior of molecules at the air-water interface and enables greater understanding of the complex chemical and physical properties in a controlled experiment. Utilizing spectroscopy to determine composition ultimately reduces the organic waste created during extraction; samples are analyzed directly.

The combination of quantitative ML, qualitative ML, and global modeling provides an advanced avenue for monitoring the ocean's surface. In conjunction with necessary laboratory experiments, such as IRRAS of the air-water interface, the fundamental and complex processes of the ocean's surface can be unraveled. With ML, the observation and quantification of SSML organics can be more frequent and more informative.

### 1.3. Dissertation Highlights

Chapter 2 includes background details on the specific methods, instrumentation, computational, and ML techniques utilized throughout the dissertation work. Included in this section is a brief theory of gas-liquid anisotropy, FTIR, ATR-FTIR, and IRRAS. Principal component analysis (**PCA**), linear regression (**LR**), support vector regression (**SVR**), and CNN mathematical background is detailed along with the computational background for modeling the SSnL.

Chapter 3 reports on the work to develop image-based ML models to predict functional groups from FTIR spectra. It was found that successful CNN models were made for 15 common organic functional groups and high predictive accuracy was achieved on unknown spectra. Since each functional group had a unique model, the dependence on training spectra was reduced and many functional groups with limited training data were able to perform with similar predictive power as groups that have thousands of examples.

Chapter 4 approaches the question of modeling the SSnL carbon from a computational direction by using E3SM model output to map ocean surface coverage of carbon. The calculated amount of DOC by the new model is compared to literature values of DOC from field studies and is comparable to the ranges. Global maps of carbon in the

SSnL and the total value of carbon over a month reveal that while there is regional variability, the total amount of carbon does not significantly vary.

Chapter 5 summarizes the laboratory experiments directed at understanding the dynamics of protein surface adsorption to the air-water interface. Using IRRAS different subphases of pure water, NaCl, and artificial sea water at variable temperatures were used to investigate the proclivity of bovine serum albumin to partition to the surface. Amide bands of greater intensity were observed for bovine serum albumin in artificial sea water at higher temperatures, indicative of the protein being salted out.

Chapter 6 contains the culminating work of this dissertation. Real SSML samples are examined via ATR-FTIR spectroscopy and pre-trained ML models to achieve quantitative prediction of sugar concentration. The most effective ML model is determined by examining complex samples after training with proxy-solutions of varying protein and sugar concentrations. SVR predicts the correct sugar concentration within tens of millimolar, providing a methodology that quickly and efficiently enables analysis of the SSML for improved climatological and atmospheric modeling.

Lastly, Chapter 7 summarizes the findings presented in this dissertation. The atmospheric and climatological implications are briefly discussed to contextualize the work completed for this dissertation. Ultimately, it is posited that the findings from each chapter are unique and provide an avenue for sustainable environmental chemistry research of the ocean.

## Chapter 2. Theoretical Background, Instrumentation, and Computational Methods

An overview of the theory, instrumentation, and computational methods used throughout this dissertation is detailed herein. Relevant background is provided for the SSML, FTIR, ATR-FTIR, IRRAS, PCA, LR, SVR, and CNNs.

### 2.1. Surface Tension Theory

Surface tension, formally described as force per distance, or Newton per meter, results from the excess free energy at the interface between two bulk phases.<sup>1,2</sup> Gas, liquid, and solid matter can form five distinct interfaces: gas-liquid, gas-solid, liquid-liquid, liquid-solid, and solid-solid. The free energy of formation must be positive ( $G > 0$ ), where free energy is equal to or less than zero, the surface region of the matters would expand infinitely and form a disperse, indistinguishable material. While the derivations and detailed equations are documented in detail elsewhere, a brief summary of the theory is provided.<sup>1-3</sup>

The change in internal energy,  $U$ , in a two-phase system is described by

$$dU = TdS - PdV + \sum \mu_i dN_i + \gamma dA$$

Equation 1

where  $T$  is temperature,  $S$  is entropy,  $P$  is pressure,  $V$  is volume,  $\mu$  is chemical potential,  $N$  is number of molecules,  $\gamma$  is force, and  $A$  is area.

Given that Gibbs free energy is defined as

$$G = U + PV - TS$$

Equation 2

and assuming a system at constant temperature, volume, and moles, the force,  $\gamma$ , or surface tension, is expressed as

$$\gamma_0 = \left( \frac{dG}{dA} \right)_{T,V,N}$$

Equation 3

Under the assumed conditions, a decrease in area ( $-\delta A$ ) will decrease  $G$ , if  $\gamma$  is positive. Stable liquid phases generally see a decrease in surface area, attractive forces (e.g., dipole-dipole, London dispersion, Hydrogen bonding, and induced-dipole) of phases decrease the distance between individual molecules and thus the overall area decreases, such that surface tension is always positive. Sometimes in the literature, Gibbs ( $G$ ) and Helmholtz ( $F$ ) free energy are interchanged in expression for surface tension; the small changes in pressure and volume make the values essentially equivalent.

## 2.2. Surface Pressure Theory

The surface tension of pure water,  $\gamma_0$ , is 72 mN/m under standard conditions.<sup>4</sup> It is made so high by the hydrogen bonding network. These intermolecular forces are disrupted upon the addition of a surfactant and formation of a monolayer. The disruption to the surface tension,  $\gamma$ , from the monolayer is often recorded as surface pressure,  $\Pi$ .

$$\Pi = \gamma_0 - \gamma$$

Equation 4

### 2.3. Air-Water Interface

Unlike the isotropic bulk of water, the orientational anisotropy of the interface creates a unique region.<sup>5,6</sup> Bulk water, for example, has an ideal organization in the bulk that is disrupted by the presence of a surface, for example. The interface itself is not a definitive thickness or infinitesimally small region; it is instead variable depending on the physical and chemical properties of the system. Air-water interfaces are unique in that molecules may diffuse from the liquid to the interface, diffuse from the interface to the bulk, or adsorb to the surface from the air.

#### 2.3.1. Air-Sea Interface/Sea Surface Microlayer

The addition of the ocean's physical and chemical properties, including wave breaking and diverse DOC from biogenic<sup>7,8</sup> and anthropogenic<sup>9</sup> sources, respectively, complicate the interface. Specifically, wave breaking leads to aerosolization<sup>10</sup> of organics enriched at the interface and within the SSML.<sup>11</sup> The thickness is often referred to as about 1 mm,<sup>12</sup> however inconsistencies in reported thickness throughout the literature are attributed to the variability of the interface globally. Despite lacking a definite thickness, the SSML exerts a unique control over global processes given its role as a boundary between Earth's atmosphere and ocean. It's hypothesized that the SSML contributes



greatly to the production of ice nucleating particles,<sup>7,13-15</sup> or **INPs**, and cloud condensation nuclei<sup>10,16-18</sup> (**CCNs**) both released in the form of a SSA, which in turn affects climate.

### 2.3.2. Enrichment Factor, Surface Activity, Monolayer Formation

The mechanism in which INPs and CCNs are released to the atmosphere is through wave breaking and bubble bursting, producing SSAs, as previously noted. However, their prevalence at the interface and specifically the SSnL is due to the physical and chemical phenomena arising from the anisotropy and solubility. Enrichment factors<sup>19-23</sup> (**EFs**) express the significance of a compound, molecular class, or inorganic ion presence in the SSML, such that an EF of 1 indicates no enrichment of a species occurs between bulk or interface. The mechanism or cause of enrichment is usually attributable to solubility, anisotropy, surface activity, and a compound's source.

Solubility is controlled by entropy and intermolecular forces. Carbon-dense molecules, such as lipids, are less soluble in water and have a higher affinity to aggregate in the SSML, with its less structured water, and form monolayers on the surface (SSnL). The formation of a monolayer, or sea slick, on the ocean's surface was first documented by Benjamin Franklin in 1785.<sup>24</sup> However, the chemical properties of the surfactant that caused the formation were not understood at the time.

Early measurements of the effect oil or surfactants had on the physical properties of the water surface were made by Agnes Pockels.<sup>25</sup> Using only rudimentary tools, she nevertheless revealed that the surface tension of "impure" water was greatly diminished as the surface area per molecule was decreased. In other words, the surface pressure increased when a monolayer was formed. This work was published in *Nature* with the assistance of

Lord Rayleigh in 1891.<sup>25</sup> Irving Langmuir's work in the field of surface science led to the discovery of the reason for surfactant organization at the surface.<sup>26,27</sup> His findings were that the oily tails (hydrophobic region) oriented out of the water into the air while the hydrophilic head groups interacted with the water. Katharine Blodgett worked closely with Langmuir and her work on the transfer of monolayers on an air-water interface to a solid substrate is integral to our understanding of modern surface science.<sup>28</sup>

#### 2.4. Infrared Spectroscopy

IR light, with wavelengths longer than visible light and shorter than microwaves, probes vibrational motion of covalently bonded atoms by exciting from one vibrational state to another.<sup>29,30</sup> Only vibrational motions with a temporary change in dipole moment are IR active. Additionally, the transition energy must be resonant with the infrared energy to cause a change. The approximate wavenumber, proportional to frequency, where bond vibrations are observed, can be determined.

Molecular stretching vibrations can be approximated as two masses connected by a spring; moving one mass on the same axis as the spring has a resulting vibration that is expressed as

$$F = -ky$$

Equation 5

Where the restoring force,  $F$ , is proportional to the force constant,  $k$ , and displacement,  $y$ . Commonly known as Hooke's law, this expression of simple harmonic motion is the basis for classical atomic vibration.

The potential energy,  $E$ , of the mass and spring can be assigned a value of zero at equilibrium and compressing or extending the spring changes the potential energy,  $dE$ , where

$$dE = -Fdy$$

Equation 6

such that the force and change in distance is equal to the change in potential energy.

Substituting Equation 5 in and integrating results in

$$E = \frac{1}{2}ky^2$$

Equation 7

which gives the potential energy of a harmonic oscillator.

Vibrational frequency is deduced starting from Newton's second law

$$F = ma$$

Equation 8

where  $m$  is mass and  $a$  is acceleration. Acceleration is defined as the second derivative of distance with respect to time

$$a = \frac{d^2y}{dt^2}$$

Equation 9

Equation 5, Equation 8, and Equation 9 are combined to give

$$\frac{d^2y}{dt^2} = -\frac{ky}{m}$$

Equation 10

wherein solutions require that the second derivative is equal to the original function multiplied by  $k/m$ . A cosine function would satisfy this requirement; such that a displacement function of

$$y = A \cos 2\pi\nu_m t$$

Equation 11

where  $A$  is the maximum amplitude and  $\nu_m$  is the vibrational frequency. Given the second derivative

$$\frac{d^2y}{dt^2} = -4\pi^2\nu_m^2 A \cos 2\pi\nu_m t$$

Equation 12

Substituting Equation 11 into Equation 10, and using the second derivative produces, after canceling terms, converting from frequency to wavenumber, and rearranging,

$$\nu_m = \frac{1}{2\pi c} \sqrt{\frac{k}{\mu}}$$

Equation 13

the spring constant,  $k$ , and reduced mass,  $\mu$ , provide an estimated vibrational mode derived from the harmonic oscillator energy level equation. A selection rule of allowable transitions ( $\Delta n$ ) from only  $\pm 1$  state limits the vibrational transitions that occur; however, overtones present in spectra are evidence that this rule is not infallible. In addition, the selection rule enables estimation of the number of modes, or normal modes, which is based on the degrees

of freedom present in a molecule. Linear molecules have  $3N-5$  modes and nonlinear molecules have  $3N-6$ .

## 2.5. Absorption of Light

Spectra, recorded from spectrophotometers, are presented in either transmission or absorbance mode, which are mathematically related to each other. Transmission spectra are produced by dividing the detected light from the sample by the detected light from a blank measurement.

$$T = \frac{R_s}{R_o}$$

Equation 14

Derived from Beer's Law, absorbance can be defined by a simple relationship to transmission.

$$A = -\log(T)$$

Equation 15

Most FTIR spectra are presented in absorbance because, below an absorbance value of '1', the relationship of absorbance and concentration is linear. This enables the analysis that is explored in Chapter 6.

## 2.6. Fourier Transform Infrared Spectroscopy (FTIR)

Most modern infrared spectrometers utilize Fourier transform to improve upon the signal-to-noise ratio and throughput.<sup>29</sup> The Fellgett, or multiplex, advantage results in improved signal-to-noise ratio as all wavenumbers are sampled at once.<sup>31</sup> The signal-to-noise ratio for an average of  $n$  measurements is given by

$$\left(\frac{S}{N}\right)_n = \sqrt{n} \left(\frac{S}{N}\right)_i$$

Equation 16

where S/N is improved by the square root of n. The disadvantage to this is that to improve the signal-to-noise ratio by a factor of two, the acquisition time must increase four times. A spectrum can be acquired much more quickly than having to sample each wavenumber individually and record the response. Instead, spectra are recorded in the time domain. However, in consideration with the Jacquinot, or throughput, advantage where there are fewer optical elements and the overall time to complete a scan is greatly reduced, the use of Fourier transform in infrared spectroscopy has greatly advanced the field and application of the instrumentation technique.<sup>32,33</sup>

While conventional spectroscopy records in the frequency domain, time domain spectroscopy changes with radiant power over time and is achieved by Fourier transform. The time domain signal is converted from the frequency domain by

$$P(t) = k \cos(2\pi\nu_1 t) + k \cos(2\pi\nu_2 t)$$

Equation 17

where k is constant, and t is time. The time domain contains all the same information as a frequency domain spectrum; conversion is done automatically by the recording computer.

Time domain spectra for the optical region are acquired using a Michelson interferometer. The frequency range of optical spectroscopy ( $10^{12}$ - $10^{15}$  Hz) necessitates modulation because transducers cannot record radiant power changes at such high frequencies. Instead, frequencies must be modulated proportionally to produce measurable frequencies; Michelson interferometers accomplish this by splitting one collimated beam

to two of nearly equal power. These two beams are recombined, and the modulated beam is measured as a function of differences in the lengths of the paths of the two beams.

One path length is fixed while a movable mirror creates a variable path length. The difference in path length for the two beams is referred to as retardation,  $\delta$ . The unconverted spectra obtained by the interferometer is power output as a function of  $\delta$ , from which a relationship of the two frequencies can be derived. One cycle of the interferometer is completed when the movable mirror travels a distance that corresponds to one half wavelength ( $\lambda/2$ ). Where the mirror is traveling at a constant velocity,  $v_m$ , and time,  $\tau$ , required to move the mirror  $\lambda/2$ , then we can express one cycle as

$$v_m \tau = \frac{\lambda}{2}$$

Equation 18

The frequency,  $f$ , of the signal at the detector is the reciprocal of  $\tau$ , such that Equation 18 can be written as

$$f = \frac{1}{\tau} = \frac{2v_m}{\lambda} = \frac{2v_m \nu}{c}$$

Equation 19

where frequency is related to wavenumber,  $\nu$ , and the interferogram frequency is related to optical frequency. When  $v_m$  is constant, the interferogram and optical frequency is directly proportional.

The cosine wave of the interferogram is described by

$$P(\delta) = \frac{1}{2} P(\bar{\nu}) \cos 2\pi f t$$

Equation 20

where  $P(\bar{\nu})$  is radiant power of the incident beam. One power in frequency domain and the other in time domain enables Fourier transform. Introducing the term  $B(\bar{\nu})$  accounts for an imperfect split of the beam, and substituting the relationship from Equation 19 and using the relationship of retardation and mirror velocity,

$$v_m = \frac{\delta}{2t}$$

Equation 21

yields:

$$P(\delta) = B(\bar{\nu}) \cos 2\pi\delta\bar{\nu}$$

Equation 22

Integrating over all wavenumbers,  $\bar{\nu}$ ,

$$P(\delta) = \int_{-\infty}^{\infty} B(\bar{\nu}) \cos 2\pi\delta\bar{\nu} d\bar{\nu}$$

Equation 23

and a Fourier transform enables the transition from one domain to another. The Fourier transform of Equation 23 is

$$B(\bar{\nu}) = \int_{-\infty}^{\infty} P(\delta) \cos 2\pi\delta\bar{\nu} d\delta$$

Equation 24

a rather elegant mathematical phenomenon. A complete Fourier transform requires both the real (cosine) and imaginary (sine) components. However, only the real function is shown here.



A few caveats require attention. The integrals, as written, are not the precise equations used by the computer because Equation 23 assumes infinite sampling range. Computational calculations require that  $\delta$  be infinitely small (small sampling interval). However, this is impractical in application. Only a finite sampling range can be summed over a finite  $\delta$  (a few centimeters). The result is a restriction in the resolution and sampled frequency range.

In general, resolution can be defined as the distance between two lines that is barely resolvable by the instrument, which can be written as

$$\Delta\bar{\nu} = \bar{\nu}_2 - \bar{\nu}_1$$

Equation 25

to give the resolution of the instrument. Measuring from arbitrary peak '1' to peak '2', where 1 has  $\delta=0$  and the waves are in phase at 2, the maximum of b occurs when

$$\delta\bar{\nu}_2 - \delta\bar{\nu}_1 = 1 \text{ or } \frac{1}{\delta} = \bar{\nu}_2 - \bar{\nu}_1$$

Equation 26

Combining Equation 25 and Equation 26 through substitution results in

$$\Delta\bar{\nu} = \frac{1}{\delta}$$

Equation 27

such that the resolution improves to the reciprocal of the distance the mirror travels.

### 2.6.1. Attenuated Total Reflectance

With a crystal positioned horizontally, mirrors direct the incident infrared light to the crystal where internal reflection causes the beam to travel across and to the detector

after sampling the crystal and source. ATR uses total internal reflection and the creation of an evanescent wave that probes the crystal and sample by just a few micrometers.<sup>29,30,34</sup> The penetration depth of the evanescent wave is determined by the wavelength, critical angle, and refractive indices of the crystal and sample.

$$d_p = \frac{\lambda}{2\pi\sqrt{n_c^2 \sin^2 \theta_c - n_s^2}}$$

Equation 28

The critical angle is given by

$$\theta_c = \sin^{-1} \frac{n_s}{n_c}$$

Equation 29

where  $n$  is the refractive index of the sample,  $s$ , and crystal,  $c$ . The infrared beam, after reflecting off the crystal, is directed to the detector. Some of the key requirements and subsequent benefits of ATR are that the crystal material must have a higher refractive index than the sample, such that the sample will not absorb the light without returning to the detector. ATR fundamentally has a short path length, which is beneficial in highly absorbing samples, such as water.

### 2.6.2. Infrared Reflection Absorbance Spectroscopy

Much like ATR, IRRAS uses reflection to sample monolayers on reflective surfaces, such as water.<sup>35</sup> These spectra provide information about monolayer structure and interaction with the subphase. The incident IR source is reflected from a gold mirror at an angle of  $48^\circ$  relative to surface normal (Figure 1). The angle is chosen specifically for its closeness to the Brewster angle of water (where all light is transmitted rather than reflected)

to reduce sampling of the subphase in the presence of a monolayer. After the IR light has probed the interfacial region, it is reflected to the detector via a gold mirror. The background, or reference, spectrum for IRRAS is the pure subphase before the monolayer is deposited. Absorbance is calculated with the same equations used for transmission FTIR to produce a reflectance-absorbance spectrum. Only about 6 % of the original IR intensity is reflected from the surface to the detector, requiring a sensitive detector. HgCdTe (MCT) detectors are most often implemented to collect the minute amount of signal. IRRAS is generally regarded as a surface sensitive technique, wherein the topmost micrometer of the subphase is probed along with the monolayer.

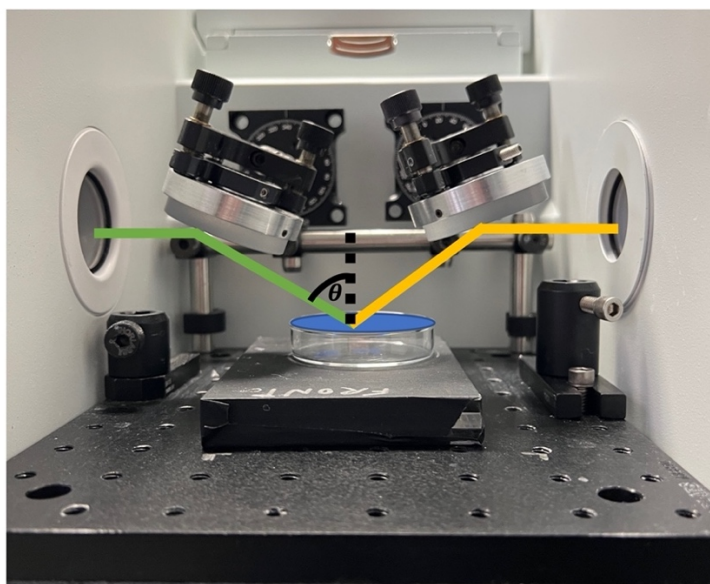


Figure 1. IRRAS configuration used in the Allen lab.  $\theta$  is angle relative to surface normal that the incident beam (green) is directed onto the surface (blue). Reflected light (yellow) is directed to the detector.

## 2.7. Computational Method Theory

The novelty of this dissertation is in the application of well-established laboratory techniques to ML for the advancement of achievable chemical analysis. Each chapter

includes background and brief theory of the applied ML techniques. Here, a rigorous detailing of the mathematical background is provided for the four main ML approaches used throughout each chapter: LR, PCA, SVR, and CNNs.

### 2.7.1. Linear Regression

LR is a mathematically simple yet useful technique in chemistry applications.<sup>36-38</sup>

Data with linear relationships may be well suited for LR. Using the relationship

$$y = mx + b$$

Equation 30

where  $y$  is the dependent variable,  $m$  is the slope of the line,  $x$  is the independent variable, and  $b$  is the intercept, concentration of an analyte in solution could be predicted, for example.

### 2.7.2. Principal Component Analysis

Broadly, PCA provides a reduction in data dimensionality.<sup>39-41</sup> It can be employed to determine underlying relationships among data, establish the mathematical components of which the data is constructed, or utilized as a preprocessing technique to eliminate superfluous information in the dataset.<sup>39,42</sup> PCA is accomplished, mathematically, by determining the best linear transformation of the data in which the dimensionality is reduced but the dataset is still described.

PCA is produced using an orthogonal linear transformation with singular value decomposition to project the data onto a new coordinate system.<sup>40,43</sup> The scalar values with the greatest variance are projected on the first coordinate and second greatest variance on the second principal component, until all variance in the data is explained or a set threshold

is met (e.g., 99% variance explained). Consider a dataset,  $x$ , that has several observations,  $d$ , and  $X$  is an  $x \times d$  matrix. The rows,  $x$ , represent each sample (e.g., a range of spectra from an experiment) and the columns,  $d$ , represent each datapoint (e.g., wavenumbers). A variance maximization function will identify the optimal linear combination of  $d$  such that variances are maximized. These linear combinations of  $d$  are known as the principal component (**PC**) scores,  $s$ , and the weights,  $w$ , are PC loadings.

The mathematical transformation is described by  $l$   $d$ -dimensional vectors with  $w$  that match the vector  $x$  of  $X$  to a new vector of  $s$ . Where  $s$  is

$$s_i = x_i \times w_k$$

Equation 31

and  $s$  inherits the greatest variance over  $X$ ,  $k$  is one through  $l$ ,  $l$  is less than  $d$  such that dimensionality reduction is achieved, and  $w$  is a unit vector.

Variance is maximized by establishing that solutions to  $w$  of the first component must satisfy

$$w_1 = \arg \max_{\|w\|=1} \left\{ \sum_i (x_i \times w)^2 \right\}$$

Equation 32

and when written in matrix form, the value being maximized is equivalent to the Rayleigh quotient. In other words, the maximum value of  $w$  is the matrix's largest eigenvalue, occurring when  $w$  is the eigenvector. The first PC is given as

$$s_{1l} = x_l \times w_1$$

Equation 33

and subsequent components are determined by

$$X_k = X - \sum_1^{k-1} Xww^T$$

Equation 34

Figure 2 provides a visualization of the matrix organization described herein.

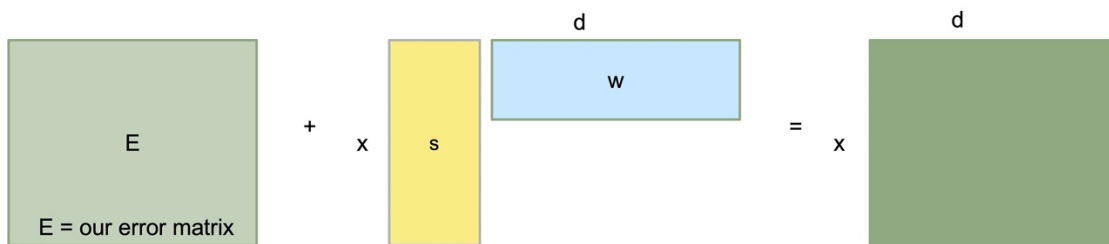


Figure 2. Principal component analysis general schema.

### 2.7.3. Support Vector Regression

First introduced by Drucker and colleagues in 1996, SVR is a subset of support vector machines (SVM).<sup>44</sup> SVMs are a supervised machine learning technique, meaning data is labeled during the training. SVMs are useful for classification, regression, and identifying outliers. In general terms, SVRs work to find the best hyperplane where the most data exists and distance from the plane is minimized for any data not in the hyperplane. A successful model will predict discrete values given input data. The overall benefit is that the data does not need to have linear correlations; hyperplanes of any function can be identified using an SVR.

In this dissertation, a linear SVR is employed. A linear model is determined by the minimization of

$$\frac{1}{2} \|w\|^2 \rightarrow 0$$

Equation 35

in the context of the loss function:

$$|y_i - \langle w, x_i \rangle - b| \leq \varepsilon$$

Equation 36

where  $y$  is the target value,  $w$  is the normal vector to the hyperplane,  $x$  is the input value,  $b$  is the intercept, and  $\varepsilon$  is a set tolerance threshold (i.e., how much error is acceptable in the model). The loss function of SVR, like SVM, does not include any data that is close to the model prediction such that training time is saved to optimize the model for data that deviates the most from the average.

#### 2.7.4. Convolutional Neural Networks

By far the most computationally intense ML technique considered, CNNs are an advanced image-based ML approach.<sup>45-47</sup> CNN theory is discussed in detail in the background section of Chapter 3 and the process of backpropagation is provided. Image-based ML is an incredibly advanced field of artificial intelligence, likely because of the plethora of applications that are afforded to an ML algorithm that can process information from images. Just a few examples include self-driving cars, playing chess, identifying faces in a crowd, scanning databases against a license photo, images from doorbell video being used to assist in solving a crime, or determining what kind of bird was seen on a walk. The

applications are unique, diverse, and complex, which is why image-based ML is a perfect tool to integrate into chemical problems.

Using transfer learning, a process of taking pretrained models where weights of the neural network have been adjusted and training only the last hidden layer, the advances of the image-based ML field can be harnessed in datasets with fewer examples.

Backpropagation in CNNs, and neural networks in general, improves the model while decreasing the overall training time required. Consider two nodes, connected to one another, as in Figure 3.



Figure 3. Nodes  $a^l$  and  $a^{l-1}$  are connected to one another. The variable  $l$  denotes location regarding each node.

A target node, where predicted values,  $y_i$ , are provided, is somewhere to the right of the node labeled  $a^l$ , for activation of layer  $l$ . A cost function can then be written as

$$c_o = (a^l - y_i)^2 \quad \text{Equation 37}$$

and the value of  $a^l$  is determined by

$$a^l = \sigma z^l \quad \text{Equation 38}$$

where  $z^l$  is

$$z^l = w^l a^{l-1} + b^l \quad \text{Equation 39}$$



These preceding functions comprise backpropagation; where  $\sigma$  is the nonlinear sigmoid function,  $b^l$  is the node bias. Backpropagation uses partial derivatives

$$\frac{\partial c_o}{\partial w^l} = \frac{\partial z^l}{\partial w^l} \frac{\partial a^l}{\partial z^l} \frac{\partial c_o}{\partial a^l}$$

Equation 40

to determine how a change in weight,  $w^l$ , effects the cost function,  $c_o$ , or loss. Given the first derivatives of Equation 37, Equation 38, and Equation 39,

$$\frac{\partial c_o}{\partial w^l} = 2(a^l - y_i)a^{l-1}z^l$$

Equation 41

a change in the weight is affected by the activation of the previous two nodes, true value, and  $z^l$ . A few different routes of interpretation can be followed from this solution. The cost of the change in weight, when substituting in a rearranged Equation 38, would be affected by the sigmoid function, or

$$\frac{\partial c_o}{\partial w^l} = \frac{2(a^l - y_i)a^{l-1}a^l}{\sigma}$$

Equation 42

An equally important observation comes from substituting Equation 39, which yields

$$\frac{\partial c_o}{\partial w^l} = 2(a^l - y_i)a^{l-1}(w^l a^{l-1} + b^l)$$

Equation 43

which illustrates that the value of the weight affects the cost function of the model.

## Chapter 3. Functional Group Identification for FTIR Spectra Using Image-based Machine Learning Models

Reproduced in part with permission from 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. Copyright 2021 American Chemical Society.

### 3.1. Introduction

The anthropogenic impact on the climate and environment has prompted the analysis and detection of pollutants or contaminants with FTIR such as microplastics in waters<sup>48,49</sup> and table salts,<sup>50</sup> nitrates from agricultural fertilizers in soil,<sup>51–53</sup> and polyaromatic hydrocarbons in the ocean’s surface<sup>54,55</sup>. The diversity of the chemical composition of the pollutants and the central fundamental technique of FTIR underscore the importance of a computational method for improved throughput of spectral analysis. The bottleneck is most frequently the assignment of peaks to relevant functional groups.<sup>56,57</sup>

Functional groups describe and define the physical and chemical properties of compounds.<sup>58,59</sup> Identification of many organic groups is accomplished via FTIR due to the associated unique vibrational frequencies.<sup>60,61</sup> Large numbers of spectra are time consuming to analyze and require expert chemist analysis to determine present composition. This limits the application of FTIR spectral techniques as a sampling method for functional group elucidation. There is thus an unexplored, yet applicable field of FTIR

spectra interpretation through statistical methods. Progress towards ML methods for environmental pollutant analysis has been explored for specific, targeted applications.<sup>56,62–</sup>

<sup>64</sup> Generalizable functional group ML models would increase the utility of FTIR sample screening in environmental and other chemistry applications.<sup>65,66</sup>

In this study, we investigate the implementation of CNNs<sup>47</sup> to identify functional groups present in FTIR spectra. By limiting spectral preprocessing, we explore a minimalistic approach to allow the network to learn spectral patterns for successful recognition of the 15 most common organic functional groups (Table 1). ML serves to address a need for quick identification of spectral components.<sup>67</sup> To date, the use of a CNN to broadly classify functional groups has not been reported. CNNs work by having layers of nodes called neurons, these neurons can be trained on data to identify spectral components that were observed in the training data in new spectra. The algorithm works to minimize a loss function; this is done by comparing answers given by the CNN to the true answers from a training dataset. The difference between the reported and the true presence of a group constitutes the loss function. The training dataset is a randomly segmented subset of spectra that the CNN uses to learn and adjust neuron weights.

CNNs expand upon artificial neural networks (ANNs) by using mathematical convolutions to provide convolved data to the following neuron. Each neuron has a receptive field for which it convolves the information, similar to how a human brain has regions of neurons designated for processing specific information.<sup>68</sup> CNNs significantly reduce the number of neurons per pixel that a traditional feed-forward network requires to capture the complexity of an image. Thus, CNNs are a sophisticated solution to the

alternative complex network required to machine learn images by capturing the spatial and temporal uniqueness of images. FTIR spectra offer unique “images” to evaluate using the sophisticated ML advancements.

We probe the effectiveness of image recognition ML as a facile solution to FTIR spectra interpretation. The information contained in a spectrum is most often presented to a chemist as a 2D image, therefore it is desirable to develop models that learn via similar spectral visualization.<sup>69</sup> Determining functional groups present in spectra requires analysis of both peak location on the frequency axis and shape; training models of spectra as images allows for an elegant approach that utilizes the totality of information obtained from a spectrum. Previous implementations of FTIR ML for functional group identification have limited,<sup>70</sup> averaged,<sup>71</sup> and segmented<sup>70,72</sup> spectral data to reduce information used during training. The computational resources available today make this an unnecessary and limiting feature. We include all available spectral data from 4000 to 600  $\text{cm}^{-1}$  to reduce any biases on the learning process.

Current methods for spectral processing and interpretation are limited to library searching software<sup>73</sup> and highly specific questions using implementations of ML including: Support Vector Machines,<sup>56,74</sup> k-Nearest Neighbors,<sup>75,76</sup> and Principal Component Analysis (PCA)<sup>56,74,75</sup> or Factor Analysis<sup>77</sup>. Library searching methods require a pre-existing and transferrable database for searching spectra. The initial creation of libraries requires an intensive endeavor for collecting a large enough spectral repository. Once implemented, libraries cannot extrapolate beyond those included in the software. The size of libraries is not of significant concern for storage, but it is a cumbersome feature for

application compatibility and relative use-to-memory consumption. ML does not require transfer of training data to the user and can predict beyond the data used for training. The use of ML to resolve challenging implementations of FTIR spectra (e.g., extremely large datasets, continuous analysis) has become of interest as increased processing power makes it possible to train and infer (interpret an unknown spectra) with complex algorithms.<sup>78-81</sup> However, these highly specific models are only applicable in the setting in which they are developed because the training is completed on a narrow range of examples. To increase the amount of available training spectra or improve further calculations, ML algorithms in tandem with molecular dynamics have been explored.<sup>81-83</sup>

Previous applications<sup>84-86</sup> of ML have employed data preprocessing prior to training with unsupervised ML methods, such as PCA<sup>74</sup>, which reduces the information in the training data. Significant mathematical spectral preprocessing is becoming an unnecessary component with the advances in ML. Data dimensionality reduction limits the transferability of the final model to broader applications. Deep learning results in feature extraction within the model before and during learning. ML requires any feature extraction (e.g., selecting peaks of interest) be completed by a user (manually or automatically) before training the algorithm. Results from recent studies identify little variability in prediction success between preprocessed data (e.g., removing wavenumber regions, derivative spectra, and components resulting from PCA) and raw data pipelined to sophisticated ML methods.<sup>46,87</sup> A recent application of ML successfully implemented broader methods for functional group analysis, however the authors utilize a multilayer perceptron ML method with an autoencoder and train using two sources of data: FTIR and MS spectra.<sup>88</sup>

Implementing methods such as selecting spectral regions of interest<sup>70</sup> can result in learning becoming memorization by the model; rather than making a generalizable algorithm that can inference on novel spectra, the model overfits the training data. An overfit model does well on spectra it has seen before but performs poorly on new data. Showing select data based on human evaluation increases the time required by an expert and introduces additional bias. While there are regions of relative disinterest to the chemist, it is not sufficient to ignore them in training. The absence of a peak is equally informative as the presence of another.

In our work, we create separate functional group models that are executed simultaneously, resulting in complete analysis of FTIR spectra. We obtain spectra from the NIST Chemistry Webbook; peaks are not labeled in this dataset. The use of individual functional group models presents a robust approach to establish a broad but precise computational analysis of spectra. Training a model for each functional group improves the overall accuracy attainable because each model is focused on a binary question: is this functional group present? The training of individual models does not impede speed of spectrum analysis achieved and results are provided succinctly. By approaching the classification of spectra via the proposed method, we reduce the likelihood that the model learns a connection between functional groups that is not chemically relevant. In other words, one present functional group does not indicate another group's presence or absence. Individually trained models reduce the potential for this and improve the overall accuracy by posing a simplified question. Here we develop effective and accurate FTIR ML models

that apply to broader questions, limit spectral preprocessing, and provide the entire spectrum to the algorithm.

## 3.2. Methods

### 3.2.1. Python Scripts

All Python scripts can be accessed from our repository at this address: <https://github.com/Ohio-State-Allen-Lab/FTIRMachineLearning>. The FTIR spectra are property of NIST and can be accessed through their website. The implementation of Inception V3<sup>47</sup> is modified for our use and the original source is linked on our repository with the published modified version. The computational procedure is described in detail in Appendix A and is documented in each Python script.

### 3.2.2. Spectra Collection

Data was obtained from the National Institute for Science and Technology Chemistry Webbook via a web scraping implementation in Selenium using the CAS number identifier from the official list of compounds in the WebBook.<sup>89</sup> When a compound had an FTIR spectrum, the file, in jcamp-dx format, was retrieved and stored with the CAS number as the filename. A total of 8,728 spectra from pure compounds in gas phase were obtained (Figure 4). Each spectrum's InChI key was saved in a collective text file.

### 3.2.3. Data Pre-Processing

Only spectra in absorbance and wavenumbers were used for training models. Each spectrum was evaluated to ensure it was in absorbance and wavenumbers via a Python script. Files in transmission or wavelength were relocated to a distinct directory to preserve

all spectra obtained from web scraping. Files in the correct mode were converted from jcamp-dx to csv. Once converted, each spectrum was normalized so that the maximum peak height was 1. Normalized spectra were saved as jpg images.

#### 3.2.4. Labeling

Functional groups were identified via the InChI key. Using SMARTS functional group identifiers, each spectrum's key was parsed to return binary indicators. Present functional groups are labelled as "1" and absent as "0". For example, a molecule containing R-COOH would have a "1" in the carboxylic acid field. Results were saved in one spreadsheet with CAS numbers as spectrum and file identifiers. Spectra were copied into directories based on presence or absence of a functional group. This method allows one compound with multiple functional groups present to be copied into the directory for each group. Each of the 17 functional groups had two directories: positive and negative cases. Positive cases include the functional group and negative cases do not contain the group. Randomly ten photos, five from positive and negative, for each functional group were reserved for validation. Then, the directory containing more instances for a given group was reduced randomly until both directories contained the same number of spectra.

#### 3.2.5. Machine Learning

A CNN for image recognition was employed. A unique model (independent of all other models) was trained for each functional group to predict on two classes: present or absent. The functional groups and the number of images in the positive cases are presented in Table 2. The architecture, Inception V3, was accessed from the available models on the



Google TensorFlow library. Each model was trained for 10,000 epochs at a learning rate of 0.01, using an initialized version of Inception V3 and training the last layer of the model graph (Figure 5).

Inception V3, a 42 layer CNN, employs several techniques that ultimately led to an increase in accuracy in the final model results. The model begins with preprocessing the input spectra, which includes decoding and reducing spectra to 299x299x3 or pixels by pixels by RGB channel. The RMSProp optimizer results in the greatest accuracy. The specific equations describing the available optimizers are provided (Appendix B). As a post-optimization step, exponential moving average is employed (Appendix B). Batch normalization aids in reducing the time to convergence and occurs after convolutions in Inception V3; specific details, including the equations used, are provided (Appendix B). Additionally, learning rate adaptation is utilized to efficiently train the algorithms. Using gradual learning rate ramp-up, the initial learning rate for the model is approximately 10% of the defined rate; after initializing, the rate is linearly increased until the slope of the decay rate intersects with the theoretically defined exponential decay rate. More specific details are included in the Supporting Information.

Models are trained and validated using an 80/20 split. For example, if 100 spectra are provided, 80 are used for training and 20 are used for validation. The preprocessing methods includes distortion of the spectra such that after each backpropagation, the 80 spectra are altered and are not identical between iterations. Validation spectra are used to guide backpropagation but do not affect the training directly. The final train and validation accuracies presented are from the described source. After the models converge, the

reserved ten spectra are used to test the models further. These spectra are not used to adjust the model and present a more accurate representation of the model accuracy beyond the current dataset.

Parameters are initialized to reduce the time and computational power required to train a custom model. Models converge within the 10,000 training steps and early termination of training is not employed. It took five hours to train the 15 models. Classification of an unknown spectrum requires one minute.

### 3.2.6. Accuracy and Loss

Accuracy and cross entropy (loss) for both training and test models was obtained and saved as csv files. The final accuracies and entropies for training and test results from each model are investigated to identify any anomalies.

### 3.2.7. Classification of Validation Data

When spectra were classified, the models were called upon to infer (determine the functional groups present) and a result was provided. Each functional group model was trained and validated separately, and the predictions were not used in conjunction to attempt ensemble classification. Models were evaluated independently of each other and from the embedded validation methods to further analyze prediction accuracy. The ten reserved spectra were analyzed via the respective models they were withheld from to examine the learning quality of the algorithm. Confusion matrices<sup>90</sup> were used to represent the true and predicted functional group for the 17 models.

### 3.3. Results and Discussion

Using approximately 9,000 gas phase FTIR spectra, we train 17 functional group models using image-based ML. Our methods result in 15 effectively identified functional groups (Table 1). Each model is trained independently. When a molecule has more than one functional group present, the models for the relevant functional group will identify the presence from the spectrum resulting in a complete analysis for the 15 trained functional groups. Accuracy and cross entropy results from the last step of training are reported for the train and validation process. The two functional group models that underperform are aldehyde and nitrile, based on model prediction of untrained spectra. We define underperforming as misidentifying more than 60% of test cases. Nitrile vibrational modes, for example, are less than  $100\text{ cm}^{-1}$  from carbon dioxide vibrational modes. In general, characteristic nitrile peaks are easily identifiable from an IR spectrum. Yet, the models difficulty in identifying the functional group likely arises from a convoluted wavenumber region. The training accuracy is a measure of how well the model classifies the training data, which it used to train the network. A higher training accuracy indicates that the model is learning the training spectra. Validation accuracy expresses the ability of the model to generalize to untrained spectra, which is determined by the number of correctly classified validation spectra. Thus, it is more meaningful to have a higher validation accuracy, albeit not a requirement for a successful inferencing model. Cross entropy is the loss function used to evaluate the final model and is defined as the logarithm of the likelihood of a correct assignment. Smaller cross entropy values indicate a model is well trained. We observe cross entropy for training is less than validation. Models are more likely to correctly

inference spectra that have been used to train and adjust weights, in comparison to the validation spectra.

A confusion matrix for each model was created by using spectra that have been withheld from training and testing data. A confusion matrix compares model assignments to the actual identities of the samples; it shows correct assignments along the trace of a matrix and false assignments off the trace. Four models have perfect confusion matrices from classification of ten withheld images, five containing and not-containing functional group spectra examples. The presence or absence of carboxylic acid, aromatic, methyl, and ester functional groups are correctly identified in the withheld spectra (Figure 6).

The number of instances of each functional group occurring in the spectra varies significantly, with aromatic-containing spectra occurring most frequently with 3,467 images. In contrast, acyl halide has 85 spectra for training and testing the model. We explored the relationship between the number of images and the cross entropy and accuracy for training and testing results (Figure 7). Training accuracy decreases with increasing number of spectra (Table 3). However, the final accuracy, determined by evaluating the unknown spectra for functional group identification, is not correlated to the number of images used for training. Our results indicate that the total number of training spectra does not affect the final performance of the models. The scattered, non-uniformity exhibited in Figure 7 (a) and (b) depict the deviation from a linear relationship between the number of spectra and accuracy and cross entropy for validation, confirming the number of images is not influencing the performance of the models. Training accuracy provides insight into how well the model has learned the training images for a functional group model.

Counterintuitively, few spectra being trained for a functional group will result in a higher training accuracy because the model trains on the same spectra more frequently. This model memorizes or overfits functional groups, resulting in a model incapable of extrapolating to new spectra.

However, from our results the challenges of limited training spectra do not result in less accurate models. We confirmed this by investigating the relationship of number of images per class as a function of validation accuracy and cross entropy (Table 3). Models that have more spectra to train on have lower overall training accuracy but still perform well when analyzing unknown spectra. To investigate the linear correlation between the number of spectra used for training and the final accuracy and cross entropy, the Pearson's correlation coefficient is used (Figure 7). More linearly correlated relationships have a coefficient closer to one, where positive coefficients indicate a positive correlation and negative coefficients indicate a negative correlation. The coefficient for training accuracy and number of training spectra indicate they are indirectly correlated, whereas the coefficient for training cross entropy and number of training spectra is positive, or positively correlated. However, the models with less training spectra show no correlation between final accuracy and ability to classify unknown spectra. Furthermore, both validation accuracy and cross entropy do not have significant linear correlation. While training results display correlation with the number of images, the validation data indicates that models are successful with a range of number of training spectra.

We can determine some of the underlying shortcomings in the model, from both spectroscopic and computational perspectives, by investigating two functional groups:

aldehyde and carboxylic acid. The model results for aldehyde are promising for the training data but do not perform as effectively in validation and testing (Table 3). The confusion matrix for carboxylic acid describes how well the model performs on spectra that have not been used for training or validation. We observe that the IR mode frequencies for the carboxylic acid and aldehyde affect the performance of the model, in addition to the number of spectral examples available for training and validation.

Aldehyde C-H stretching frequency (2830- 2695  $\text{cm}^{-1}$ ) is commonly overlapping in organic spectra with other C-H bonds because it is a weaker mode (Table 3). The carbonyl stretch is also frequently unresolved in compounds that contain multiple oxygen atoms. The C-H bending mode is often weak, in addition to being in the fingerprint region, which is a challenge to interpret due to the complexity. With these stretching and bending modes considered, it is reasonable to anticipate that an aldehyde functional group is challenging for the model to identify in spectra. In comparison, carboxylic acid functional groups are always correctly identified in spectra by the model. The model for carboxylic acids is well trained. As observed by the validation accuracy and cross entropy (Table 4), the carboxylic acid model has a more robust transferability to spectra it has never observed. We confirm the effectiveness of the model with a correct assignment of unknown spectra. In totality, there are more carboxylic acid training spectra, and the IR modes are better resolved, especially the strong COO-H stretching, in comparison to aldehydes (Table 4).

We do not specifically probe the temperature and pressure dependence of model success. However, we hypothesize that the models would perform at high accuracy because pressure effects have been shown to have minimal effect on IR response.<sup>91</sup> In general, the IR transition moment strength for hydrocarbon bonds decreases with increasing temperature and this may affect model accuracy.

From our results, we observe that the models are more accurate for functional groups when there are more training spectra examples for the functional group and IR peaks are well resolved. Albeit this is an intuitive result for a trained spectroscopist with respect to accuracy correlating to peak resolution, yet there is no precedent using a machine learning approach.

#### 3.4. Conclusion

We present a novel method for FTIR spectral interpretation using CNNs and the NIST database. Fifteen functional group models successfully and effectively classify unknown spectra in a facile method for spectral submission to interpretation. We find that the image recognition features inherent in CNNs are transferrable to a chemical-identification application. From our observations, we can conclude that CNNs are effective at identifying spectral features for classification and generalizable models are achievable with ample spectral examples. In future work, optimization for functional group identification with fewer spectral examples should be investigated to improve accuracy. Further investigation of the models could include determining the ability of the model to predict binary and higher-order mixture compositions as well as shifts along the frequency axis.

Table 1. Functional groups for which successful models were trained.





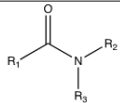



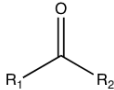
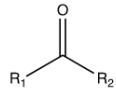

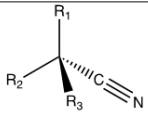
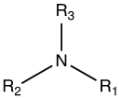
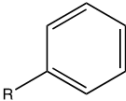
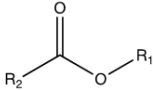
alcohol	alkane	alkene	alkyne	amide
				
ether	acyl halide	alkyl halide	ketone	ketone
				
carboxylic acid	nitrile	amine	aromatic	ester
				



Table 2. Functional groups are presented with the total number of positive spectra examples used in training as well as the total number of functional group examples used.

<b>Functional Group</b>	<b>Number of Positive Spectra</b>
alcohol	2,212
alkane	2,542
alkene	1,095
alkyne	209
amide	152
amine	770
aromatic	3,467
carboxylic acid	581
ester	904
ether	2,033
acyl halide	85
alkyl halide	2,266
aldehyde	198
methyl	2,941
nitro	414
ketone	743
nitrile	345
<b>Total Instances (spectra)</b>	<b>20,957</b>

Table 3. Final accuracy and cross entropy for train and validation of each functional group model is presented in order of increasing number of training images.

	<b>Accuracy</b>		<b>Cross Entropy</b>	
	<b>Train (%)</b>	<b>Validation (%)</b>	<b>Train</b>	<b>Validation</b>
acyl halide	100	98	0.025347	0.143665
amide	100	70	0.060037	0.900095
aldehyde	100	80	0.04735	0.385665
alkyne	99	80	0.079596	0.332745
nitrile	97	65	0.17019	0.668148
nitro	98	89	0.12624	0.668148
carboxylic acid	98	98	0.070173	0.076216
ketone	93	76	0.228837	0.501178
amine	93	80	0.24136	0.494815
ester	97	83	0.111057	0.323917
alkene	85	68	0.407803	0.743543
ether	89	81	0.27397	0.443644
alcohol	90	86	0.236544	0.330591
alkyl halide	85	73	0.350733	0.531302
alkane	85	90	0.327755	0.265718
methyl	81	84	0.384048	0.358021
aromatic	92	89	0.199645	0.259044

Table 4. Aldehyde and carboxylic acid IR stretching and bending mode frequencies.

	Mode	Frequency (cm <sup>-1</sup> )	Appearance
Aldehyde	C-H stretch	2830-2695	Weak, medium
	C=O stretch	1740-1720	Medium, strong
	C-H bend	1390-1380	Weak, medium
Carboxylic acid	O-H stretch	3300-2500	Strong, broad
	C=O stretch	1760	Strong
	O-H bend	1440-1395	Medium

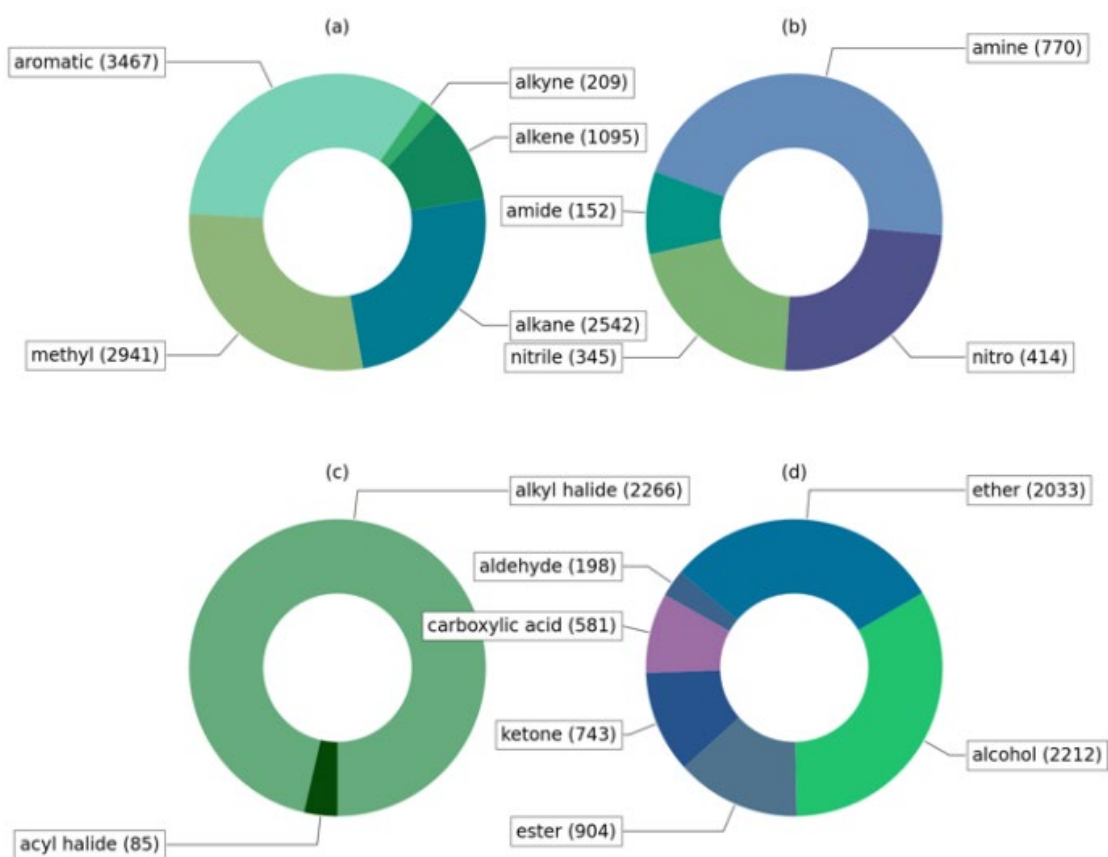


Figure 4. Number of spectra used to train each functional group model, (a) carbon-containing, (b) nitrogen-containing, (c) halide-containing, and (d) oxygen-containing. The number of images is equivalent for the positive and negative cases used in training and testing.

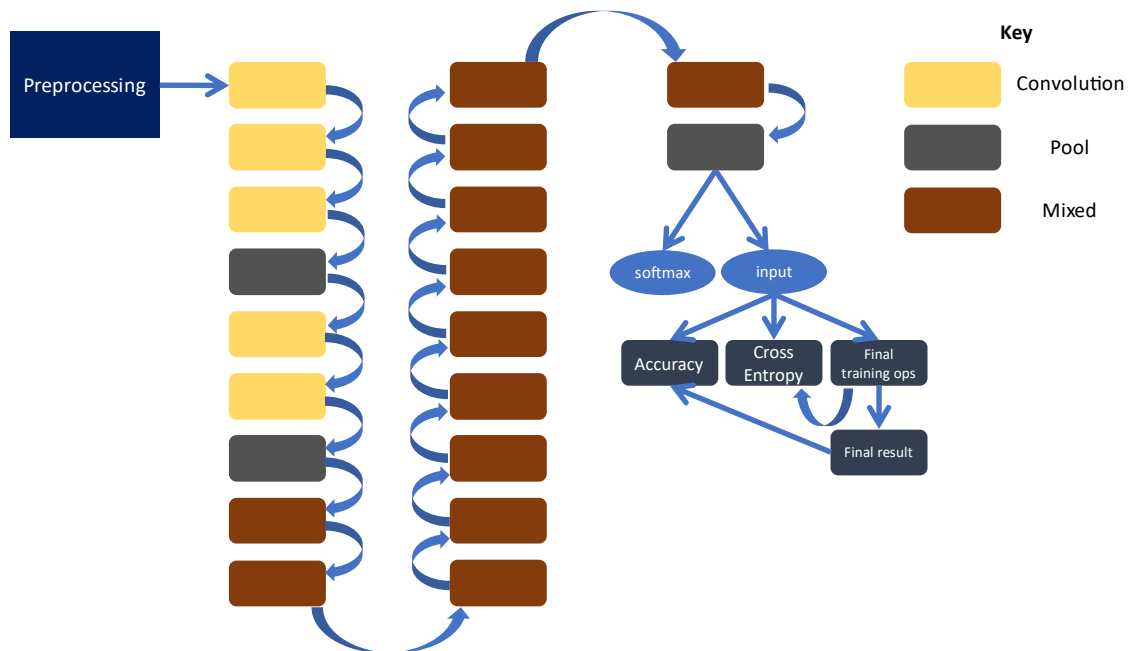


Figure 5. General summary of Inception V3 architecture. Additional details are provided in the Supporting Information, including a summary of the model preprocessing parameters.

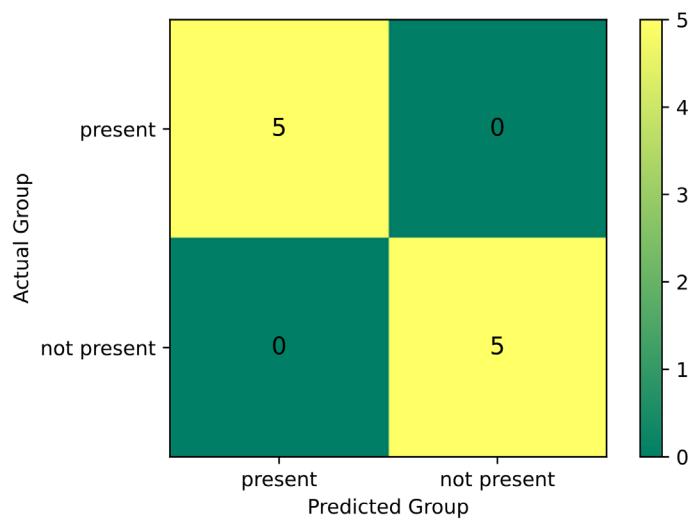


Figure 6. Confusion matrix is identical for carboxylic acid, aromatic, methyl, and ether functional group models.

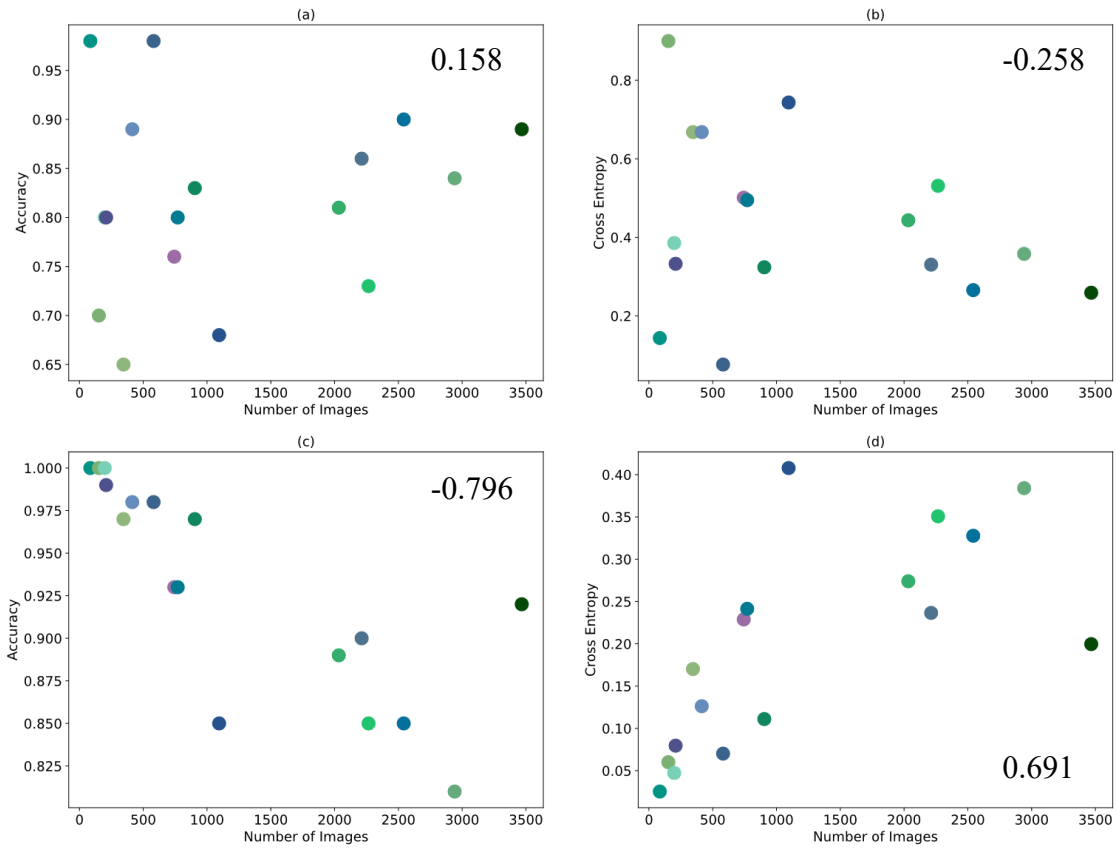


Figure 7. Final train and validation accuracy and cross entropy as a function of the number of spectra used to train each functional group. Insets (a) and (b) show validation results; (c) and (d) show training results. Pearson's correlation coefficients (PCC) are inset in the plots for final accuracy and cross entropy of training and validation as a function of the number of spectra. The coefficients closer to  $\pm 1$  indicate that the train accuracy and cross entropy are linearly correlated (negative is inversely correlated and positive is directly correlated) to the number of spectra used in training. Validation accuracy and cross entropy are not linearly correlated to the spectral examples used.

## Chapter 4. Carbon on the Ocean Surface: Temporal and Geographical Investigation

Reproduced in part with permission from Enders, A.A.; Elliott, S. M.; Allen, H.C. “Carbon on the ocean surface: Temporal and geographical investigation”. *ACS Earth Space Chem.* **2023**, 7, 2, 360–369. Copyright 2023 American Chemical Society.

### 4.1. Introduction

Carbon is the single most essential element for existence of life on Earth.<sup>58</sup> From polymeric backbone support of diverse biomacromolecules to the varietal self-bonding of which this atom is capable, its impact seems almost infinite.<sup>58,59,92</sup> This is especially true when considering the effect of increased atmospheric gas concentrations on Earth as a unified ecosystem.<sup>93,94</sup> Since the Industrial Revolution, global usage of fossil fuels has steadily increased.<sup>95,96</sup> A clear display of increased carbon consumption over the last few decades is the stark rise in atmospheric carbon dioxide; CO<sub>2</sub> is the thermochemical fate for all hydrocarbons in the oxidizing atmosphere.<sup>97</sup> Prior to the 1850s, the global concentration of CO<sub>2</sub> was around 280 ppm, but by November of 2020 measurements exceeded 410 ppm.<sup>98</sup> While terrestrial sources of carbon are primarily anthropogenic,<sup>95</sup> the ocean also has a vital role in the carbon cycle through several biogeochemical mechanisms.

Relative to geocycling, emphasis is frequently placed on ocean acidification since the dominant aqueous form of CO<sub>2</sub> is carbonic acid.<sup>99</sup> Yet, the relationship between carbon and Earth’s ocean is multi-faceted. Seawater is both a sink and source for carbon, because



of uptake and return by aquatic organisms extending from mixed layer depths to the sea surface and beyond (Figure 8. Simplified schematics illustrating (a) relationships between the marine boundary layer (white), sea surface nanolayer (light blue), and sea surface microlayer (dark blue) and (b) highlighting some of the major oceanic processes that occur including vertical transport from the bulk, enrichment of organics at the surface nanolayer, adsorption of atmospheric aerosols and gases, and release of sea spray aerosols from the ocean to the atmosphere.(Figure 8).<sup>12</sup> The sea surface nanolayer (SSnL) is about 1 nm of thickness (i.e. the thickness of one molecular layer) and is the topmost surface of the sea surface microlayer (SSML, ~1-1000  $\mu\text{m}$  thickness). The SSnL is enriched in organic molecules due to their hydrophobic properties, and these molecules tend to form in an ordered layer at the ocean surface, in a monolayer; the carbon within this monolayer is what we model globally. Since the 1960s, enrichment of organics in the SSML and SSnL has been evidenced by analyses of proteins, lipids, and carbohydrates throughout multiple field studies<sup>100-102</sup> and laboratory experiments.<sup>103-108</sup> However, precise organic composition is difficult to characterize for many reasons, including variable biogeochemistry and transport dynamics within the SSML and SSnL.

Research into the carbon present in the SSML and SSnL is either from field measurements or models that use satellite data or field data to map global estimates.<sup>8,109-</sup>  
<sup>112</sup> We invoke a previously established method of using phytoplankton from chlorophyll satellite data to model carbon.<sup>113</sup> We present a method to calculate net amounts of carbon resident in the SSnL utilizing Energy Exascale Earth System Model (E3SM) model output; our results are reported as normalized carbon and the well-known global budgeting unit of

gigatons. Our effort is equally complicated and enriched by the ocean's inherent biogeographical diversity; details of which are directly reflected in the SSML and SSnL.<sup>114,115</sup> We estimate the overall SSnL mass as a reference through the implementation of previously defined physicochemical properties of the air-sea interface. Specifically, we apply Gibb's equations for surfactant thermodynamics utilizing a modified Langmuir isotherm expression,<sup>116</sup> the concept of an oceanic equation of state,<sup>117</sup> and equilibrium expressions for adsorption and desorption of complex organics at air-liquid interfaces.

Our model parameterization is highly simplified and thus we acknowledge that the carbon distributions used here are not fully consistent with the reported DOC and particulate organic carbon (POC) pools, and their surface prevalence. This disconnect is unavoidable. Specifically, there is a knowledge gap in knowing adsorption properties of the total DOC and POC pools because of the inherent dynamic nature of the ocean and the chemical complexity of the pools. For the first modeling attempt in the present work, we assume that the DOC and POC pools retain properties like their initial biomolecular forms in phytoplankton, such that a carbon distribution of 60 % protein and 20 % lipid exists. Our percentages do not explicitly account for the lifetime, decay, or the resultant aged products and the effect that it has on surface activity or enrichment, in general. For example, we assume lifetimes of lipids and proteins based on literature values,<sup>118</sup> but depending on the conditions the decay rate can vary significantly as presented by Duffy and colleagues for proteins<sup>119</sup> and He et al. reported decay rates under hypoxic conditions.<sup>120</sup> There are limitations with that assumption because we know from the literature that lipids, proteins, and carbohydrates are broken down at different rates,<sup>121</sup> and future work will have to refine

and improve our model as more data is published. The model we have developed herein requires these fundamental assumptions, as described, to result in a global approximation of SSnL carbon.

Two specific compounds serve as proxies for the protein and lipid adsorptive contributions, bovine serum albumin and stearic acid. We explicitly address the cyclic nature of marine biogeochemical underpinnings; growth, conservation, release, and equilibrium-adsorptive reorganization are simulated for the organic composition of our model nanolayer. Moreover, all quantities are subject to geographic and temporal scaling. Calculated values are compared to total global carbon budgets to provide a perspective on the influence of contributions from biomacromolecules when constrained to the ocean SSnL. Our model values represent a low-end estimation of the global carbon because we only include two molecular classes. To the best of our knowledge, these computations are the first of their kind and they represent a unique portrait of the marine SSnL, the topmost surface layer of the SSML. Overall, we provide insight into the variability of the ocean nanolayer chemistry on a global scale.

#### 4.2. Methods

For our model approximation of carbon in the SSnL we use an assembly of proteins and lipids, which all enrich this region of the ocean.<sup>12</sup> We identify these components based on the experimental evidence from Cochran et al.,<sup>122</sup> Schiffer et al.,<sup>123</sup> and Pham et al.,<sup>124</sup> among other literature in the field. Specifically, sea spray aerosol formation is dependent on the SSnL and SSML composition, which we are focused on; Cochran and colleagues identify almost 300 surfactants that are best classified as lipids.<sup>122</sup> Work conducted by

Schiffer et al. and Pham et al. investigated the biological impact on sea spray aerosols, providing us with a framework to understand how the SSnL is influenced by marine biology.<sup>123,124</sup>

Enrichment of the SSnL occurs, in large part, because of gas and liquid interfacial phenomena; that is, unfavorable interactions between non-polar, carbon-rich organics and polar water.<sup>3,116</sup> The protein and lipid concentrations are driven by their release from phytoplankton. Here, we use a simplified model system consisting of proteins and lipids as a baseline. From that baseline, we then scale up to a global representation to provide flexible estimates for carbon mass in the SSnL.

#### 4.2.1. Proteins

Proteins are a common exudate or lysate from aquatic species. They are primarily injected into the water column when organisms are disrupted by grazing or senescing.<sup>104</sup> Graham and Phillips investigated the behavior of three model proteins:  $\beta$ -casein, bovine serum albumin, and lysozyme, with data taken relative to laboratory air-water interfaces.<sup>106</sup> Their results indicate a partly irreversible adsorption of proteins to the air-sea interface but with two-phase equilibration occurring as well, and the total resulting in a surface pressure maximum of 20 mN/m and film thicknesses of up to 50-60 Å. Certain processes, such as salting-out, which we expect to occur, can result in greater SSnL coverage and a lowering of measured surface tension (increased surface pressure).<sup>125,126</sup>

#### 4.2.2. Lipids

Lipids, such as fatty acids, phospholipids, and cholesterol, disrupt the surface tension of water through specific amphiphilic enrichment.<sup>3,116</sup> The non-polar tail chains of fatty acids, for example, orient into the air while polar headgroups interact with water itself resulting in a stable monolayer.<sup>127</sup> Headgroups also interact with inorganic cations, further stabilizing the organic films in the SSnL.<sup>19,128</sup> A characteristic monolayer (one molecule thickness) of stearic acid on pure water can reach surface pressures upward of 65-70 mN/m.<sup>129</sup> Organic films tend to stabilize and calm rough seas through their capacity to dampen waves. Wave breakage is a source of bubble bursting and aerosolization of surfactants.

#### 4.2.3. Carbohydrates

Carbohydrates are more soluble and less surface active in aqueous solutions relative to proteins or lipids, but they are still observed in the interfacial region, particularly in the SSML, although it is debated as to its concentration in the SSnL.<sup>20,103,130</sup> Satellite<sup>110</sup> and field<sup>131</sup> studies suggest that carbohydrates are an estimated 20% of the dissolved organic carbon (DOC) in the SSML.<sup>130</sup> By comparison and on an absolute basis, carbohydrate adsorption is weak. Only through processes such as co-adsorption are carbohydrates significantly adsorbed to the SSnL.<sup>20,130</sup> Burrows et al. were able to connect the co-adsorption process through a two-layer Langmuir model for fractional surface coverage that ultimately improves their sea spray model.<sup>108</sup> There are no well accepted satellite proxy data for the co-adsorption in the presence of mixed monolayers to establish carbohydrate

SSnL concentrations; therefore, we must neglect their net contributions. We acknowledge that their presence and geochemical role is non-negligible and should be incorporated in the future after further laboratory experiments on mixed monolayer co-adsorption have been validated. Thus, our estimate is a lower limit in establishing total carbon.

#### 4.2.4. Chlorophyll Data and Plankton Concentration

Chlorophyll is monitored at the planetary scale by several satellite instruments (e.g., NASA MODIS) and modelled through E3SM, among others. Model output is compared to chlorophyll satellite data to confirm its accuracy.<sup>132</sup> Monthly averages provide a convenient means to understand geographic and seasonal variability of upper ocean biomass. Phytoplankton are the “primary producers” of the sea, and cell densities are proportional to remotely measured chlorophyll. Specifically, pigments are essential to the photosynthetic process; light-absorbing conjugated bond systems constitute a relatively constant proportion of intracellular compounds. Autotrophs can only live near the ocean surface, because of the obligate need for sunlight.<sup>39</sup> Dissolved organic carbon primarily originates from primary production, or phytoplankton.<sup>110,112,134–136</sup> Literature suggests that using chlorophyll to model carbon and phytoplankton is a viable approach.<sup>110,118,136–140</sup>

#### 4.2.5. Carbon Calculations

All calculations and figures were done using Python scripts. E3SM chlorophyll data were averaged per month for 2005 on a scale of approximately 0.5 by 1 degrees latitude by longitude, respectively.<sup>132</sup> The satellite results are converted from chlorophyll to planktonic carbon concentration ( $C_p$ ) using the standard ratio of 50:1 (planktonic C to chlorophyll by

mass in grams).<sup>137</sup> The ratio is an imperfect representation of all phases of plankton blooms and we assume carbon averages across a bloom to the defined ratio. Equation 44 accounts for several ocean biogeochemical dynamic processes and ultimately provides a dissolved carbon concentration ( $C_i$ ) for the  $i^{\text{th}}$  biomacromolecule. Here  $i = 1$  for protein and  $i = 2$  for lipid, but this vector can be expected to lengthen in future studies to account for the diverse chemical pool. Zooplankton maximum growth rate ( $g$ ) is estimated from literature values<sup>141</sup> and zooplankton ( $C_z$ ) values obtained from E3SM output.<sup>132</sup> The model provides zooplankton in moles carbon. We average monthly from daily outputs of E3SM for the calendar year of 2005, which corresponds with available satellite measurements.

$$C_i = gC_z \left( \frac{C_p}{K_{ingest} + C_p} \right) (1 - \gamma)\tau_i p_{i,\%}$$

Equation 44

The consumption of phytoplankton due to grazing is limited on a kinetic basis and is accounted for through the relationship ( $C_p/(K_{ingest} + C_p)$ ), where  $K_{ingest}$  is the half saturation constant for ingestion by zooplankton. We adopt rough global average values for key variables (i.e.,  $K_{ingest}$ ,  $\tau$ ) established in early ocean models.<sup>141</sup> The local concentration of carbon is further modulated by grazing assimilation efficiency ( $\gamma$ ), and steady state is achieved with the mixed layer lifetime ( $\tau$ ) of the  $i^{\text{th}}$  species. Lastly, we address the fraction or percentage of each molecule (protein and lipid) initially residing in an autotrophic cell. Using literature values, we approximate the carbon associated with proteins and lipids ( $p_{i,\%}$ ) to be 60% and 20% of biomass, respectively.<sup>117</sup> All of the parameters are summarized in Table 5.

The conversion from concentration to SSnL mass requires realistic models of molecular adsorption to the air-water interface. Classic Langmuir isotherms are capable of modeling relevant population competition to zeroth order.<sup>3</sup> Extending beyond the idealization of Langmuir allows us to account for deviations in isothermal surface excess between proteins and lipids, which are attributable to bonding, functionality, and site configuration. We assume that partial coating of the SSnL is limited to one molecular thickness.<sup>3,116,142</sup> Several literature equilibrium constants are adopted in the form of inverse half saturation carbon atom concentrations, as a first approximation.<sup>117</sup> The expressions we use are derived from typical laboratory adsorption isotherm behavior so that the monolayer can be modeled as accurately as possible. An effective fractional surface coverage relevant to the excess is determined via the relationship:

$$\theta_i = \frac{(1/C_{i,Ref})^{n_i} \times (C_i)^{n_i}}{1 + \sum_{i=1}^2 (1/C_{i,Ref})^{n_i} \times (C_i)^{n_i}}$$

Equation 45

Here  $C_{i,Ref}$  is the half saturation carbon concentration of the  $i^{\text{th}}$  biomacromolecule (1 = proteins and 2 = lipids) and noting as stated above that there is no established literature value of  $C_{i,Ref}$  for carbohydrates. The variable  $n_i$  is an adjustable coverage parameter permitting us to set the isotherm shape to typical experimental adsorption curves for each species, and  $C_i$  is the calculated bulk seawater carbon concentration of the  $i^{\text{th}}$  contributor.

The calculations are performed over the entire global ocean, distributed about every one-half degree latitude and one degree longitude. We assume that local marine biogeochemical dynamics are at steady state on this scale because horizontal diffusivities



are about  $3 \times 10^7 \text{ cm}^2/\text{s}$ .<sup>143</sup> Results are also grouped by Longhurst province to summarize SSnL carbon masses regionally.<sup>114</sup> The nanolayer total is determined using a summation over each calculated value (Equation 46, see Table 5 for definitions). Our calculations encompass biogeochemical diversity using chlorophyll-a modeling, zooplankton concentrations, and grazing rates of the upper ocean, while also presenting integrated amounts in a concise form. The total SSnL mass for each evaluated month is presented in gigatons (Gt) of carbon to provide a convenient reference for comparison to budget figures.

$$\sum_{lat} \sum_{long} \sum_i A_{cell} \Gamma_i \theta_i = M_{C,SSnL}$$

Equation 46

Areas are scaled appropriately to the cosine of latitude so that they decrease toward the poles. Based on the model results and excess maxima for surrogate proteins and lipids, we present SSnL carbon estimates. Parameter values inserted into the described equations are presented in Table 5 with relevant references. Maps are normalized to the largest observed value of carbon overall such that the most intense value is calculated as one.

#### 4.3. Results and Discussion

The carbon SSnL mass is estimated to compare to global budget values, with potential to improve current understanding of how the surface modulates gas transfer, micrometeorology, and global biogeochemistry. Using established relationships between chlorophyll and phytoplanktonic production and consumption, model measurements are converted to dissolved concentrations of carbon for proteins and lipids, the major marine surfactants. Maps of carbon after calculating Equation 44 are presented in Appendix C;

these reflect the calculation of  $C_i$  for lipids and proteins. Fractional surface coverage is also mapped for each month (Appendix C).

We first compare our calculated concentrations of lipids and proteins (from Equation 44) to field measurements of DOC to confirm our model is representative of physical measurements; these values do not take into account surface adsorption and are not representative of the SSnL.<sup>109</sup> Briefly, we sum over  $C_i$  for a given pixel to give us lipid and protein mass and we average over the year to get an estimated carbon concentration. We confirm that our calculated carbon concentration is close to experimentally determined DOC concentrations, indicating that the model methods are sufficient for global analysis, albeit they should be an underestimate as shown in Figure 9. Experimental amounts of DOC are likely divergent from our calculations for two main reasons. Our calculations do not include carbohydrate contributions because of insufficient information about their surface adsorption kinetics, so these calculations have not been included in our model. Second, the dates studied vary. Our model, for example, is for the year 2005. However, the field studies are from a range of years. DOC is dynamic and varies throughout the day, let alone yearly. Despite the variations, our calculated value from the modeled carbon is close to field observations of DOC concentrations, which confirms that our approach of using chl-a is viable for modelling applications.

SSnL adsorption is accounted for through experimental isotherm relationships to calculate the fractional carbon coverage. Total carbon mass of the SSnL for a one-month average is determined to be  $\sim 10^{-4}$  Gt, indicating simultaneously that there is significant amount of carbon, but that amount is small relative to other contributions in the

biogeochemical cycle. For example, most recently in 2020 total CO<sub>2</sub> emissions were estimated at 10.2 Gt carbon and the carbon sink into the ocean was about 3 Gt.<sup>144</sup> Carbon in the SSnL is normalized to its greatest single pixel value;  $\sim 10^7$  g carbon is represented by the value of 1 and bright yellow regions on maps. Results for all months are mapped to emphasize the variability throughout the calendar year (Appendix C).

We observe the seasonal variability more closely in the months of change (Figure 10). Maps for March, June, September, and December emphasize the hemisphere separation and seasonality. In June and September, southern oceans are significantly darker blue/purple, which is indicative of less carbon in the southern hemisphere winter and spring. The equator is defined by consistently high values of carbon, but seasonal variations are still observed, most notably between March and December. May and November 2005 monthly averages provide Northern hemisphere mid-spring and mid-fall references, with six months separation (Figure 11).

Our results indicate that coastlines support higher carbon biomass, which is supported by literature observations reported for field studies.<sup>115</sup> For example, the yellow-green coloring offshore Chile is consistent with well-known eastern basin upwelling and associated biological activity. Similarly, regions of remote open ocean exhibit significantly lower masses in their respective fall seasons, and this can be seen most notably by contrasting May and November shifting central minimum of the South Pacific and North Pacific Gyres. Even equatorial variability and continuity is of significance (Figure 5a). The equator has a more constant yearly temperature and increased upwelling of nutrients

through ocean circulation, which creates ideal conditions for marine biota to thrive and ultimately produce more carbon (bright yellow on maps).

Figure 12 emphasize longitudinal variability and a degree of similarity between the two central sample months. Normalized carbon for spring and fall is overlaid for latitudes of  $0^\circ$  and  $25^\circ$ . Regional overlap of red and black is an indication that SSnL carbon remains consistent over time. Figure 12 highlights the detailed seasonal variability we observe along the equator. Carbon mass is lower in May when viewed along this longitudinal axis. Greater November SSnL carbon is explained by the behavior of global trade winds.

As the adjoining hemispheres move in and out of summer/winter, the carbon mass rises and falls by five percent or more (Figure 12). As summer ends in the two hemispheres, low latitude aquatic productivity increases since there is more upward mixing, resulting in injection of nutrient material. This ultimately contributes to a relatively concentrated SSnL. In the midrange northern hemisphere (above  $40^\circ$  latitude), carbon decreases from May to November and this effect is driven by the standard productivity decreases; seasonal algal blooms do not persist into late fall. This is supported by the higher masses around  $-40^\circ$  (southern hemisphere) in November relative to May.

At  $25^\circ$  north latitude, Figure 12b, we observe greater carbon mass in northern hemispheric spring across most longitudes. November displays lower or unchanged surface carbon between the two sample months. There are few longitudes where northern November outpaces May at this selected mid-latitude; mostly occurring in coastal areas. Our observations confirm regular oceanographic seasonal variations known to modulate the contemporary distribution of primary productivity. For example, as the northern

hemisphere approaches summer a strong upwelling of nutrients becomes stabilized in the mixed layer. Combined with seasonally increased temperature and daylight hours, this change gives rise to plankton blooms and the ecosystem is subsequently enriched with carbon.

Calculated masses were evaluated over the named ecological regions defined by Longhurst in order to evaluate biome diversity, because the sum total global SSnL carbon does not vary seasonally (Figure 13).<sup>114</sup> Regional mass totals summarize month to month variability. Observed carbon decreased in the Northwest Arabian Sea upwelling province (ARAB) from May to November. We also noted a decrease in Western Tropical Atlantic (WTRA) between the selected months. The South Pacific Subtropical Gyre (SPSG) is a relatively biologically quiescent region, but it is vast, so effects are amplified (Appendix C). By contrast, the California current (CCAL) is narrow and restricted but it is subject to coastal nutrient upwelling.

Total SSnL carbon mass was determined by summing over all pixels of non-zero chlorophyll measurements. In both months, the calculated value was  $\sim 10^{-4}$  Gt. Monthly variability is minimal when we integrate across the entire globe since the hemispheres offset one another. We observe only a 4% difference. Therefore, the biogeochemically-driven variability is seasonally symmetrical. Ultimately, we attribute consistent carbon concentrations to adsorption equilibria exerting primary control over the ratio of bulk concentration to SSnL carbon.

Due to local seasonal variability shifting globally, the change in individual biomes does not have a strong impact on the global values. Seasonality reflects the geographic

scope of mixed layer productivity. The SSnL and SSML reservoirs must possess an ecological geography of their own consistent with several recent regional analyses.<sup>115,117</sup> The results assist in understanding the geocycling of surface pressure (reduced interfacial tension), which likely affects micrometeorological phenomena. From these results, we assert that global fluid dynamic parameters, such as the drag coefficient, are likely highly variable because there is significant carbon at the SSnL.

These carbon enrichments are important to consider alongside other reservoirs within the Earth System, due to aerosol composition and boundary layer turbulence through interfacial roughness.<sup>22,24</sup> If the contributions from the SSnL are omitted in determining environmental pathways of carbon, a planetary self-regulatory mechanism is neglected. We assert that molecules in the SSnL may have critical links between carbon, micrometeorology, biogeochemistry, and climate.

#### 4.4. Conclusions

The role of a carbon rich region dividing the ocean and atmosphere remained vague in global studies recently, despite the tendency for carbon-rich molecules to organize at the SSnL. The ocean surface has unique carbon sequestration capability; therefore, we examined distributions and variability for the regional to global masses involved. Chlorophyll and zooplankton data from E3SM were used to calculate an integrated SSnL carbon mass for proxy compounds selected from among natural proteins and lipids. The computations were controlled by parameterizations for phytoplanktonic (primary) and zooplanktonic (secondary) production. We determined that SSnL carbon varies both temporally and geographically; however, the monthly total remains consistent at

approximately  $10^{-4}$  Gt. Emphasis is placed here upon temporal and geographical fluctuations, which are estimated from remotely sensed observations.

The small value we calculate is indicative of a large relative global geocycling impact. The molecular reservoir of the SSnL functions as a direct physical barrier and physicochemical moderator between ocean and atmosphere.<sup>15,104,145,146</sup> For example, the roughness-driven drag coefficient is likely affected by surface pressure, since friction elements are reduced in the presence of amphiphilic species. Winds passing over the ocean “grip” the surface less effectively when the chemical complexity is enriched.<sup>147</sup> The local variability of the SSnL thus results in a wide array of effects on global micrometeorology. The SSnL carbon mass is many orders of magnitude less than present day global anthropogenic CO<sub>2</sub> emissions (about 10 Gt).<sup>96,144</sup> Yet, the thin organic nanolayer<sup>12</sup> appears to control the transfer of organics from ocean to atmosphere through aerosols. In aerosols, the organic molecules may act as ice nucleators or cloud condensation nuclei and will become acidified.<sup>148</sup>

Overall, we conclude that impacts of the carbon enriched SSnL should be more thoroughly investigated. The model presented here relies heavily on simple assumptions regarding the molecular structure of typical surfactants, chlorophyll-a, and the food web concentrations. We know that the non-equilibrium of the ocean departs from our assumptions. A more complete understanding of organic chemistry focused at the SSnL requires that future iterations address, for example, the variability of phytoplankton taxonomy throughout upper layers, along with detailed species-by-species composition. Additionally, use of chlorophyll-a as a proxy for biological activity should be improved

upon, to account for the structural changes that may be relevant during ecological succession. A more complete view of chemical complexity should be adopted and tested by large-scale systems modelers. Comparison of the carbon models presented to global estimations of the SSnL and surface tension reduction should be done in the future to further advance our understanding of the ocean surface chemistry. The interaction of dissolved and adsorbed carbon through planetary scale surface chemistry is of direct relevance to many aspects of evolving contemporary climate, and we are hopeful that this initial examination of mass variation will provide a starting point.



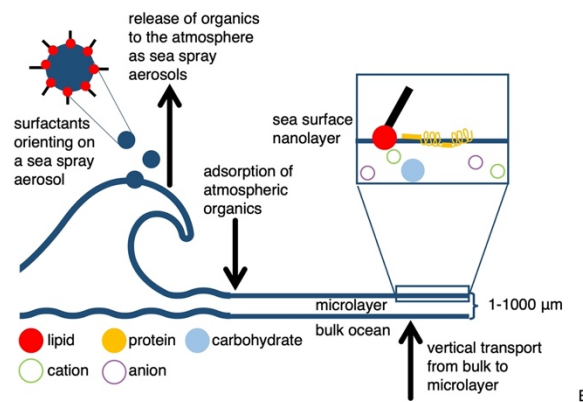
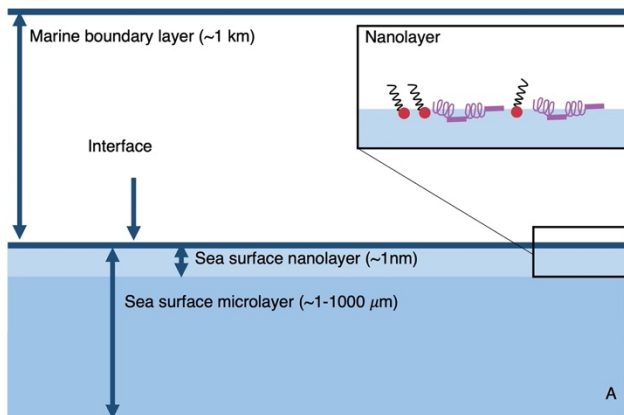


Figure 8. Simplified schematics illustrating (a) relationships between the marine boundary layer (white), sea surface nanolayer (light blue), and sea surface microlayer (dark blue) and (b) highlighting some of the major oceanic processes that occur including vertical transport from the bulk, enrichment of organics at the surface nanolayer, adsorption of atmospheric aerosols and gases, and release of sea spray aerosols from the ocean to the atmosphere.

Table 5. Summary of variables used in calculation of total surface carbon mass including relevant references for literature values.

Variable	Description	Value	Unit	Reference
$i$	Component molecule	1 = protein 2 = lipid	-	-
$C_i$	Carbon atom concentration of the $i^{\text{th}}$ molecule	<i>calculated</i>	$\mu\text{M}$ carbon	-
$g$	Zooplanktonic growth rate	1.0	$\text{d}^{-1}$	Sarmiento 1993 <sup>141</sup>
$C_z$	Concentration of carbon within zooplankton	<i>calculated</i>	$\mu\text{M}$ carbon	Gibson 2020 <sup>132</sup>
$C_p$	Concentration of carbon within plankton	<i>calculated</i>	$\mu\text{M}$ carbon	Gibson 2020 <sup>132</sup>
$K_{inges}$	Half saturation of ingestion	7.0	$\mu\text{M}$ carbon	Sarmiento 1993 <sup>141</sup>
$\gamma$	Assimilation efficiency	0.75	-	Sarmiento 1993 <sup>141</sup>
$p_{i,\%}$	Percentage of macromolecule within the SSnL	$p_{1,\%} = 60$ $p_{2,\%} = 20$	-	Elliott 2019 <sup>117</sup>
$\tau_i$	Lifetime of molecule	$\tau_1 = 10$ $\tau_2 = 2$	d	Ogunro 2015 <sup>118</sup>
$C_{i,Ref}$	Half saturation carbon atom concentration for the $i^{\text{th}}$ molecule	$C_{1,Ref} = 10$ $C_{2,Ref} = 0.5$	$\mu\text{M}$ carbon	Elliott 2019 <sup>117</sup>
$n_i$	Effective shape of adsorption isotherm	$n_1 = 0.5$ $n_2 = 1$	-	Elliott 2019 <sup>117</sup>
$\theta_i$	Fractional surface coverage for the $i^{\text{th}}$ molecule	<i>calculated</i>	-	-
$M_{surf}$	Carbon in the SSnL	<i>calculated</i>	g carbon	-
$A_{pixel}$	Surface area of a pixel	$\sim 10^{10}$	$\text{m}^2$	-
$\Gamma_i$	Maximum surface excess of the $i^{\text{th}}$ molecule	$\Gamma_1 = 2 \times 10^{-3}$ $\Gamma_2 = 2.5 \times 10^{-3}$	$\text{g}/\text{m}^2$	Graham 1979 <sup>106</sup>

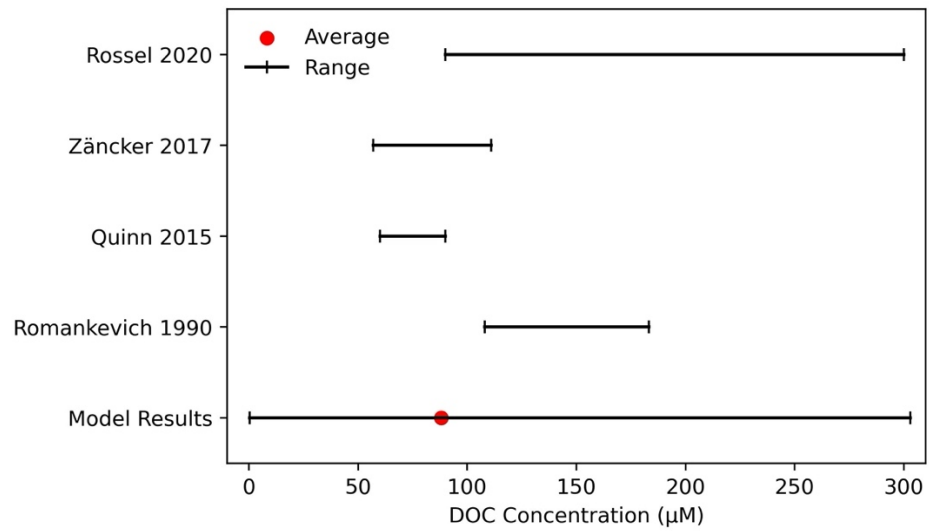


Figure 9. Comparison of four field study DOC concentrations<sup>109,149-151</sup> and the average mass of carbon each month in the region of 0-10 east longitudes and 78-80 north latitudes, which aligns with the regions studied in the Rossel et al., 2020 study.

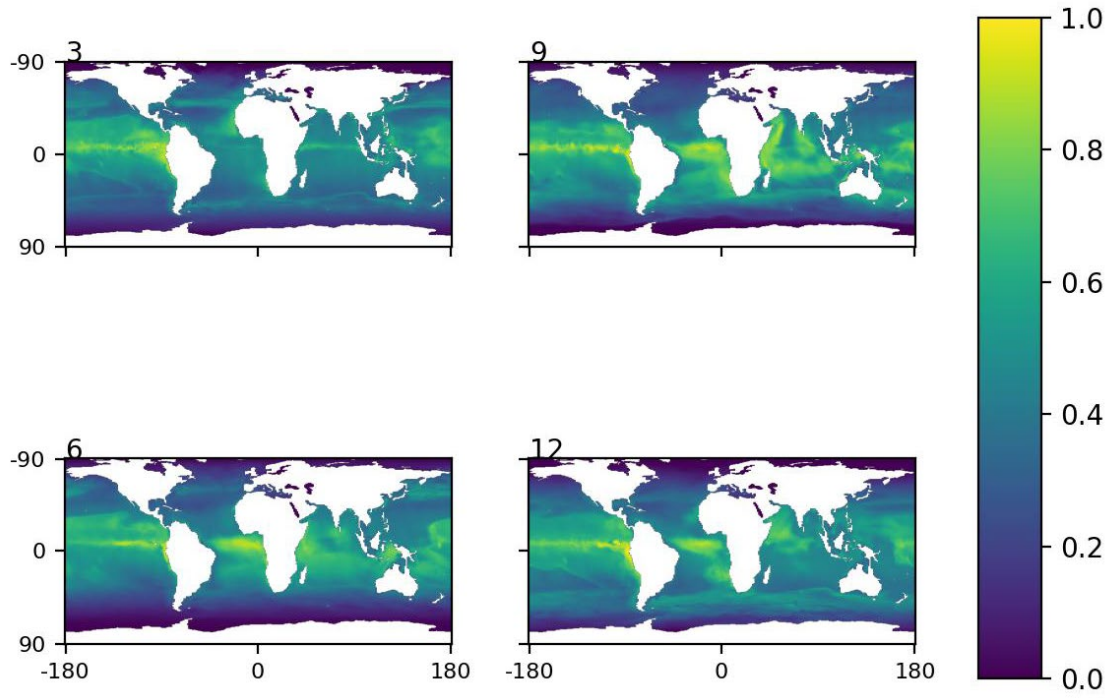


Figure 10. Maps of normalized SSnL carbon for the months of March (3), June (6), September (9), and December (12) from E3SM output for the year 2005 are presented.

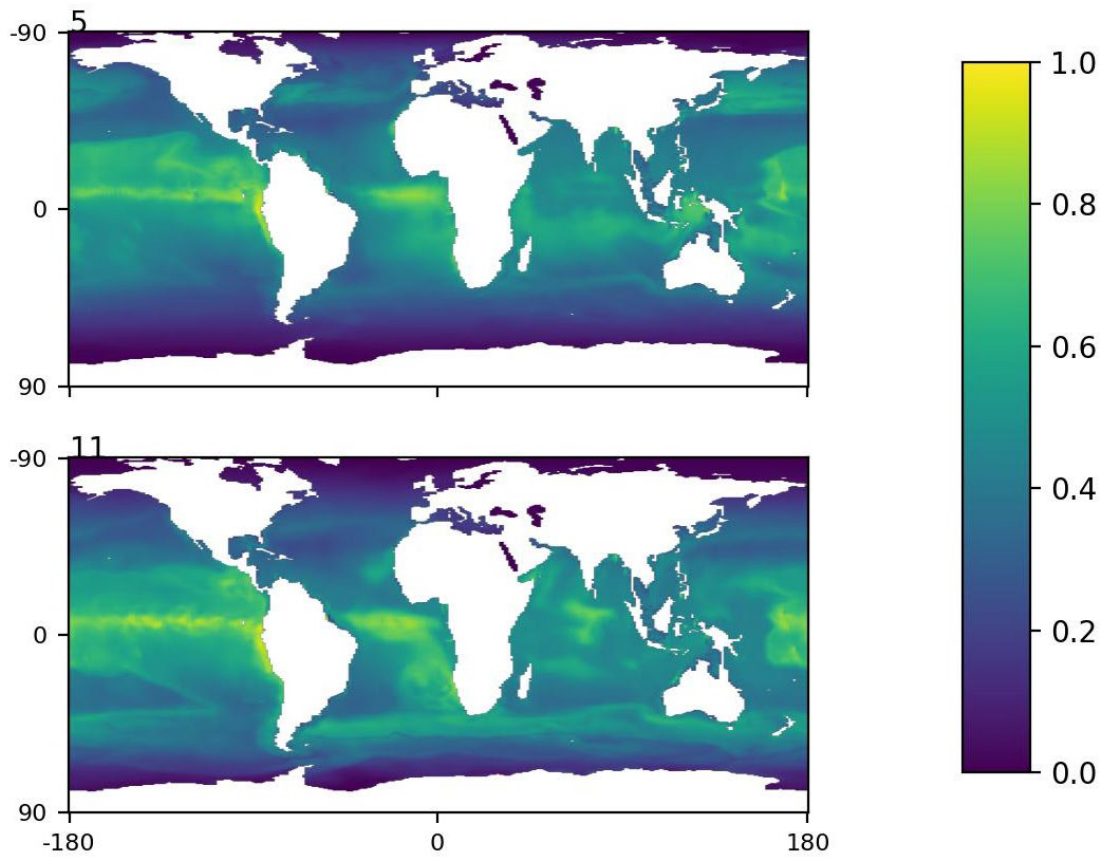


Figure 11. Normalized SSnL carbon for May 2005 (top, '5') and November 2005 (bottom, '11') calculated from E3SM chlorophyll-a and zooplankton output.

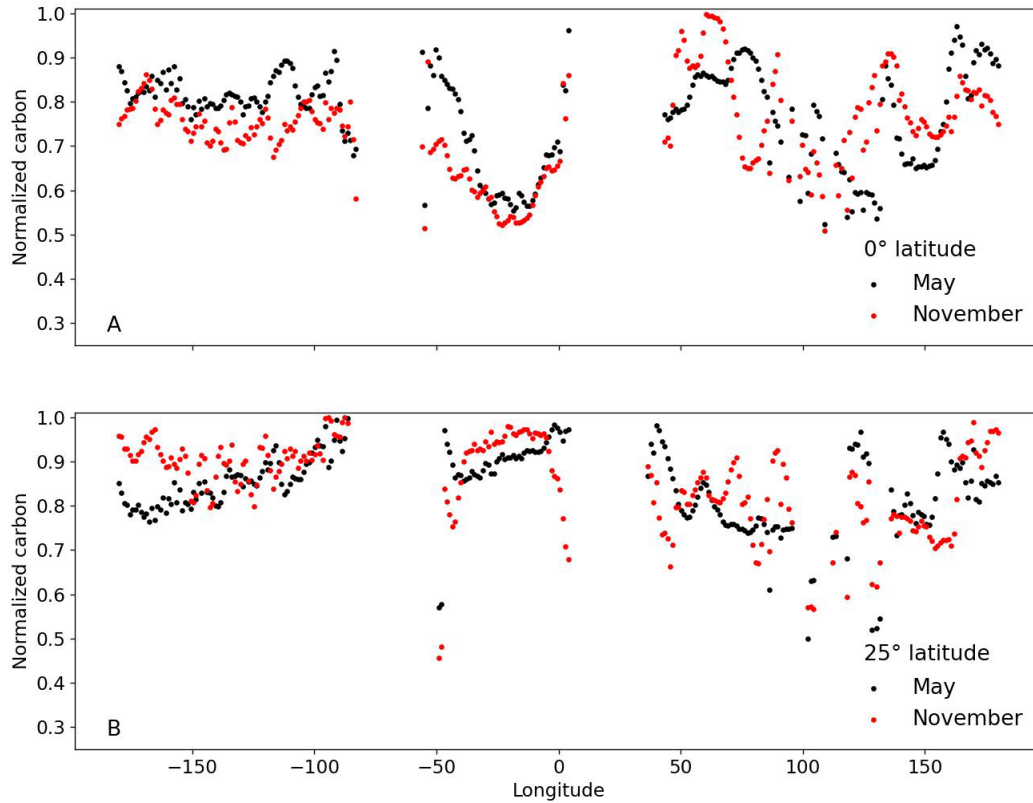


Figure 12. Normalized SSnL carbon across (a) 0° and (b) 25 north latitudes for May and November 2021. Only locations where estimates are greater than zero are included (plots exclude land). Less seasonal variability is observed at the equator in (a), and coastal regions in (b) are well emphasized by the uptick in calculated surface carbon. As we approach land, carbon increases and then decreases in more open ocean regions.

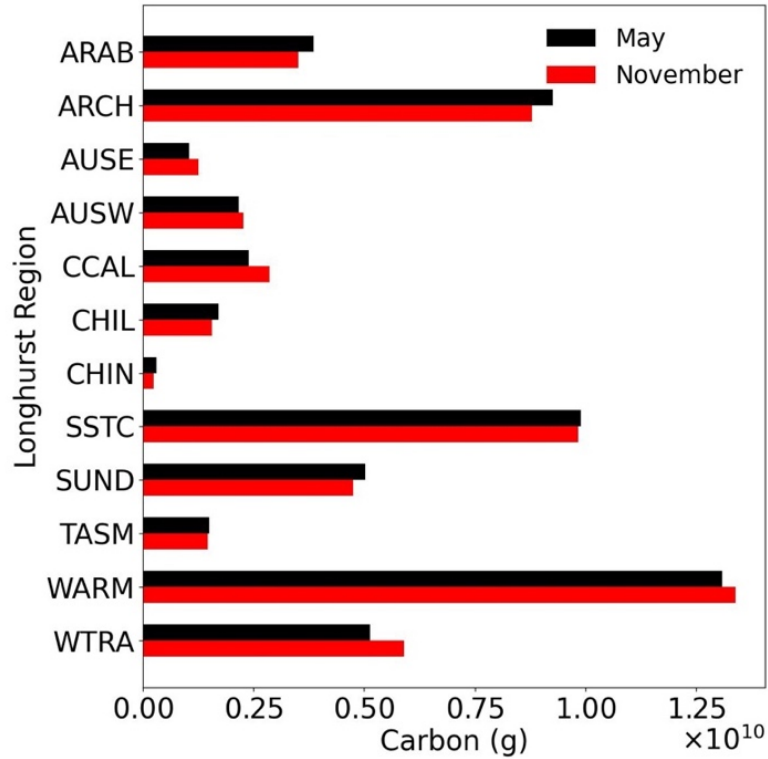


Figure 13. Subset of Longhurst regions with SSnL carbon mass (g) between May and November 2021. Regions are ordered alphabetically by their four-letter standard codes. All values are presented in Appendix C and a subset is discussed in text to underscore key observations. Province acronyms are defined as follows: Northwest Arabian Sea Upwelling (ARAB), Archipelagic Deep Basins (ARCH), East Australian Coastal (AUSE), Western Australian and Indonesian Coast (AUSW), California Current (CCAL), Chile Current Coastal (CHIL), China Sea Coastal (CHIN), South Subtropical Convergence (SSTC), Sunda-Arafura Shelves (SUND), Tasman Sea (TASM), Western Pacific Warm Pool (WARM), and Western Tropical Atlantic (WTRA).

## **Chapter 5. New insights into cation and temperature driven protein adsorption to the air-water interface through infrared reflection studies of bovine serum albumin**

Reproduced in part with permission from Enders, A.A.; Clark, J.B.; Elliott, S. M.; Allen, H.C. “New insights into cation and temperature driven protein adsorption to the air-water interface through infrared reflection studies of bovine serum albumin”. *Langmuir*, resubmitted with revisions, **2023**. Copyright 2023 American Chemical Society.

### 5.1. Introduction

Proteins are an abundant macromolecule in the ocean<sup>9,110,152</sup> and are enriched within the sea surface microlayer (SSML).<sup>23,153</sup> The SSML is operationally defined as the topmost 1-1,000  $\mu\text{m}$  of the ocean, where the atmosphere and ocean have a multitude of transport mechanisms between each other.<sup>12</sup> Proteins and peptides are transferred into the atmosphere via sea spray aerosols (SSAs) that form through wave breaking and bubble bursting at the air-ocean interface.<sup>7,15,153–155</sup> SSAs are known to contain particles that nucleate cloud<sup>16,156</sup> and ice formation.<sup>7</sup> The chemical composition of SSAs is directly controlled by the SSML chemistry.<sup>17,123,150,157,158</sup> However, the processes controlling surface adsorption of proteins are not well understood,<sup>125,153,159</sup> and thus requires further investigation into how surface-active compounds, in general, are promoted to the air-sea interface under variable conditions, including temperature, salinity, and presence of other organic molecules including surfactants. The observations presented herein may improve



climate models by providing insight into the chemical complexity of the ocean and its surface and aid in understanding the biogeochemical processes affecting the ocean surface.<sup>7,10,18,115,117,118</sup>

Protein promotion, orientation, and organization at the air-water interface is well documented in the literature, which largely suggests denaturation as proteins form a monolayer on the water surface.<sup>155,160–162</sup> Yet, knowledge is limited about the effect of physical properties on protein promotion to the air-water interface. Surface adsorption of proteins is a complicated phenomenon because of their tertiary structures<sup>163</sup> which are controlled, in part, by hydrophobicity and hydrophilicity.<sup>164</sup> This likely influences the propensity for monolayer formation by proteins at the air-water interface.<sup>163,164</sup> The structure is also affected by the system's temperature,<sup>165</sup> which is variable across the ocean.<sup>166</sup> Furthermore, the function of a protein may impact the adsorption process. For example, bovine serum albumin (BSA), our chosen proxy-protein, is a transport protein with well-known binding affinity for fatty acids, including stearic acid.<sup>167,168</sup>

BSA is our proxy-protein of choice because of the abundance of previous literature investigations indicating that there is an abundance of proteins and amino acids in the ocean.<sup>169–172</sup> BSA is frequently used in the literature as a proxy protein for surface studies<sup>100,173–175</sup> because BSA captures slow adsorption within the air-ocean surface. While it does not completely capture the nuance of the ocean surface chemistry, the literature results and precedence strongly suggest that it is a suitable proxy for understanding fundamental adsorption changes. Most notably, Jarvis and colleagues presented surface pressure-area isotherms of BSA and upon comparison to ocean slick samples, their

observed surface pressure responses were strikingly similar, producing analogous isotherms.<sup>100</sup> Meinders and coworkers previously reported observations and conformational information of protein surface adsorption, including studies of  $\beta$ -casein and egg serum ovalbumin, to the air-water interface via infrared reflection-absorbance spectroscopy (IRRAS).<sup>176–180</sup> Additionally, the interaction with cations at neutral pH has been documented, but the effect they have on modulating protein surface adsorption varies in the literature.<sup>27,100,125</sup> Langmuir and Waugh presented the first interfacial investigation of albumins, highlighting the irreversibility of films formed by proteins on pure water.<sup>27</sup> More recently, Li and coworkers presented evidence that BSA populates the surface in saline solutions more readily than in pure water.<sup>125</sup> The unknown propensity for BSA to adsorb at the air-water interface yields motivation to probe more deeply the temperature, salinity, and monolayer effects and the relationships to interfacial structure.

The tertiary structure of BSA at the air-water interface was studied through dilational surface rheology by Noskov and colleagues in 2010.<sup>173</sup> They found that the protein underwent a conformational change at the surface where the protein unfolded and elongated along the x-axis, until surface pressures of about 12 mN/m when it began to “loop” into the water. Further work by Yuan and coworkers expanded on understanding the surface behavior of BSA by determining surface excess at variable concentrations of protein and sodium chloride salt.<sup>174</sup> The observed “salting up” and “salting down” effects were a function of protein concentration; low and high concentrations have salting up effects and mid-range concentrations salting down. Ulaganathan and colleagues found that  $\beta$ -lactoglobulin was more readily promoted to the surface in ionic solutions, but that the

effect diminished at higher concentrations (mM scale).<sup>181</sup> These results indicate that there is at first a stabilizing effect that promotes proteins to the air-water interface in ionic solutions and the effect is not linear with ionic strength. The well-studied Hofmeister effect is further evidence of the salting out and salting in effects on proteins.<sup>182–186</sup> Generally, anions have a greater effect than cations, yet the mechanism is still not fully understood. Current works hypothesized the disordering of water structure by the ions is the cause for the observed effect.<sup>187</sup>

BSA surface adsorption at the air-water interface is also affected by surfactants. Noskov and Mikhailovskaya reported on the reduced globule charge density of BSA when sodium dodecyl sulfate (SDS) monolayers were formed at the interface.<sup>188</sup> At very low surfactant concentrations, the surface adsorption of BSA was observed to decrease and the surface tension was more representative of expected values for SDS monolayers. Further surfactant work by Pedraz and colleagues characterized a two-step mechanism in which Langmuir biofilms were formed with arachidic acid and BSA.<sup>175</sup> The process ultimately resulted in co-adsorption, where BSA interacted with the surfactant head groups in the solution phase rather than assembly at the interface.

Global ocean modelling efforts have long used bovine serum albumin and stearic acid as stand ins for ocean organic matter stemming from seminal studies,<sup>100,115,132,189</sup> and thus our proxy system implements these molecules. However, our experimental design is ultimately not comprehensive or exhaustive in its inclusion of all oceanic and atmospheric biological, chemical, or physical components. For example, our experiments exclude the influence of microbial communities,<sup>55,190–193</sup> the diverse and complex array of molecules

adsorbed from biogenic<sup>15,191,194</sup> or anthropogenic sources,<sup>9</sup> and wind dynamics<sup>195</sup> that are destabilized by decreased surface tension from adsorbed organic monolayers. Instead, the inclusion of salinity and temperature are used to understand the fundamental process of protein surface adsorption. Our results provide experimental evidence about these factors, which can be used to guide models that rely on laboratory experiments. Importantly, our results are not all encompassing of the oceanic and atmospheric non-equilibrium systems.

The investigation presented herein requires the assumption of limited contributions from biological, chemical, or physical influence on the air-water interface. Our assumption is a limitation of the study and requires additional experimental parameterization be included in future investigations. Despite this, the use of a proxy that mimics some surface measurements (e.g., surface tension<sup>100</sup>) and conditions of the ocean provides necessary insight into the sparsely measurable ocean surface. Our results provide insight into the effect of temperature and ionic strength on surface adsorption of the ocean through the proxies of BSA and stearic acid.

We present an investigation into the dynamic air-aqueous interfacial adsorption process that occurs when BSA is introduced into an aqueous system. To our knowledge, this is the first study that evaluates the impact on adsorption as a function of amide peak intensity with varying solution ionic strength, temperature, and adsorbed monolayer. We examine the surface structure and its changes through IRRAS measurements, which enables surface-sensitive characterization of the interfacial chemical composition. Our approach builds from the simplest system of BSA injected into pure water at 20°C and we compare the observed changes relative to this system. We extend our fundamental system

to include ionic strength (e.g., artificial sea water (ASW) solution) and temperature (e.g., surface temperature of 10°C), which is relevant to ocean systems. Ultimately, our findings provide insight into the fundamental chemical and physical properties affecting surface adsorption.

## 5.2. Methods

### 5.2.1. Materials and Sample Preparation

Aqueous solutions were made using ultrapure water (MilliQ Advantage A10, resistivity 18.2 MΩ). All materials were used as received except for sodium chloride (Fisher Chemical, ≥ 99%, certified ACS) which was baked at 600°C for at least 10 hours to remove residual impurities and used to make 0.45 M NaCl in water.<sup>196</sup> About 35.5 g of Instant Ocean salt (ion concentrations from manufacturer reported in Appendix E) was used to make 1 L solutions of artificial sea water (ASW) based on label recommendations. The ASW solution has 19,290 ppm Cl<sup>-</sup>, 10,780 ppm Na<sup>+</sup>, 2,660 ppm SO<sub>4</sub><sup>2-</sup>, 1,320 ppm Mg<sup>2+</sup>, 420 K<sup>+</sup>, and 400 ppm Ca<sup>2+</sup>, and all other included ions or elements are less than 400 ppm. A 1 mM BSA (Sigma Aldrich, ≥ 98%, heat shock fraction, pH 7) solution in water was prepared for injection. BSA solutions of concentrations 1, 50, 100, 250, 500, and 750 μM were also prepared for analysis via ATR-FTIR and IRRAS. Stearic acid (Sigma Aldrich, ≥ 98.5%, capillary GC) was dissolved in chloroform (ACROS Organics, ≥ 99.8%, ACS Reagent) to make a 3 mM solution for spreading. Solutions were prepared the day before measurements, stored in a glass Pyrex container, and equilibrated to lab conditions overnight.

### 5.2.2. Infrared Spectroscopy

A PerkinElmer micro-ATR assembly (one bounce, diamond/KRS-5) and PerkinElmer FTIR (Spectrum 3, DTGS detector, and KBr windows) was used to obtain bulk measurements of the BSA solutions to confirm amide peak assignment. Infrared Reflection-Absorbance Spectroscopy (IRRAS) spectra were collected with a modified PerkinElmer Frontier FT-IR Spectrometer using a custom, lab-built reflection system with two gold mirrors (Figure 14). The sampling stage is contained within the sample compartment of the commercial instrument. After the IR source is modulated by the interferometer, it is directed to the sample compartment. The IR beam is incident on the first gold mirror which is angled such that the light is directed toward the sample surface at 48° relative to surface normal. The light reflected off the sample surface is collected using a second gold mirror and returned to the instrument and directed to a liquid-nitrogen cooled HgCdTe (MCT) detector.

Surface-sensitive spectra are obtained by calculating reflectance-absorbance (RA) which is given as  $RA = -\log\left(\frac{R_M}{R_0}\right)$ , where  $R_M$  is the reflectivity of the sample surface and  $R_0$  is the reflectivity of the reference surface (background). As a result of the mathematical relationship, positive and negative vibrational modes are observed in the spectra. Negative RA response occurs when the sample surface reflectance is greater than the reference reflectance ( $R_M/R_0 > 1$ ). For backgrounds with reflectance greater than the sample ( $R_M/R_0 < 1$ ), positive RA responses are observed.<sup>197</sup> Measurements were taken using unpolarized light in the single-beam mode and averaged over 400 scans at a resolution of

4  $\text{cm}^{-1}$ . IR response was recorded from 4000-450  $\text{cm}^{-1}$  at every 0.5  $\text{cm}^{-1}$ . Samples were collected using a KSV NIMA Langmuir trough (~135 mL total volume) equipped with a surface tensiometer.

Solutions at 20°C were added to the trough and equilibrated for 10 minutes prior to acquiring a background measurement. In lower temperature studies, a ThermoFisher Recirculating Chiller was used with a 60:40 water and ethylene glycol solution. The equilibration time was increased to 30 minutes to ensure complete cooling. For trials with fatty acid monolayers, 18  $\mu\text{L}$  of 3 mM stearic acid in chloroform was spread dropwise onto the surface using a Hamilton gas-tight syringe and 10 minutes were allowed for solvent evaporation prior to measurement. To all solutions, 50  $\mu\text{L}$  of 1 mM bovine serum albumin in MilliQ  $\text{H}_2\text{O}$  was injected, which resulted in a final concentration of 0.37  $\mu\text{M}$ . Measurements were taken immediately after injection. All data was taken in triplicate. Spectra presented were analyzed using custom Python codes, which included conversion from single beam to RA, linear background subtraction, and averaging. Experiments performed are summarized in Table 1. The concentration dependence of BSA surface adsorption was evaluated via IRRAS measurements after injection of a series of BSA concentrations at 20°C in  $\text{H}_2\text{O}$  (Appendix E). The calculated value corresponds with the weight percent that has stable surface pressure observed in work by Graham and Phillips in 1979.

### 5.3. Results and Discussion

We examine protein adsorption to the surface as a function of temperature, ionic strength, and surface structure (preexisting monolayer). The amount of BSA adsorbed to

the aqueous surface corresponds to the intensity of the RA response. Our results indicate that BSA is surface active under all conditions, however the structure of the surface and amount that adsorbs is variable as conditions of the system are modified. Briefly, we evaluate temperature change in the system to understand the thermodynamic dependence of protein adsorption. In addition, we determine if ionic strength or presence of a monolayer (stearic acid) enables co-adsorption to the surface. The results provided insight into surface adsorption under variable conditions. Ultimately, these experimental data improve the accuracy of ocean-relevant computational models<sup>117,137,189,198</sup> that rely on fundamental laboratory experiments for input parameters. Specifically, our results provide increased understanding of what promotes organics to the surface and the variation in adsorption that ultimately affects SSA organic fractionation.

We first evaluated solution-phase measurements of BSA to confirm amide peaks for assignment (Figure 15). Amide I is assigned to  $1660\text{ cm}^{-1}$  ( $\nu_{\text{C=O}}$ ) and amide II to  $1540\text{ cm}^{-1}$  ( $\delta_{\text{N-H}}$ ).<sup>199,200</sup> We also note that the integrated absorbance is linear with increasing concentration and provide details regarding the nature of ATR as a bulk phase measurement with variable path length (Appendix E). Additionally, we ensure the ions in each solution does not significantly affect the O-H stretching or bending modes (Appendix E).

### 5.3.1. Solution Effect

The solution effect on surface adsorption is observed at  $20^{\circ}\text{C}$  (Figure 16a). The standard deviation of amide I peak intensities are provided in the SI. Surface adsorption is observed when BSA is injected into all three systems, however the intensity varies as a



function of ionic composition. At ocean relevant sodium chloride concentration (0.45 M), adsorption is increased resulting in more intense negative bands ( $1660\text{ cm}^{-1}$  and  $1540\text{ cm}^{-1}$ ). The addition of divalent cations in ASW results in a relatively small increase in intensity as compared to the NaCl solution consistent with the greater concentration of sodium relative to magnesium and calcium cation concentrations. We observe the intensity of the amide I and amide II bands increase with increasing ionic composition ( $\text{H}_2\text{O} < \text{NaCl} < \text{ASW}$ ), which is consistent with observations described in the literature.<sup>174,181</sup>

### 5.3.2. Temperature Effect

A decrease in the solution temperature resulted in decreased peak intensity in both amide bands (Figure 16c). We observe a slightly less intense amide I band at  $1660\text{ cm}^{-1}$  from carbonyl stretching ( $\nu_{\text{C=O}}$ ) at  $20^\circ\text{C}$  for BSA in pure water. This slight change could be attributed to strengthening of  $\text{H}_2\text{O}$ -BSA hydrogen bonds as the temperature is decreased.<sup>200</sup> Between  $20^\circ\text{C}$  and  $10^\circ\text{C}$  for the water solution, the amide I and II peak intensities have only a  $\sim 1\%$  difference. As shown in Figure 2c, the temperature effect on surface adsorption in pure water is minimal yet has a greater impact on ionic solutions. We observe a decrease in the  $1660\text{ cm}^{-1}$   $\nu_{\text{C=O}}$  peak intensity at  $10^\circ\text{C}$  compared to  $20^\circ\text{C}$  for both 0.45 M NaCl and ASW. The percent difference between peak intensities at  $20^\circ\text{C}$  and  $10^\circ\text{C}$  for both 0.45 M NaCl and ASW are  $-5.5\%$  and  $-5\%$ , respectively.

Compared with ultrapure water at  $20^\circ\text{C}$ , a solution with sodium chloride greatly increases the intensity of amide I and II bands (Figure 16a). Similar trends of increased intensity in ionic solutions are observed at  $10^\circ\text{C}$ , but the overall intensity is decreased

(Figure 16b), demonstrating that the cation stabilization of BSA at the interface is disrupted by the removal of thermal energy from the system. Our observation indicates that the addition of monovalent cations increases the propensity for protein surface adsorption in direct alignment with the “salting-out” effect that is described by the Hofmeister series.<sup>125,183,184</sup> However, the ionic interactions and interfacial surface structure is destabilized at decreased temperatures. The effect of the solutions is further emphasized in Figure 17, where the area under the curve is analyzed for both amide bands. We assume constant transition moment dipole strength based on our ATR analysis (Appendix E); the observed intensity corresponds to surface adsorption as a result.

BSA is negatively charged under neutral and alkaline conditions and appears to be stabilized at the surface by the sodium cations, in 0.45 M NaCl solution, at neutral pH.<sup>125,174,181</sup> Increased intensity in the amide bands is attributed to the ionic strength of the solution. At 10°C, the increase in peak area is consistent with increasing ionic strength; from NaCl to ASW, the increase is relatively small. However, at 20°C, the increase is much larger between the two salt solutions. Also of note is the relatively equal peak area observed between the two temperatures in the sodium chloride solution. In the difference spectra we observe an amide I band consistent with a change in the adsorption at varying temperatures (Figure 16c). From previous literature studies, we assert this observation is occurring as a result of changing tertiary protein structure, including unfolding or denaturing, leading to a change in the observed intensity.<sup>201</sup> We conclude that the surface adsorption of BSA is affected by both temperature and ionic strength and that there is an observed synergistic effect.

### 5.3.3. Monolayer Effect

The presence of a stearic acid monolayer ( $\sim 45 \text{ \AA}^2/\text{molecule}$ , ‘gaseous’ phase) changes the surface adsorption of BSA at  $20^\circ\text{C}$  compared to the system with no monolayer at the same temperature (Figure 18a). The surface-IR spectra of the BSA/water system at the air-water interface with and without a surface adsorbed monolayer show minimal changes in peak intensity and shape. The presence of a surface adsorbed monolayer results in red-shifting of spectral peaks belonging to both the stearic acid monolayer and proxy-protein, BSA. We observe a red-shift in the  $\nu_{\text{C=O}}$  of  $2.5 \text{ cm}^{-1}$  (convolution of amide from BSA and stearic acid head group) and a  $5 \text{ cm}^{-1}$  red-shift in the  $\delta_{\text{N-H}}$  mode, belonging to BSA, when a stearic acid monolayer is present. Vibrational shifts are generally attributed to changes in the intermolecular interactions (e.g., hydrogen bonding or ion-dipole interactions) or molecular environment (e.g., reduced thermal energy). The IRRAS spectra presented herein are surface sensitive therefore such shifts indicate stronger interactions are occurring at the surface when BSA and a stearic acid monolayer is present. This is indicative of greater surface organization, which is established in the literature through water surface measurements using Brewster angle microscopy (BAM)<sup>6,191</sup> and surface pressure-area Langmuir isotherms.<sup>106,127,197,202</sup>

The observed wavenumber shifts are not necessarily related to tertiary protein structure conformation change, but to a change in the water surface structure at the air-water interface. When a stearic acid monolayer is present, our results are consistent with previous results presented in the literature for greater surface organization.<sup>180,203</sup> The effect is reversed in a  $0.45 \text{ M NaCl}$  solution at  $20^\circ\text{C}$  and we observe a blue shift in each peak,

indicating the surface becomes more disordered. This is likely from cationic destabilization and disruption of the stearic acid headgroup hydrogen bonding with water molecules. For example, cations will interact with stearic acid at the interface through ion-dipole interactions that include ionic bonds, bridging, and chelation.<sup>20,197,202,204</sup> Yet the stronger hydrogen bonding network achievable in non-ionic solutions is still disrupted in comparison. We note a 7% difference between ASW with and without a fatty acid monolayer, where the lipid monolayer results in less intense amide bands.

The surface structure and IRRAS response becomes more complicated when the system is cooled to 10°C and a monolayer is spread (Figure 18). We observe similar  $\nu_{\text{C=O}}$  and  $\delta_{\text{N-H}}$  modes in water with similar intensities to 20°C with and without a monolayer. The 0.45 M NaCl solution has a much less intense  $\delta_{\text{N-H}}$  mode and positive carbonyl stretch, which indicates that the protein is below the surface.<sup>130</sup> We also observe a stronger response in the ASW solution compared to 0.45 M NaCl at 10°C; the  $\nu_{\text{C=O}}$  mode is positive and  $\delta_{\text{N-H}}$  is more intense. As noted above, the positive and negative peaks originate from the mathematical conversion from single beam data to RA spectra.<sup>130</sup> Therefore, positive peaks are indicative of a greater reflectance in the background.

The observed derivative-like peak shape (Figure 18b) has been previously observed in the literature.<sup>176,203,205</sup> Meinders and colleagues conducted a comprehensive examination of the IR response and specifically attribute the abnormal features to external reflection optical effects, not tertiary protein structure variations.<sup>206</sup> In general, the peak shape is interpreted as a convolution of the bending mode of water (1650  $\text{cm}^{-1}$  in neat water<sup>207</sup>) from the reference signal ( $R_0$ ) and perturbed O-H bending mode and amide I vibrational

contributions from the sample ( $R_M$ ). The resulting peak shape has an indeterminate origin. We interpret the spectral response to indicate a complex interfacial environment where equilibrated water organization is disrupted by proteins being promoted to the surface.

From our experimental results, we assert that the BSA surface adsorption is modulated by salinity and temperature; this result is supported by literature.<sup>160,179,208,209</sup> The changes that occur are not simple or independent of other system variables. Ionic strength, temperature, and lipid monolayers have compounding effects on the observed IR response. The least complex system is ultrapure water; amide I band intensity varies only slightly when temperature and monolayers are considered. Red shifting is observed, as noted, when the stearic acid monolayer is spread and when the temperature is reduced to 10°C. Thermal energy reduction of the system results in this observed wavenumber shift, which is visualized in Figure 19. While only a small change, observing the thermodynamic effect in water indicates that the temperature does alter the system and affects how surface adsorption and organization occurs.

The observed results are summarized in Figure 20. Overall, we observe smaller negative intensities on the water and have more IR response in ionic solutions (more negative RA). The temperature dependence is evident when comparing 20°C and 10°C; decreasing temperature decreases the amount of BSA adsorbing to the surface. Our results indicate that temperature affects the surface structure: decreasing the system's temperature decreases the protein adsorption to the surface as reflected in the lower IR response.

#### 5.3.4. Application to Climate Models

We assert that variable temperatures over the ocean surface must influence adsorption to the air-ocean interface and the effect should be considered in relevant ocean and climate models. Organic Compounds from Ecosystems to Aerosols: Natural Films and Interfaces via Langmuir Molecular Surfactants (OCEANFILMS),<sup>189</sup> Energy Exascale Earth Systems Model (E3SM) research and development,<sup>115,132</sup> and offline simulations of macromolecule surface activity<sup>118</sup> among other models<sup>117,210</sup> exclude or assume constant temperature. Our experimental results indicate that temperature does influence surface adsorption. It follows that decreased organic enrichment would result in SSAs formed with decreased enrichment of ice nucleating particles or cloud condensation nuclei. Other factors convolute ocean surface adsorption; for example, algal blooms and their subsequent senescence result in injection of organic molecules into the water column and enrichment of the surface layer.<sup>191</sup> While not presently investigated, their effect is nonetheless present and impacting SSA chemistry.

The methodology of our model system probing the proxy-protein BSA surface adsorption should be expanded to investigate temperature effects on more complex systems. In general, including temperature in models, such as measurements of surface temperature made by satellite imagery (e.g., NASA MODIS), would likely result in more relevant model output, especially for monolayer and surface chemistry modeling. As previously stated, our results are not exhaustive, but instead a framework for further investigating the role of temperature on surface adsorption.

#### 5.4. Conclusions

Herein, we have presented evidence supporting the surface activity and adsorption process of BSA to the interface under varying ionic strength and surface conditions. The amount of adsorbed BSA is dependent on the interfacial chemistry and solution ionic strength at the time it is injected into the system. We determine that even in pure water, BSA is surface active. However, ions promote more BSA to the surface; an observation consistent with results previously presented in the literature.<sup>125,168</sup> Importantly, our results exhibit a cooperative effect of the monovalent and divalent ions from artificial sea water and increasing temperature promote greater protein surface adsorption. We also conclude that BSA is surface active and has a dynamic process through which it adsorbs to the surface that is relevant to proteins in the ocean adsorbing to the surface under variable conditions of fatty acid films and variable temperatures. Ocean relevant concentrations of monovalent and divalent cations facilitate and enhance the surface adsorption of BSA alongside increasing temperature. Stearic acid molecules adsorbed at the air-water interface constricts the adsorption process of BSA, further complicating the surface.

The results presented provide a more nuanced understanding of the effects that ocean conditions have on the SSML and interfacial structure. It is necessary to acknowledge that observations drawn from our results are dissimilar to the ocean because the ocean is a non-equilibrium system. Future work should aid in determining the effect that the ocean's dynamic nature has on the interfacial chemistry. Additional studies exploring the tertiary protein structural modifications at the surface under these conditions of variable temperature and ionic strength would provide interfacial insights. Ultimately,

greater understanding of the effects physical parameters have on adsorption of molecules to the air-water interface will provide improved global ocean and climate modeling.



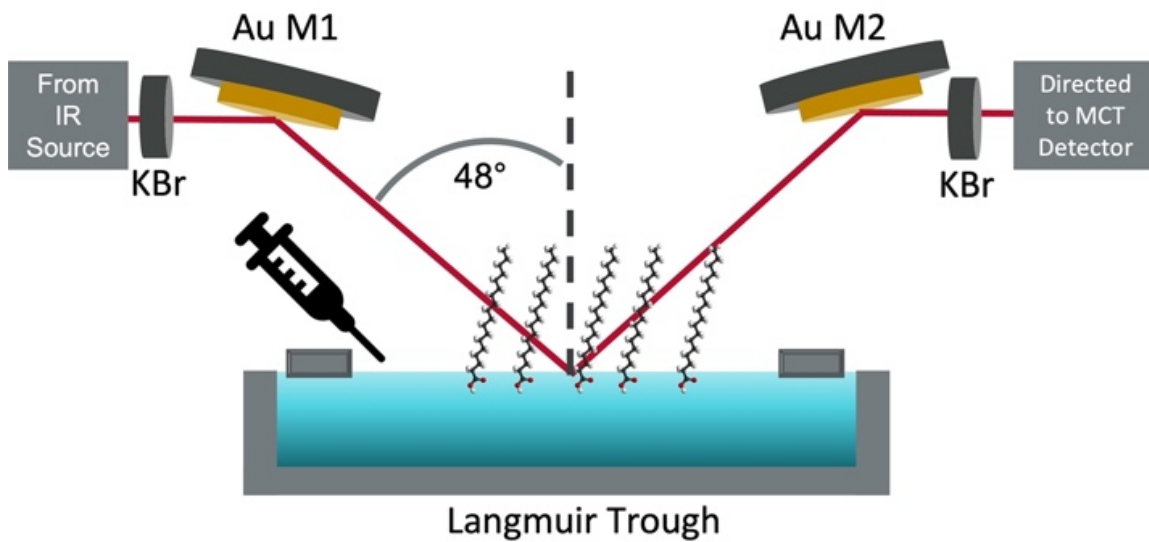


Figure 14. Experimental design of IRRAS assembly combined with a Langmuir trough and temperature control via a recirculating chiller.

Table 6. Three different ionic strengths were used to evaluate the adsorption of BSA to the surface with and without a competing stearic acid monolayer at both 10° and 20°C. Experiments outlined here are for IRRAS measurements.

MilliQ H <sub>2</sub> O				0.45 M NaCl				Instant Ocean			
No Stearic Acid		Stearic Acid		No Stearic Acid		Stearic Acid		No Stearic Acid		Stearic Acid	
10°C	20°C	10°C	20°C	10°C	20°C	10°C	20°C	10°C	20°C	10°C	20°C

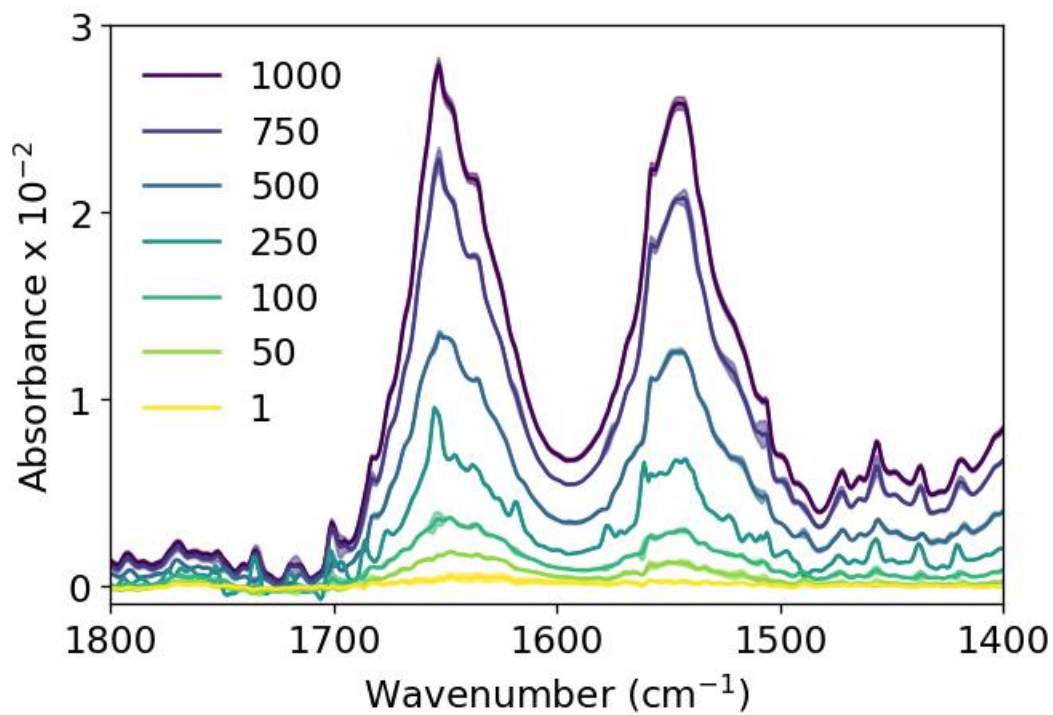


Figure 15. FTIR of BSA in water at increasing concentrations measuring solution-phase concentrations starting from 1  $\mu\text{M}$  to 1000  $\mu\text{M}$  shown in the amide region, presented with error of one standard deviation. The value of the standard deviation is so small it is approximately the thickness of the line of each spectrum. The sharp features likely result from gas phase water present in ambient conditions.

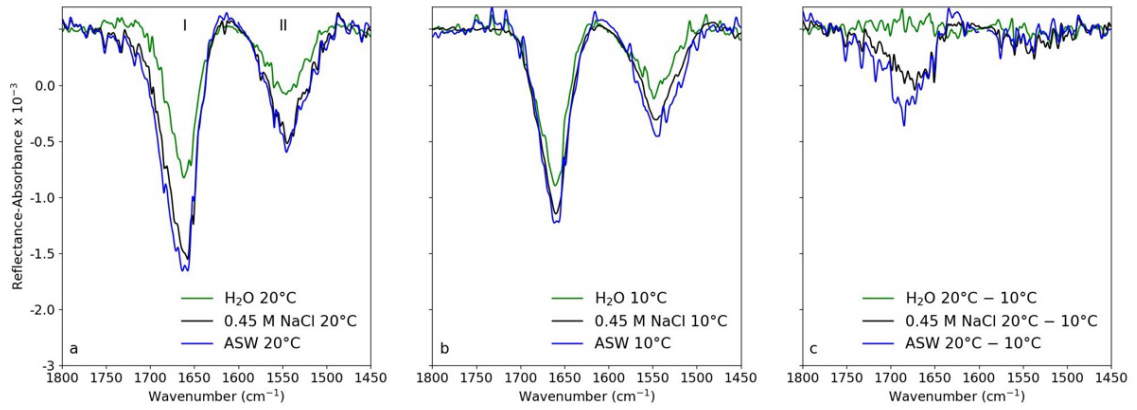


Figure 16a-c. Surface-IR responses in amide region (1800-1450  $\text{cm}^{-1}$ ) after bovine serum albumin (BSA) injection into the H<sub>2</sub>O, 0.45 M NaCl, and the artificial seawater solutions at **a)** 20°C, **b)** 10°C, **c)** difference spectra of 20°C – 10°C. Here, reflectance-absorbance bands are observed as negative peaks.

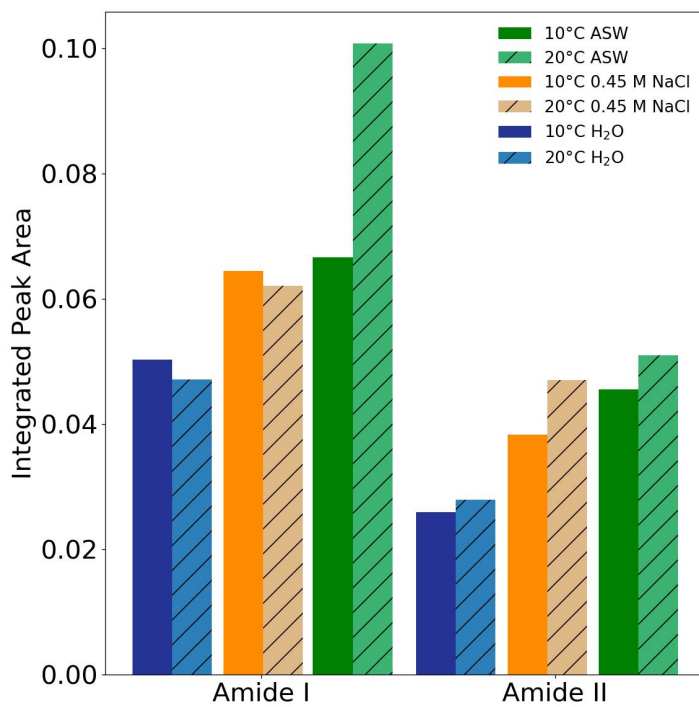


Figure 17. Integrated peak area for amide-I ( $1660\text{ cm}^{-1}$ ) and amide-II ( $1540\text{ cm}^{-1}$ ) bands at  $10^\circ\text{C}$  (dark, solid) and  $20^\circ\text{C}$  (light, diagonal lines). As noted, the peak area and intensity is a result of surface adsorbed protein as confirmed through bulk measurements via ATR-FTIR.

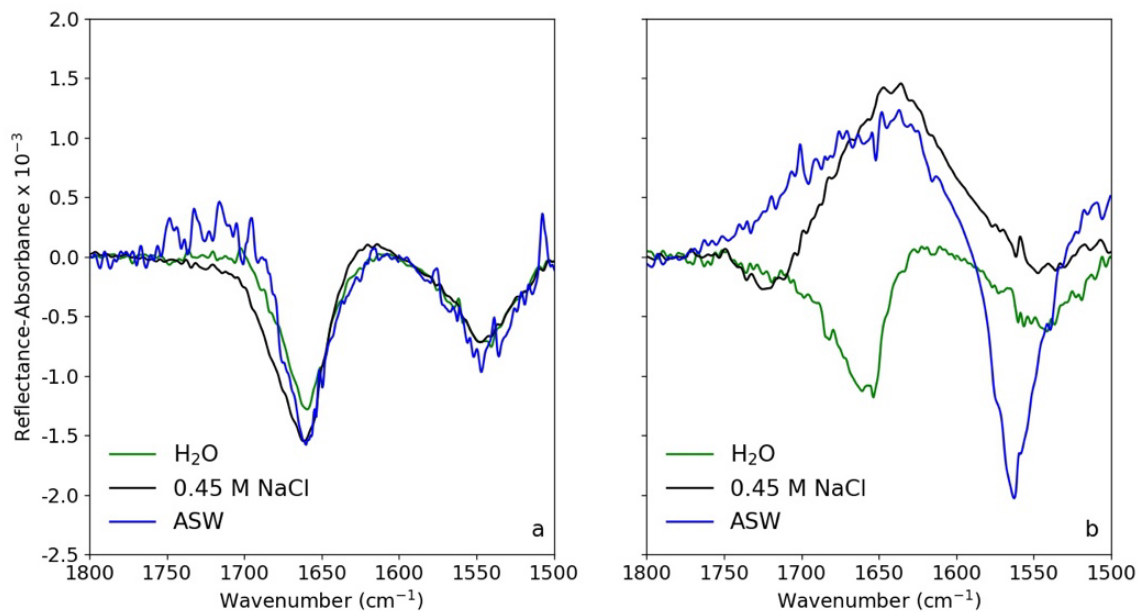


Figure 18a-b. IR response in amide region (1800-1500  $\text{cm}^{-1}$ ) after bovine serum albumin (BSA) injection into each solution with a stearic acid monolayer ( $\sim 45 \text{ \AA}^2/\text{molecule}$ ) on the surface at a) 20°C and b) 10°C.

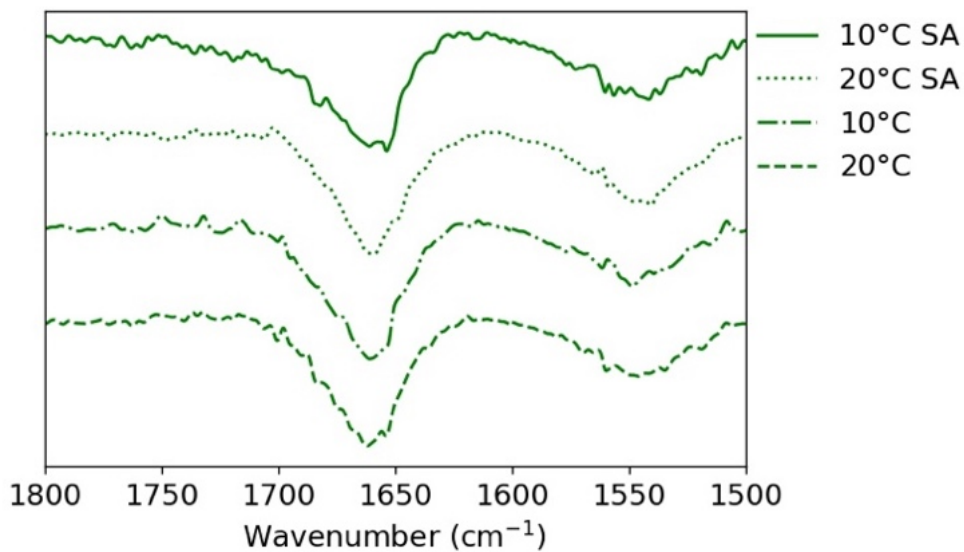


Figure 19. Spectra of ultrapure water surface after injection of bovine serum albumin (BSA) at variable temperature with and without stearic acid monolayer to emphasize the minimal intensity change in the water system. Spectra are offset horizontally for clarity.

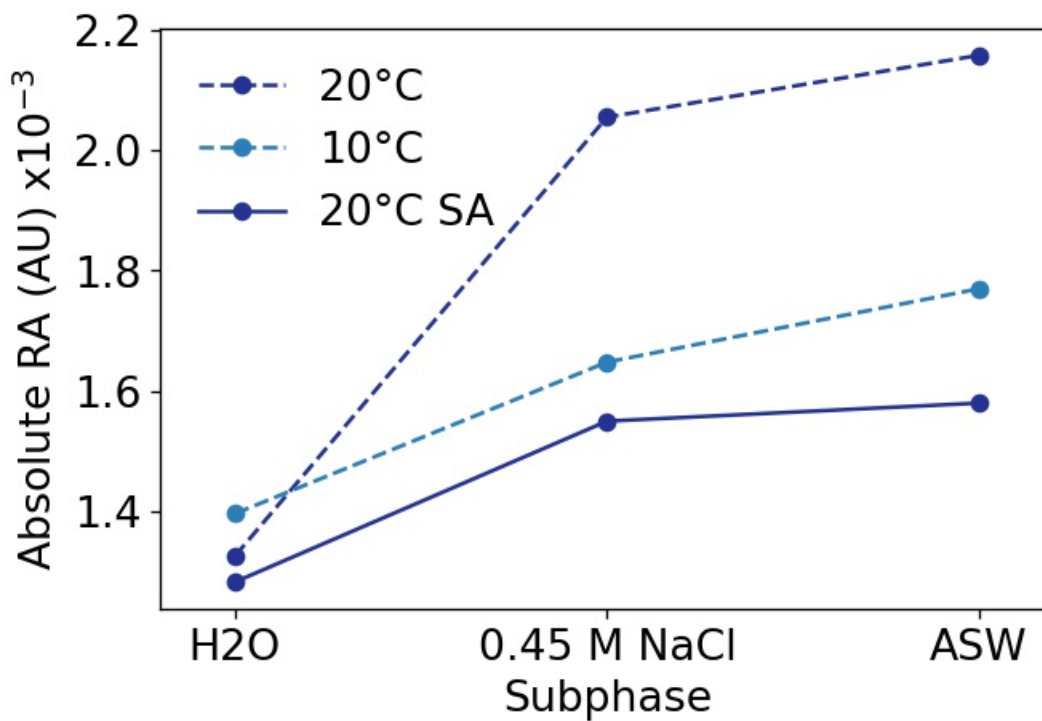


Figure 20. Observed maximum intensity of amide I band (1660 cm<sup>-1</sup>) for each solution at varying temperatures. IR responses from each solution with stearic acid monolayer at 10°C are not included due to the variability in the amide I and II region.



## **Chapter 6. Saccharide Concentration Prediction from Proxy-Sea Surface Microlayer Samples Analyzed via ATR-FTIR Spectroscopy and Quantitative Machine Learning**

### 6.1. Introduction

The sea surface microlayer (SSML) is a multifaceted, deeply complex region of the ocean.<sup>8,12,145,146,157,211,212</sup> As the interface between the Earth's atmosphere and ocean, the SSML performs vital functions that affect climate<sup>108,146,166,202</sup> and ice formation.<sup>13–15,211</sup> Because of unique interfacial anisotropy,<sup>130,197,213,214</sup> 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) that are all components of dissolved organic carbon (DOC).<sup>21,112,125,215,216</sup> Understanding the chemical composition of the SSML provides insight into the biological activity and productivity within the SSML and enables predictions of cloud condensation<sup>16</sup> or ice nucleation,<sup>211</sup> ultimately aiding climatological models.<sup>10,115,117,118</sup> Recent analyses of saccharide concentrations in SSML have shown that a concentration of about 500 nM from eight unique compounds is observed.<sup>21</sup> The dynamic nature and chemical complexity of the SSML make monitoring the region equally more difficult and more necessary.

Our work is motivated by the need for fast, accurate analysis of SSML samples to establish a method that enables exponentially more SSML chemical measurements. Current methods to analyze SSML samples are limited to mass spectrometry,<sup>122,146,217</sup>

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 attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectra to estimate the saccharide concentration via machine learning (ML) implementations. ATR-FTIR spectroscopy also provides concentration and chemical composition, although we note lower detection limits are well known for IR methods as opposed to the high sensitivity for mass spectrometry. Rather than mass separation, IR probes bond vibrational responses at specific wavenumbers.<sup>218</sup>

ML provides a unique avenue to explore relationships among data that cannot be otherwise deduced and the applications to improve or expand chemical systems are broad and present throughout all chemistry fields. Materials design,<sup>219,220</sup> novel drug discovery,<sup>221,222</sup> catalyst optimization,<sup>223,224</sup> and clean energy production<sup>225,226</sup> are some of the many fields where knowledge has expanded because of ML. 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.<sup>227,228</sup> Takamura and colleagues explored methods to identify donor biological sex from urine samples.<sup>87</sup> 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 machines (SVM) in predicting brain cancer from ATR-FTIR spectra.<sup>229</sup> Their high-throughput approach featured high sensitivity and specificity in the prediction of benign versus malignant samples.

SVMs 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), multi-layer perceptron (MLP), and random forest (RF) models.<sup>230</sup> These literature examples highlight the diverse applications of ML and develop techniques that expand the applications of chemistry, as we present herein.

We chose ML methods of increasing complexity to evaluate the training data and investigate new data, including field samples with unknown composition. More specifically, while not quantitative, principal component analysis (PCA) provides a useful unsupervised classification technique.<sup>231</sup> PCA is common in chemometrics;<sup>40</sup> examples in the literature include identifying trace elements in wheat,<sup>232</sup> analysis of time of flight-secondary ion mass spectra from organic monolayers,<sup>42</sup> detecting sparse compounds via FTIR spectra,<sup>39</sup> and identifying peak shape changes in chromatography.<sup>41</sup> Specifically, PCA does not mathematically consider a known value, such as concentration, when fitting data. Instead, the matrix of wavenumbers and corresponding intensity for each sample spectrum goes through a dimensionality reduction such that the most variance is explained by the first component. Successive components explain less variance than the previous component. In chemistry applications, the chemical system has some known, or estimated,

number of species that provide a baseline for determining the number of principal components.

Fitting data to a linear model, or LR, is common for absorbance data, such as fitting to the Beer-Lambert Law to determine physical constants or identify concentrations of unknown samples.<sup>233</sup> Absorbance FTIR spectra follow a linear relationship of intensity with respect to concentration, which is advantageous for determining new sample composition. Recent work has utilized multiple LR to identify heavy metals, including investigating the effect of surface chemistry on vanadium<sup>37</sup> and lead<sup>38</sup> toxicity. However, the simplicity of the method ultimately restricts the model usefulness in more complex, dynamic systems.

Of the techniques considered, SVR is the most mathematically advanced ML model.<sup>234</sup> SVR fits training data to the best function by minimizing the distance of each value from the fitting equation to be able to predict discrete values, rather than a group assignment. Not all data is appropriate for SVR, but in cases where concentration is being predicted and it is linearly correlated with absorbance SVR 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.<sup>235</sup> As another example, ATR-FTIR and SVR were employed by Chen et al. 2022 to predict bio-oil characteristics quickly.<sup>236</sup> Our review of the literature and ML methods indicates that the SVR model will perform best for predicting saccharide concentration.

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 proxy-samples of glucose and egg serum albumin (ESA) with a relatively simple training dataset. 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.<sup>22,217,237</sup> 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.<sup>27,178,182</sup> While an unlikely protein to find in field samples, ESA provides a complex framework of amino acids that are abundant in the ocean's water column.<sup>146,169,171,212,238</sup> The utilization 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 proxy-SSML samples using ML.

## 6.2. Methods

### 6.2.1. 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 m $\Omega$ ) from a MilliQ system. For training spectra, stock solutions of 1M glucose (Sigma Aldrich,  $\geq 99.5\%$  (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 relevant amount of water. Specific details of each solution, including concentration, relative ratio, and volume of stock

solution are provided in the SI. Briefly, we selected this system and concentrations to have reasonable complexity.

Both the protein and saccharide have IR responses from 1800 to 900  $\text{cm}^{-1}$ . The peaks were well resolved, with minimal convolution of responses. 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 response. Concentrations were selected based on literature precedent from field study results.<sup>115,117,122</sup> Solutions were measured in triplicate via ATR-FTIR spectroscopy on a 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 a liquid nitrogen cooled HgCdTe (MCT) detector over 32 scans (approximately one minute) from 4000 to 450  $\text{cm}^{-1}$  with a resolution of 1  $\text{cm}^{-1}$ . Spectra were converted to absorbance with a water background using the established relationship of  $-\log(R/R_0)$ . Background correction was done using a linear fit model for the baseline to correct for inconsistent baseline between measurements. Water 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 20).

### 6.2.2. Proxy-Sample and Real Sea Surface Water Preparation and Sampling

For test data, stock 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%). 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.

### 6.2.3. Field Sampling

We operationally define the surface microlayer (SML) as the top 1 mm of the sampled water and bulk surface water (BSW) as the top 1 m of the sampled water. 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. 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. All samples were stored at room temperature and shipped; once received, samples were stored at 2°C until analyzed.

Sea and river BSW 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 10<sup>th</sup> and 11<sup>th</sup> 2023. BSW was collected by first copiously rinsing a glass jar, replacing the lid, submerging the covered jar, and finally removing the lid underwater. Jars were filled to avoid head space.

SML water was collected according to methods detailed by Harvey and Burzell.<sup>239</sup> 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 silicon squeegee into a copiously rinsed glass jar.

ATR-FTIR spectra were acquired for all field samples as described in the data collection and preprocessing methods section. In addition, DOC was extracted from the samples using the method detailed by Dittmar et al. and described in the SI.<sup>240</sup> Extracted DOC was analyzed via gas chromatography-mass spectrometry (GC-MS) to identify organic components (Appendix F).

#### 6.2.4. Machine Learning Methods

All machine learning (ML) methods were implemented using Python scripts. These are available online at the Allen Lab GitHub: <https://github.com/Ohio-State-Allen-Lab/Sea-Surface-Microlayer-MachineLearning>. PCA was used to elucidate any relationships between the data in the training set as a qualitative approach. Using the PCA method in the SciKit-Learn decomposition package, the principal components were determined based on the chemical system having four known components. We estimate that the glucose, ESA, and perturbed water contribute three components and a fourth component is included for error. The components were compared to each other to determine if a relationship exists for concentration or relative ratio.

To provide quantitative analysis of the sample concentrations, we implement linear regression (LR) of the FTIR training data set. The linear model method from SciKit-Learn was used to fit absorbance and concentration for the data. A SVR model was initialized using the support vector machine package from SciKit-Learn and trained using the FTIR training data set. Proxy and real SSML samples were evaluated via the SVR model to



predict concentration. The SVR model parameters were optimized by evaluating  $\epsilon$ , threshold tolerance, and C, regularization parameters to reach a minimization of mean squared error (MSE) (Appendix F). For the LR and SVR, a train-test split of 80:20 was used to randomly withhold data, which was determined by minimizing MSE and based on literature evidence (Appendix F). The MSE and  $R^2$  values were calculated using the SciKit-Learn Metrics package to compare all models. New data, including the proxy and real SSML samples, were evaluated with both LR and SVR models to predict concentration.

We evaluate a proxy solution and a real sample spectrum using pre-trained models from our previous work<sup>1</sup> to determine the functional groups present and confirm the predictive accuracy of the prior model on liquid-phase, mixtures samples. Previously, convolutional neural networks (CNN) were trained on gas-phase FTIR spectra to predict present and absent functional groups, and we expand on this in detail in the Supporting Information. We compare the known functional groups in our proxy solution and the model predictions to gain insight into the generalizability of the CNN models and deduce information about our unknown field sample.

In addition, feature extraction is presently being performed to further analyze the discrepancies between the chemically important wavenumbers and the ML relevant wavenumbers. These results will be included in the final publication of this work.

### 6.3. Results and Discussion

The chemical complexity of the SSML is explored via ATR-FTIR spectroscopy and quantitative machine learning approaches to develop a simple 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 was not of concern in the presently studied concentration ranges. Figure 21 is an example spectrum of a training sample with peaks assigned to the protein and glucose for reference. The two solute components of the training samples were well resolved from one another. The separation improved the likelihood that ML approaches were successful. A single figure containing all the acquired spectra is presented in Appendix F.

PCA provides a qualitative, or classification, model from an unsupervised dimensionality reduction. The resulting principal components (PCs) can be used to reconstruct a spectrum. We compare PC one and PC two to deduce information about the training spectra (Figure 22). Our relative ratio definition is such that '0' is equivalent to no glucose, or protein only, and '1' indicates that there is only glucose, or no protein. The resultant dimensionality reduction and comparison of PC1 and PC2 is expected given the input data is a gradient matrix of glucose and ESA concentrations. As a result of the input data, the PCA method provides us with less classification accuracy. We determine that PC1 largely represents the contribution of glucose to a spectrum and PC2 represents ESA contributions. Classification of a sample with more glucose, or greater relative ratio, would be concentration dependent.

LR provided a mathematically simple fitting of the training data but does not accurately predict on more complex samples (Figure 23a). We chose to evaluate the effectiveness of the fit with the data because absorbance is linear with concentration, especially in the low concentration regime of the SSML. As can be observed in Figure 23a,

the fit is exceptional for the training and testing data with an  $R^2$  value of 100 % and no mean squared error. However, when more complex samples containing both glucose and sucrose were evaluated, the model is unable to predict the concentration of sugar. Notably, the true concentration values have a slope that is greater than that of the training data. While we have selected the absorbance at  $1036\text{ cm}^{-1}$ , the LR is performed using all wavenumbers from  $1800$  to  $900\text{ cm}^{-1}$ , which eliminates feature selection biases.

In comparison to the LR, the SVR fits the training data and closely predict the concentration of sugar in more complex solutions (Figure 23b). Rather than fitting to a linear equation, SVR employs iterative fitting to find an equation that captures data and creates boundaries for which data should fall in. The higher-level mathematical complexity of the fit creates a more suitable model for predicting on more complex solutions, as observed in Figure 23b.

The SVR and LR models were directly compared via the relative difference in the predicted versus true concentration of saccharides (Table 7). Our SVR model correctly predicts the concentration for the three complex samples within tens of mM accuracy. In contrast, the LR model fails to achieve any predictive power. Despite the LR having a greater  $R^2$  value (100 %), the SVR computational complexity results in a slightly lower  $R^2$  and significantly improved regressive predictions. The positive relative difference highlights that samples A and B were under-predicted from their true concentration, while the negative relative difference for sample C indicates a predicted concentration higher than the true value.

The LR and SVR model fit results are presented in Table 8 for comparison. Despite the exceptional training metrics of the LR model, its predictive power does not translate to more complex samples. The SVR training metrics were slightly lower than the LR, but the SVR model outperforms in predictions on new, more complex data. More interestingly, the mean squared error of the SVR model, 0.02 M, is greater than the error in determining the discrete sugar concentration of the proxy samples, where the error was 10 mM or less from the true concentration value. Ultimately, these results suggest that the training metrics of the LR could be misleading as to success on future sample concentration prediction and that the decreased, but still excellent, metric values for SVR indicate the model is more suitable for applications including complex solution spectra. Thus, SVR is the model of choice for predicting sugar concentration.

We analyzed sea and river samples that were collected in January 2023 from Cocoa Beach, Florida to determine if the model could successfully identify saccharides in real ocean samples. The FTIR spectra of the samples are included in Appendix F. All the samples were predicted to have concentrations of saccharides in the mid to high mM range (70-100 mM) (Figure 25). Literature values range from 10-25 mM;<sup>150</sup> the predicted values are on the same order of magnitude albeit a factor of 2 to 4 times higher than what one might expect. The predicted concentrations for the known samples (Table 7) were within 10% of the true value, so we approximate that our predictions for unknown, real field samples may have a similar uncertainty. Further analysis via GC-MS of the unknown Florida samples was performed to investigate the samples more closely (Appendix F). Specifically, we employed GC-MS to confirm the presence of DOC and identify if

characteristic saccharide fragmentation was observed. As a general observation of the FTIR spectra, the absorbance at  $1036\text{ cm}^{-1}$  for the Cocoa Beach, Florida samples closely aligns with the training data. The alignment of the unknown samples with the data indicates that the model is suitable for saccharide concentration prediction.

The SVR feature weights are shown in Figure 26; the figure, while resembling an FTIR spectrum, is representing the importance of each wavenumber to the model success. For example, a value of zero indicates no influence in model prediction, a positive value indicates that the feature (wavenumber) is aiding in model prediction, and the converse is true for negative values. In conjunction with Figure 22, we can observe chemically relevant weighting of the SVR model. Wavenumbers above  $1500\text{ cm}^{-1}$  have a negative impact on the model; these wavenumbers are not associated with vibrational modes of the saccharide or protein in solution. The O-H bend of liquid water is centered at  $1650\text{ cm}^{-1}$ . A pure water spectrum is used as the background for the ATR measurements, so an FTIR response in this region is a result of water structure perturbation in solution. Thus, we posit that the region is ineffective for the mathematical model to predict saccharide concentration in comparison to the much more influential wavenumber region of about  $980\text{ cm}^{-1}$  to  $1180\text{ cm}^{-1}$ . Of equal interest is the importance of wavenumbers in the region of the protein amide vibrations ( $1300\text{-}1500\text{ cm}^{-1}$ ). The computational relevance may be partially explained by O-H bending of alcohol ( $1330\text{-}1420\text{ cm}^{-1}$ ) from glucose.

To further investigate the model success, we limited the wavenumbers to the primary regions of importance according to the feature extraction. Feature importance for wavenumbers from  $900\text{-}1500\text{ cm}^{-1}$  is shown in Figure 27. Reducing the input features from

900 wavenumbers to 600 wavenumbers did not impact feature importance. We observed a nominal increase in the MSE of 5 mM. Most notably, between both feature analyses, two wide bands of features (wavenumbers) remain important for the model to make saccharide concentration prediction. The analysis presented herein of one wavenumber to represent the model success is a shortcoming of our capabilities to accurately portray the model fit. It follows that the model predictions for the Florida samples are not well defined by one wavenumber and the corresponding absorbance, yet instead it is only a snapshot of one feature and the correlation to the predicted concentration.

We utilized CNN functional group assignment models from our previous work to determine if correct assignments could be achieved and explore the unknown sample (solution of bovine serum albumin, ESA, glucose, sucrose, and 1-butanol in water) further. The proxy sample with known composition is correctly assigned (Table 9). Only four functional groups were misassigned out of 17 groups; and three of those were predicted absent rather than present. The differentiation indicates that the model is underpredicting functional groups that were present (e.g., predicting alcohol is not present when it is). This incorrect assignment is likely due to the characteristic differences in the O-H vibrational peak in gas- versus liquid-phase spectroscopy. Liquid-phase O-H stretching is broadened from hydrogen bonding, which could occur between the water, protein, and other sugar molecules in the known proxy solution. The solution complexity most likely results in a broad O-H region in comparison to the neat, gas-phase spectra. Overall, the functional group assignment has 78% accuracy. Importantly, the model predicts that the Banana River FTIR spectrum has an aromatic functional group, which is consistent with the observed

mass of 77 m/z in the GC-MS (Appendix F). In addition, the CNN model predicts several nitrogen-containing functional groups (amide, nitrile, and nitro) in the Banana River sample, which is consistent with the several observed odd nominal masses (Appendix F).

Sensitivity, specificity, positive predictive value, and negative predictive value were calculated according to the definitions presented by Trevethan in 2017 (Table 10).<sup>241</sup> Specificity, or how well the model correctly assigns negative cases, is determined to be 90%. Sensitivity, with a value of 57%, indicates that the model is not optimal for identifying positive cases; however, the positive predictive value is 80%. The Florida field sample results provide insight into the composition of the spectrum and respective sample. The results provide qualitative insight about the samples and further confirms the presence of organics in the field sample. The correct functional group assignments and minimal misassignments emphasizes the utility of our prior model that was trained on neat, gas-phase spectra. A larger, more diverse mixture training data set would increase all the analyzed metrics, as well.

Overall, the results from the CNN provide contextualization of the samples without the requirement of a lengthy extraction process to identify DOC (Appendix F). The generalizable models from our 2021 publication provide a framework for improving upon the current analysis methods utilized for ocean surface samples. Furthermore, the prediction of functional groups provides qualitative insights into field samples with a simple sampling methodology. The approach detailed herein serves as a supplement to field analysis for faster qualitative observations.

Our quantitative results indicate that a computationally inexpensive model, SVR, provides predictions of sugar concentration within 10 mM of the true value. In comparison to LR, the SVR has a slightly lower coefficient of determination but provides much more accurate concentrations on elaborate test samples. Even with increased sample complexity, including additional sugar, protein, and lipid molecules, the SVR model accurately predicts the total sugar concentration. When tested on field samples, the SVR model predicts sugar concentration within the expected values that have been presented in the literature for carbohydrates.<sup>150</sup> Samples were successfully examined via the functional group assignment model previously developed, which informs as to the presence of organic carbon in unknown samples, including real field samples.

#### 6.4. Conclusions

Several ML methods were applied to ATR-FTIR spectra to determine concentration and chemical composition of aqueous samples to develop efficient, less-expensive analytical techniques for analysis of the SSML. Our multifaceted approach includes examining LR and SVR for quantitative analysis, PCA for quasi-quantitative grouping, and a CNN for qualitative assignment of functional groups. Our results indicate that SVR is viable for complex solutions, especially considering the training sample data is relatively simple. The repurposed, generalizable CNN provides valuable insight into the functional groups present in the samples and validates the SVR assignment by confirming the presence of organics in the field sample. The research presented herein provides a unique approach to studying the SSML utilizing the advanced computational tools available and reduces the time needed to perform analyses of field SSML samples. Further work should



focus on finding an optimal training data set, investigating other concentration quantification, and intercalating other spectroscopic or spectrometric data, to name a few. An improved understanding of the SSML is achievable, wherein more frequent measurements and analysis can occur, ultimately providing more information about the productivity of the SSML and its effects on our atmosphere and climate.

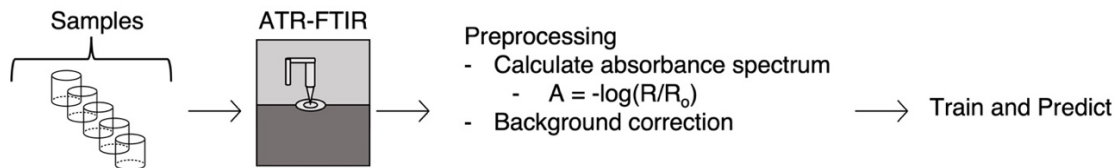


Figure 21. Schematic flow chart of data collection process to the ML pipeline.

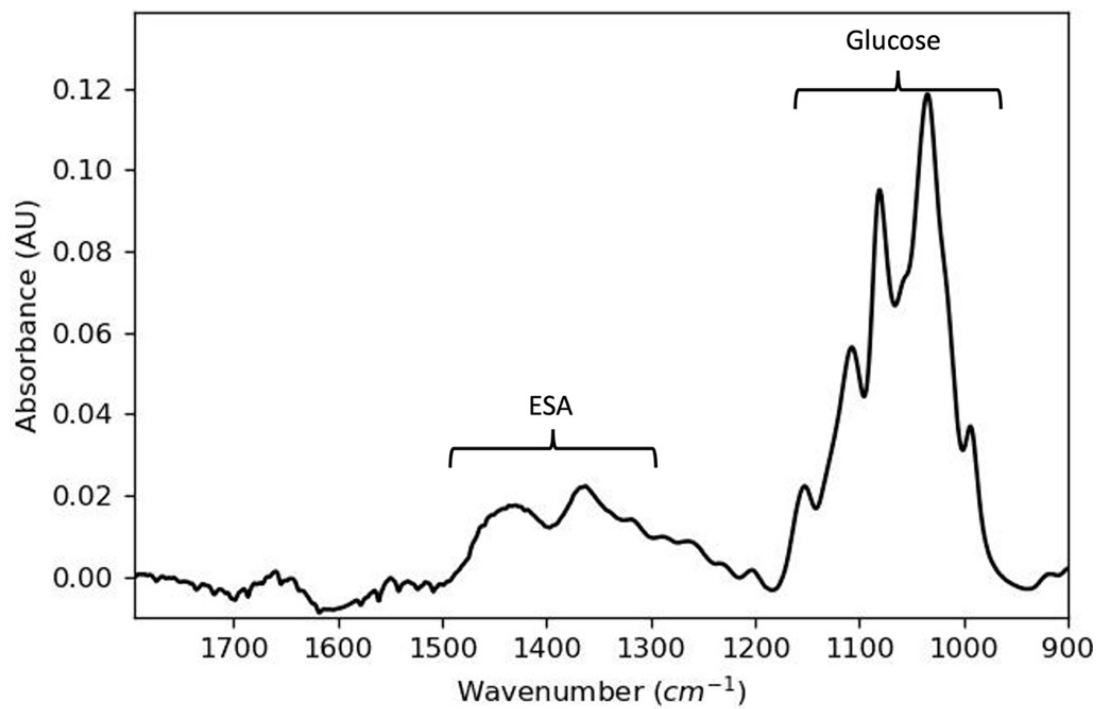


Figure 22. ATR-FTIR spectrum of 0.6 M glucose and 2 mg/mL egg serum albumin. The labels are provided to emphasize that the components do not compound on one another and are well resolved, despite being in a similar wavenumber region.

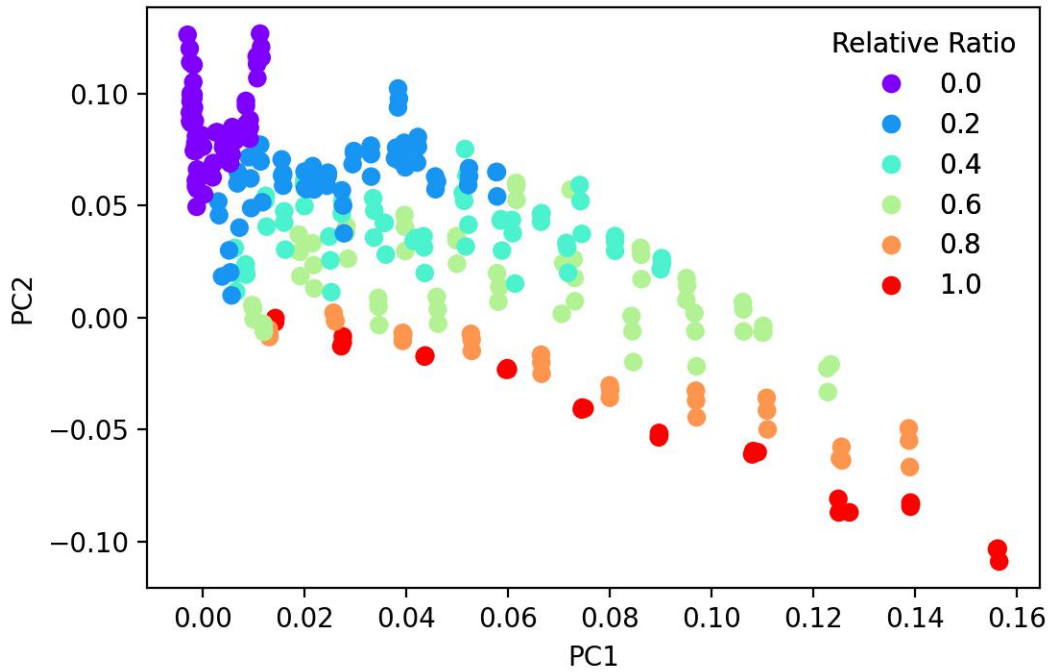


Figure 23. Principal components (PCs) one and two from the data dimensionality reduction performed using principal component analysis (PCA). The relative ratio is respective to glucose. Solutions with a relative ratio of '1' have no ESA. PC1 mainly captures glucose response and PC2 mainly captures ESA response.

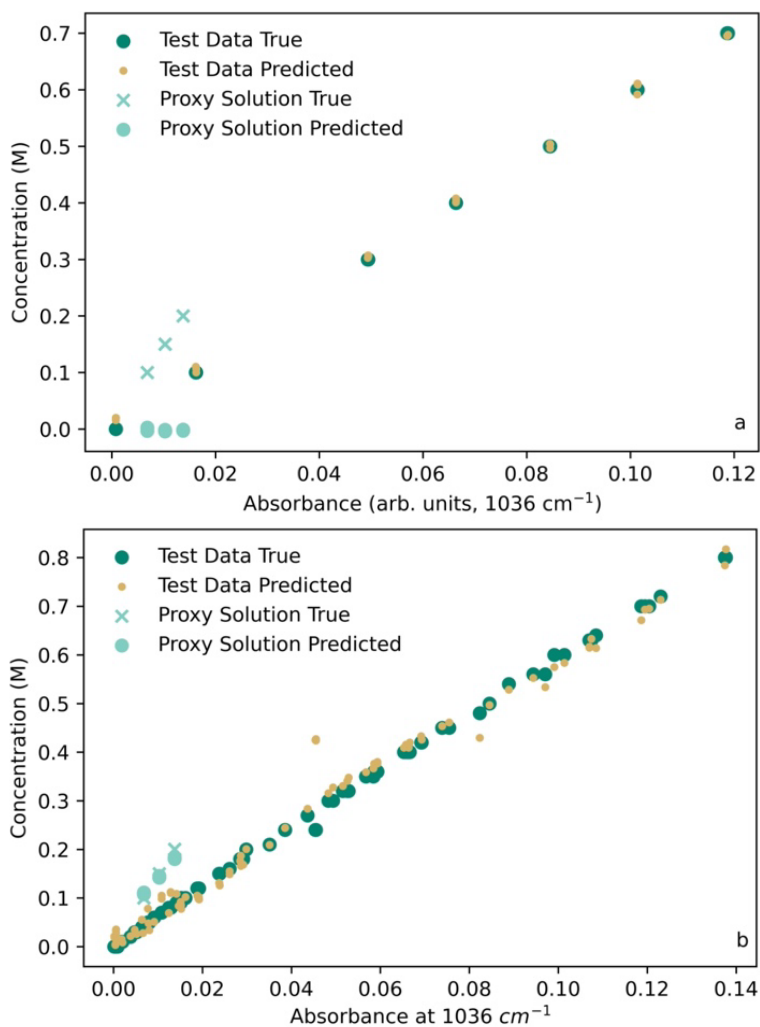


Figure 24. Linear regression (LR) (a) fits the experimental training data well with a 100 % R<sup>2</sup> and no mean squared error. ‘True’ indicates the known concentration of saccharide while ‘predicted’ is the model’s estimate. Proxy sample saccharide concentrations are not correctly predicted, as shown with the teal ‘X’ demarcating the known saccharide concentration. Support vector regression (SVR) (b) results show that the test data accurately follows the training data. Predicted concentrations for the known complex samples are much closer to the true concentration. The training results in an R<sup>2</sup> of 97.1%.

Table 7. Predicted sugar concentration (M) in more complex samples containing glucose, sucrose, ESA, bovine serum albumin (BSA), and 1-butanol are predicted by the SVR and LR model. Values are the average predicted concentration (M). The SVR model predicts reasonable concentration values in the range of the true concentration, while the LR model predictions do not provide any reasonable estimates of concentration.

Sample Label	Concentration of sugar (M)	Average Predicted SVR (M)	Average Predicted LR (M)
A	0.200	0.182	-0.002
B	0.150	0.143	-0.003
C	0.100	0.109	0

Table 8.  $R^2$  and mean squared error of linear regression (LR) and support vector regression (SVR) models after training.

Metric	LR	SVR
$R^2$ (%)	100	97.1
Mean Squared Error (M)	0.00	0.02

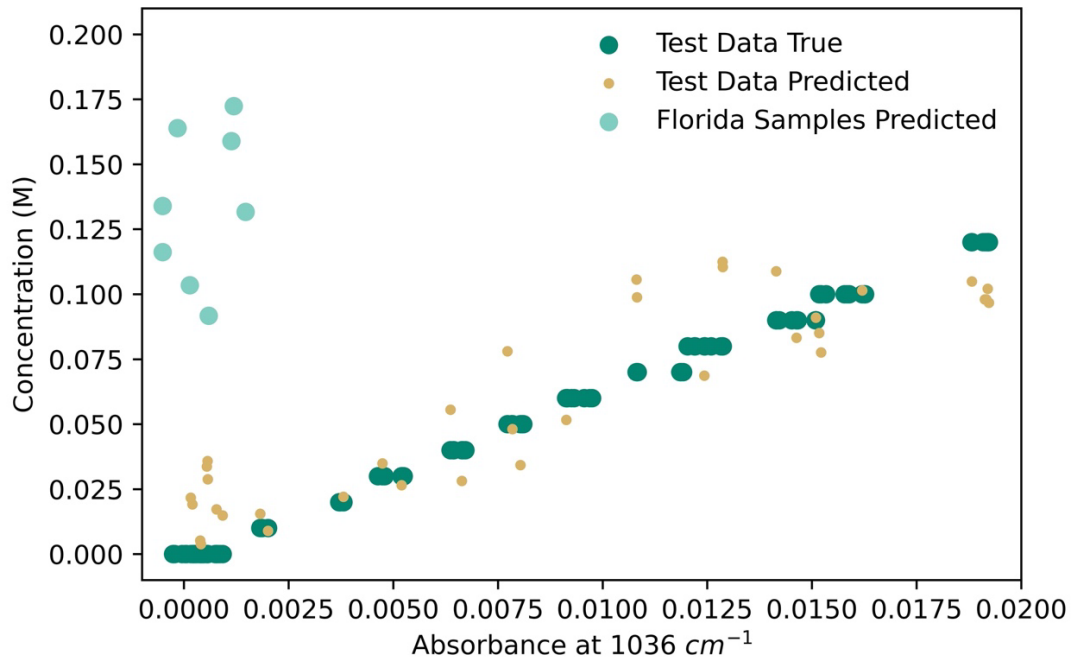


Figure 25. Support vector regression (SVR) model predictions on unknown field samples are closely aligned with the training and test data although are in the low absorbance range.



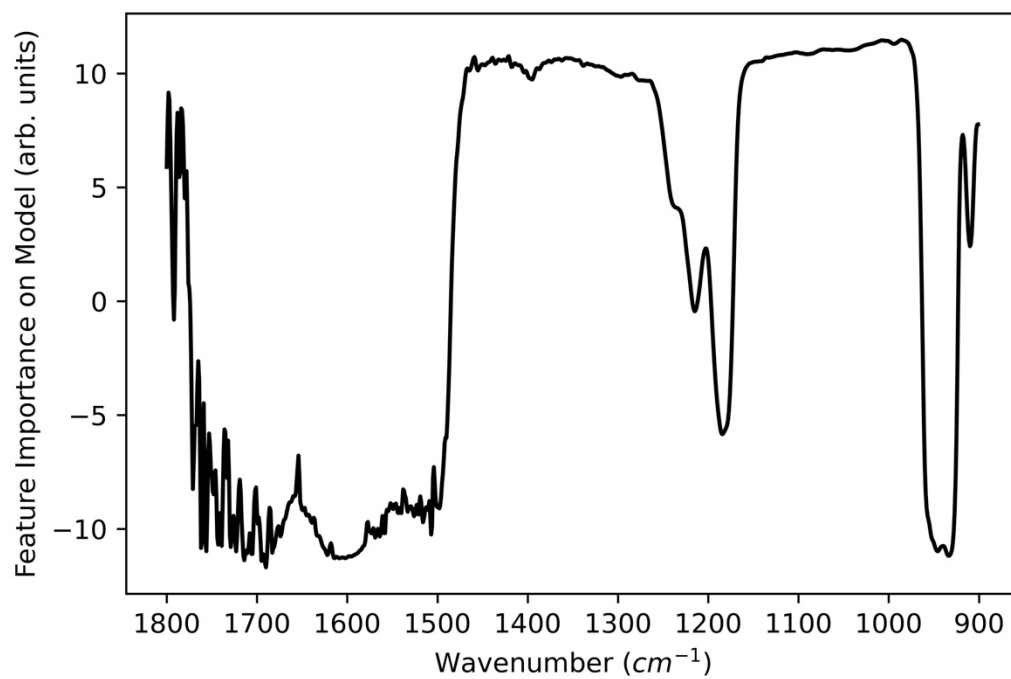


Figure 26. Feature extraction for SVR model. Positive values indicate strong influence on model prediction and negative values indicate negative impacts on model success.

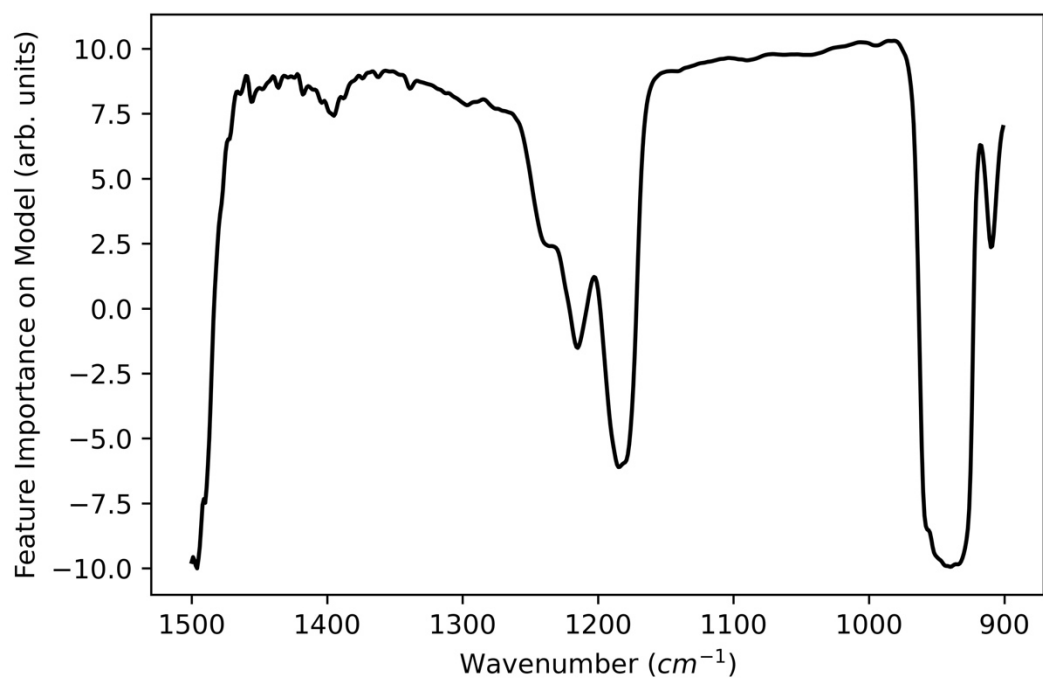


Figure 27. Feature extraction for SVR model with reduced features. The model uses similar features for prediction and the negatively impacting features have been removed.

Table 9. Functional group analysis of proxy sample and unknown SSML sample. Red text indicates that the model incorrectly predicted (e.g., nitrile is predicted present for the proxy sample, yet it is not present). An asterisk (for Banana River sample only) indicates that the GC-MS of the sample has characteristic m/z values for that functional group identification.

Prediction	Proxy Sample	Banana River Surface January 2023
Present	alkene, amide, ester, methyl, <b>nitrile</b>	alcohol, alkyne, amide*, aromatic*, nitrile*, nitro*
Absent	acyl halide, <b>alcohol</b> , aldehyde, <b>alkane</b> , alkyl halide, alkyne, amine, aromatic, carboxylic acid, <b>ether</b> , ketone, nitro	acyl halide, aldehyde, alkane, alkene, alkyl halide, amine, carboxylic acid, ester, ether, ketone, methyl

Table 10. Sensitivity, specificity, positive predictive value, and negative predictive value for model results on proxy sample prediction of functional groups. These metrics provide a more thorough analysis of how the model performs and detail the model's performance more holistically.

Metric	Value (%)
Sensitivity	57
Specificity	90
Positive predictive value	80
Negative predictive value	75

## Chapter 7. Conclusions

The presented studies provide an improved understanding of the ever-complex sea surface and interfacial region. Each approach was explicitly defined to target the various options for investigating the SSML. While complex and multifaceted, the approaches converge over a theme of computational approaches that advance the understanding of laboratory and model data from the ocean's surface.

ML provides a useful tool for producing quick, efficient, and accurate functional group predictions from gas-phase FTIR spectra. The fundamental work conducted for this study proves that the analysis of FTIR spectra is sufficiently handled by a ML model. While not full structure prediction, the primary purpose of many FTIR spectra is functional group analysis. This is improved upon by eliminating the need for a bulky searching library. The models are even capable of evaluating spectra of molecules that have not been included in the training dataset and provide correct identification.

Global models, such as E3SM, provide an avenue for predictions of the ocean surface composition through the application of Gibbs free energy, biological relationships, and surface adsorption equations. While these models are indirect measurements, they are guided by field observations and literature data to provide suitable estimations. Utilizing the global model output and novel arrangement of established relationships, global surface

carbon is estimated to  $\sim 10^{-4}$  Gt. The impact of the model being that these results can be used to further estimate physical properties of the surface and provide feedback to the model for future iterations to improve the modeling.

Experimental techniques are explored to gain additional fundamental information about the physical chemistry of an air-water interface. As expected based on literature results, the most important factor affecting surface adsorption is ionic strength. Temperature and presence of a lipid monolayer have less significant effect on the surface structure; slightly more protein adsorbs at higher temperatures and a preexisting monolayer minimally decreases surface adsorption.

The results from each study guided the investigation into the best computational method for predicting unknown sugar concentrations to reduce the amount of preparation required to study real SSML samples. SVR proved most accurate at predicting unknown sugar concentrations when trained on a labeled mixture of protein and sugar FTIR spectra dataset. When tested on increasingly complex samples with known sugar concentration, the SVR model successfully predicted total sugar concentration. These results help guide the development of techniques to evaluate the SSML more frequently while reducing sample workup required.

Ultimately, the use of computational methods, including modeling and ML, enabled a cohesive, collaborative examination of the ocean's surface. The results presented herein can be extended to increasingly more complex systems and provide a basis for future work. The SSML itself is a chemically and physically complex system that requires a holistic investigation through which laboratory and computational results guide the

comprehensive understanding of the surface and its effect and impact on the global climate, atmosphere, and environment. As humans face the inevitable changes from incomprehensible carbon emissions, the science of the SSML should not be ignored in consideration for reparations and restoration. The ocean's surface will surely have a vital role in global climate change.

## References

- (1) Adamson, A. W. *Physical Chemistry of Surface*; John Wiley & Sons, Inc., 1967.
- (2) Marchand, A.; Weijs, J. H.; Snoeijs, J. H.; Andreotti, B. Why Is Surface Tension a Force Parallel to the Interface? *Am. J. Phys.* **2011**, *79* (10), 999–1008. <https://doi.org/10.1119/1.3619866>.
- (3) Davies, J. T.; Rideal, E. K. *Interfacial Phenomena*; Academic Press, 1961.
- (4) Hauner, I. M.; Deblais, A.; Beattie, J. K.; Kellay, H.; Bonn, D. The Dynamic Surface Tension of Water. *J. Phys. Chem. Lett.* **2017**, *8* (7), 1599–1603. <https://doi.org/10.1021/acs.jpcelett.7b00267>.
- (5) Fischer, B.; Teer, E.; Knobler, C. M. Optical Measurements of the Phase Diagram of Langmuir Monolayers of Fatty Acid–Alcohol Mixtures. *J. Chem. Phys.* **1995**, *103* (6), 2365–2368. <https://doi.org/10.1063/1.469659>.
- (6) Teer, E.; Knobler, C. M.; Lautz, C.; Wurlitzer, S.; Kildae, J.; Fischer, T. M. Optical Measurements of the Phase Diagrams of Langmuir Monolayers of Fatty Acid, Ester, and Alcohol Mixtures by Brewster-Angle Microscopy. *J. Chem. Phys.* **1997**, *106* (5), 1913–1920. <https://doi.org/10.1063/1.473312>.
- (7) Steinke, I.; Demott, P. J.; Deane, G. B.; Hill, T. C. J.; Maltrud, M.; Raman, A.; Burrows, S. M. A Numerical Framework for Simulating the Atmospheric Variability of Supermicron Marine Biogenic Ice Nucleating Particles. *Atmospheric Chem. Phys.* **2022**, *22* (2), 847–859. <https://doi.org/10.5194/acp-22-847-2022>.
- (8) 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>.
- (9) Hardy, J. T. The Sea Surface Microlayer: Biology, Chemistry and Anthropogenic Enrichment. *Prog. Oceanogr.* **1982**, *11* (4), 307–328. [https://doi.org/10.1016/0079-6611\(82\)90001-5](https://doi.org/10.1016/0079-6611(82)90001-5).
- (10) 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>.
- (11) Nebbioso, A.; Piccolo, A. Molecular Characterization of Dissolved Organic Matter (DOM): A Critical Review. *Anal. Bioanal. Chem.* **2013**, *405* (1), 109–124. <https://doi.org/10.1007/s00216-012-6363-2>.
- (12) 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>.

- (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) 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>.
- (15) 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>.
- (16) 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>.
- (17) Collins, D. B.; Ault, A. P.; Moffet, R. C.; Ruppel, M. J.; Cuadra-Rodriguez, L. A.; Guasco, T. L.; Corrigan, C. E.; Pedler, B. E.; Azam, F.; Aluwihare, L. I.; Bertram, T. H.; Roberts, G. C.; Grassian, V. H.; Prather, K. A. Impact of Marine Biogeochemistry on the Chemical Mixing State and Cloud Forming Ability of Nascent Sea Spray Aerosol. *J. Geophys. Res. Atmospheres* **2013**, *118* (15), 8553–8565. <https://doi.org/10.1002/jgrd.50598>.
- (18) McCoy, D. T.; Burrows, S. M.; Wood, R.; Grosvenor, D. P.; Elliott, S. M.; Ma, P. L.; Rasch, P. J.; Hartmann, D. L. Natural Aerosols Explain Seasonal and Spatial Patterns of Southern Ocean Cloud Albedo. *Sci. Adv.* **2015**, *1* (6), 1–12.  
<https://doi.org/10.1126/sciadv.1500157>.
- (19) Adams, E. M.; Casper, C. B.; Allen, H. C. Effect of Cation Enrichment on Dipalmitoylphosphatidylcholine (DPPC) Monolayers at the Air-Water Interface. *J. Colloid Interface Sci.* **2016**, *478*, 353–364.  
<https://doi.org/10.1016/j.jcis.2016.06.016>.
- (20) Vazquez de Vasquez, M. G.; Rogers, M. M.; Carter-Fenk, K. A.; Allen, H. C. Discerning Poly- and Monosaccharide Enrichment Mechanisms: Alginate and

- Glucuronate Adsorption to a Stearic Acid Sea Surface Microlayer. *ACS Earth Space Chem.* **2022**, *6*, 1581–1595.  
<https://doi.org/10.1021/acsearthspacechem.2c00066>.
- (21) 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>.
- (22) 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>.
- (23) Mustaffa, N. I. H.; Badewien, T. H.; Ribas-Ribas, M.; Wurl, O. High-Resolution Observations on Enrichment Processes in the Sea-Surface Microlayer. *Sci. Rep.* **2018**, *8* (1), 1–12. <https://doi.org/10.1038/s41598-018-31465-8>.
- (24) Tanford, C. *Ben Franklin Stilled the Waves: An Informal History of Pouring Oil on Water with Reflections on the Ups and Downs of Scientific Life in General*, 1st ed.; Oxford University Press, 2004.
- (25) Pockels, A. Surface Tension. *Nature* **1891**, *43* (1115), 437–439.
- (26) Langmuir, I. The Constitution and Fundamental Properties of Solids and Liquids. II. Liquids. *J. Am. Chem. Soc.* **1917**, *39* (9), 1848–1906.  
<https://doi.org/10.1021/ja02254a006>.
- (27) 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>.
- (28) Blodgett, K. B. Films Built by Depositing Successive Monomolecular Layers on a Solid Surface. *J. Am. Chem. Soc.* **1935**, *57* (6), 1007–1022.  
<https://doi.org/10.1021/ja01309a011>.
- (29) Hanlan, J.; Skoog, D. A.; West, D. M. Principles of Instrumental Analysis. *Stud. Conserv.* **1973**, *18* (1), 45. <https://doi.org/10.2307/1505543>.
- (30) Harris, D. *Exploring Chemical Analysis*, 5th ed.; 2013.
- (31) Mertz, L. Fourier Spectroscopy, Past, Present, and Future. *Appl. Opt.* **1971**, *10* (2), 386–389. <https://doi.org/10.1364/AO.10.000386>.
- (32) Ferraro, J. R.; Basile, L. J. *Fourier Transform Infrared Spectrometry*; New York: Academic Press, 1978.
- (33) Griffiths, P. R.; de Haseth, J. A. *Fourier Transform Infrared Spectrometry*; Wiley-Interscience, 1986.
- (34) Averett, L. A.; Griffiths, P. R.; Nishikida, K. Effective Path Length in Attenuated Total Reflection Spectroscopy. *Anal. Chem.* **2008**, *80* (8), 3045–3049.  
<https://doi.org/10.1021/ac7025892>.

- (35) Mendelsohn, R. External Infrared Reflection Absorption Spectrometry of Monolayer Films at the Air-Water Interface. *Annu. Rev. Phys. Chem.* **1995**, *46* (1), 305–334. <https://doi.org/10.1146/annurev.physchem.46.1.305>.
- (36) Watson, G. S. Linear Least Squares Regression. *Ann. Math. Stat.* **1967**, *38* (6), 1679–1699. <https://doi.org/10.1214/aoms/1177698603>.
- (37) 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>.
- (38) 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>.
- (39) Hasegawa, T. Detection of Minute Chemical Species by Principal-Component Analysis. *Anal. Chem.* **1999**, *71* (15), 3085–3091. <https://doi.org/10.1021/ac981430z>.
- (40) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. 16.
- (41) Macnaughtan, Donald.; Rogers, L. B.; Wernimont, Grant. Principal-Component Analysis Applied to Chromatographic Data. *Anal. Chem.* **1972**, *44* (8), 1421–1427. <https://doi.org/10.1021/ac60316a016>.
- (42) Biesinger, M. C.; Paepegaey, P.-Y.; McIntyre, N. S.; Harbottle, R. R.; Petersen, N. O. Principal Component Analysis of TOF-SIMS Images of Organic Monolayers. *Anal. Chem.* **2002**, *74* (22), 5711–5716. <https://doi.org/10.1021/ac020311n>.
- (43) Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **1936**, *28* (3/4), 321–377. <https://doi.org/10.2307/2333955>.
- (44) Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. In *Advances in Neural Information Processing Systems*; MIT Press, 1996; Vol. 9.
- (45) Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. **2017**.
- (46) Malek, S.; Melgani, F.; Bazi, Y. One-Dimensional Convolutional Neural Networks for Spectroscopic Signal Regression. *J. Chemom.* **2018**, *32* (5), 1–17. <https://doi.org/10.1002/cem.2977>.
- (47) Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*; 2016.
- (48) Yonkos, L. T.; Friedel, E. A.; Perez-Reyes, A. C.; Ghosal, S.; Arthur, C. D. Microplastics in Four Estuarine Rivers in the Chesapeake Bay, U.S.A. *Environ. Sci. Technol.* **2014**, *48* (24), 14195–14202. <https://doi.org/10.1021/es5036317>.
- (49) Song, Y. K.; Hong, S. H.; Jang, M.; Kang, J. H.; Kwon, O. Y.; Han, G. M.; Shim, W. J. Large Accumulation of Micro-Sized Synthetic Polymer Particles in the Sea

- Surface Microlayer. *Environ. Sci. Technol.* **2014**, *48* (16), 9014–9021. <https://doi.org/10.1021/es501757s>.
- (50) Lee, H. J.; Song, N. S.; Kim, J. S.; Kim, S. K. Variation and Uncertainty of Microplastics in Commercial Table Salts: Critical Review and Validation. *J. Hazard. Mater.* **2021**, *402* (August 2020), 123743. <https://doi.org/10.1016/j.jhazmat.2020.123743>.
- (51) Bogard, J. S.; Johnson, S. A.; Kumar, Romesh.; Cunningham, P. T. Quantitative Analysis of Nitrate Ion in Ambient Aerosols by Fourier-Transform Infrared Spectroscopy. *Environ. Sci. Technol.* **1982**, *16* (3), 136–140. <https://doi.org/10.1021/es00097a004>.
- (52) Anil, I.; Golcuk, K.; Karaca, F. ATR-FTIR Spectroscopic Study of Functional Groups in Aerosols: The Contribution of a Saharan Dust Transport to Urban Atmosphere in Istanbul, Turkey. *Water. Air. Soil Pollut.* **2014**, *225* (3). <https://doi.org/10.1007/s11270-014-1898-9>.
- (53) Linker, R.; Shmulevich, I.; Kenny, A.; Shaviv, A. Soil Identification and Chemometrics for Direct Determination of Nitrate in Soils Using FTIR-ATR Mid-Infrared Spectroscopy. *Chemosphere* **2005**, *61* (5), 652–658. <https://doi.org/10.1016/j.chemosphere.2005.03.034>.
- (54) Hardy, J. T. The Sea Surface Microlayer: Biology, Chemistry and Anthropogenic Enrichment. *Progress in Oceanography*. 1982. [https://doi.org/10.1016/0079-6611\(82\)90001-5](https://doi.org/10.1016/0079-6611(82)90001-5).
- (55) Wurl, O.; Obbard, J. P. A Review of Pollutants in the Sea-Surface Microlayer (SML): A Unique Habitat for Marine Organisms. *Mar. Pollut. Bull.* **2004**, *48* (11–12), 1016–1030. <https://doi.org/10.1016/j.marpolbul.2004.03.016>.
- (56) P. M. Michel, A.; E. Morrison, A.; L. Preston, V.; T. Marx, C.; C. Colson, B.; K. White, H. Rapid Identification of Marine Plastic Debris via Spectroscopic Techniques and Machine Learning Classifiers. *Environ. Sci. Amp Technol.* **2020**, *54* (17), 10630–10637. <https://doi.org/10.1021/acs.est.0c02099>.
- (57) Rezania, S.; Park, J.; Md Din, M. F.; Mat Taib, S.; Talaiekhosani, A.; Kumar Yadav, K.; Kamyab, H. Microplastics Pollution in Different Aquatic Environments and Biota: A Review of Recent Studies. *Mar. Pollut. Bull.* **2018**, *133* (May), 191–208. <https://doi.org/10.1016/j.marpolbul.2018.05.022>.
- (58) Carey, F. A.; Giuliano, R. M. *Organic Chemistry*; McGraw-Hill Education, 2016.
- (59) Carey, F. A.; Sundberg, R. J. *Advanced Organic Chemistry Part A: Structure and Mechanisms*; Springer US, 2007.
- (60) Bowling Barnes, R.; Gore, R. C. Infrared Spectroscopy. *Anal. Chem.* **1949**, *21* (1), 7–12. <https://doi.org/10.1021/ac60025a003>.
- (61) Peck, R. L. Characterization of Organic Compounds. *Anal. Chem.* **1950**, *22* (1), 121–126. <https://doi.org/10.1021/ac60037a027>.
- (62) Kedzierski, M.; Falcou-Préfol, M.; Kerros, M. E.; Henry, M.; Pedrotti, M. L.; Bruzaud, S. A Machine Learning Algorithm for High Throughput Identification of FTIR Spectra: Application on Microplastics Collected in the Mediterranean Sea.

- Chemosphere* **2019**, *234*, 242–251.  
<https://doi.org/10.1016/j.chemosphere.2019.05.113>.
- (63) Galimberti, D. R.; Bougueroua, S.; Mahé, J.; Tommasini, M.; Rijs, A. M.; Gaigeot, M. P. Conformational Assignment of Gas Phase Peptides and Their H-Bonded Complexes Using Far-IR/THz: IR-UV Ion Dip Experiment, DFT-MD Spectroscopy, and Graph Theory for Mode Assignment. *Faraday Discuss.* **2019**, *217*, 67–97. <https://doi.org/10.1039/c8fd00211h>.
- (64) Bougueroua, S.; Spezia, R.; Pezzotti, S.; Vial, S.; Quessette, F.; Barth, D.; Gaigeot, M. P. Graph Theory for Automatic Structural Recognition in Molecular Dynamics Simulations. *J. Chem. Phys.* **2018**, *149* (18). <https://doi.org/10.1063/1.5045818>.
- (65) Coe, J. V.; Nystrom, S. V.; Chen, Z.; Li, R.; Verreault, D.; Hitchcock, C. L.; Martin, E. W.; Allen, H. C. Extracting Infrared Spectra of Protein Secondary Structures Using a Library of Protein Spectra and the Ramachandran Plot. *J. Phys. Chem. B* **2015**, *119* (41), 13079–13092. <https://doi.org/10.1021/acs.jpcc.5b08052>.
- (66) Geiger, A.; Cao, Z.; Song, Z.; Ulcickas, J.; Simpson, G. Chapter 18. Autonomous Science: Big Data Tools for Small Data Problems in Chemistry; 2020; pp 450–487. <https://doi.org/10.1039/9781839160233-00450>.
- (67) Nalla, R.; Pinge, R.; Narwaria, M.; Chaudhury, B. Priority Based Functional Group Identification of Organic Molecules Using Machine Learning. *ACM Int. Conf. Proceeding Ser.* **2018**, 201–209. <https://doi.org/10.1145/3152494.3152522>.
- (68) Yang, J.; Xu, J.; Zhang, X.; Wu, C.; Lin, T.; Ying, Y. Deep Learning for Vibrational Spectral Analysis: Recent Progress and a Practical Guide. *Anal. Chim. Acta* **2019**, *1081*, 6–17. <https://doi.org/10.1016/j.aca.2019.06.012>.
- (69) Kowalski, B. R.; Jurs, P. C.; Isenhour, T. L.; Reilley, C. N.; Isenhour, T. L.; Jurs, P. C. Computerized Learning Machines Applied to Chemical Problems Interpretation of Infrared Spectrometry Data. *Anal. Chem.* **1969**, *41* (14), 1945–1949. <https://doi.org/10.1021/ac50159a026>.
- (70) Fessenden, R. J.; Györgyi, L. Identifying Functional Groups in IR Spectra Using an Artificial Neural Network. *J Chem Soc Perkin Trans 2* **1991**, 1755.
- (71) van Est, Q. C.; Schoenmakers, P. J.; Smits, J. R. M.; Nijssen, W. P. M. Practical Implementation of Neural Networks for the Interpretation of Infrared Spectra. *Vib. Spectrosc.* **1993**, *4* (3), 263–272. [https://doi.org/10.1016/0924-2031\(93\)80001-V](https://doi.org/10.1016/0924-2031(93)80001-V).
- (72) Wu, W.; Massart, D. L. Artificial Neural Networks in Classification of NIR Spectral Data: Selection of the Input. *Chemom. Intell. Lab. Syst.* **1996**, *35* (1), 127–135. [https://doi.org/10.1016/S0169-7439\(96\)00034-2](https://doi.org/10.1016/S0169-7439(96)00034-2).
- (73) Averill, D. F.; Baird, K. C.; Hopkins, L. L.; Yerkes, M. J. Fourier Transform Infrared Spectroscopy without an FTIR Spectrometer: Library Searching and Concise Storage of Spectra. *J. Chem. Inf. Comput. Sci.* **1990**, *30* (2), 133–136. <https://doi.org/10.1021/ci00066a006>.
- (74) Santos, V. H. J. M. D.; Bruzza, E. D. C.; De Lima, J. E.; Lourega, R. V.; Rodrigues, L. F. Discriminant Analysis and Cluster Analysis of Biodiesel Fuel Blends Based on Fourier Transform Infrared Spectroscopy (FTIR). *Energy Fuels* **2016**, *30* (6), 4905–4915. <https://doi.org/10.1021/acs.energyfuels.6b00447>.

- (75) Judge, K.; Brown, C. W.; Hamel, L. Sensitivity of Infrared Spectra to Chemical Functional Groups. *Anal. Chem.* **2008**, *80* (11), 4186–4192. <https://doi.org/10.1021/ac8000429>.
- (76) Renner, G.; Schmidt, T. C.; Schram, J. A New Chemometric Approach for Automatic Identification of Microplastics from Environmental Compartments Based on FT-IR Spectroscopy. *Anal. Chem.* **2017**, *89* (22), 12045–12053. <https://doi.org/10.1021/acs.analchem.7b02472>.
- (77) Golz, E. K.; Vander Griend, D. A. Modeling Methylene Blue Aggregation in Acidic Solution to the Limits of Factor Analysis. *Anal. Chem.* **2013**, *85* (2), 1240–1246. <https://doi.org/10.1021/ac303271m>.
- (78) Agbaria, A. H.; Beck Rosen, G.; Lapidot, I.; Rich, D. H.; Huleihel, M.; Mordechai, S.; Salman, A.; Kapelushnik, J. Differential Diagnosis of the Etiologies of Bacterial and Viral Infections Using Infrared Microscopy of Peripheral Human Blood Samples and Multivariate Analysis. *Anal. Chem.* **2018**, *90*, 7888–7895. <https://doi.org/10.1021/acs.analchem.8b00017>.
- (79) Sharaha, U.; Rodriguez-Diaz, E.; Sagi, O.; Riesenber, K.; Lapidot, I.; Segal, Y.; Bigio, I. J.; Huleihel, M.; Salman, A. Detection of Extended-Spectrum  $\beta$ -Lactamase-Producing Escherichia Coli Using Infrared Microscopy and Machine-Learning Algorithms. *Anal. Chem.* **2019**, *91* (3), 2525–2530. <https://doi.org/10.1021/acs.analchem.8b05497>.
- (80) Kedzierski, M.; Falcou-Préfol, M.; Kerros, M. E.; Henry, M.; Pedrotti, M. L.; Bruzard, S. A Machine Learning Algorithm for High Throughput Identification of FTIR Spectra: Application on Microplastics Collected in the Mediterranean Sea. *Chemosphere* **2019**, *234*, 242–251. <https://doi.org/10.1016/j.chemosphere.2019.05.113>.
- (81) Kananenka, A. A.; Yao, K.; Corcelli, S. A.; Skinner, J. L. Machine Learning for Vibrational Spectroscopic Maps. *J. Chem. Theory Comput.* **2019**, *15* (12), 6850–6858. <https://doi.org/10.1021/acs.jctc.9b00698>.
- (82) Morawietz, T.; Urbina, A. S.; Wise, P. K.; Wu, X.; Lu, W.; Ben-Amotz, D.; Markland, T. E. Hiding in the Crowd: Spectral Signatures of Overcoordinated Hydrogen-Bond Environments. *J. Phys. Chem. Lett.* **2019**, *10* (20), 6067–6073. <https://doi.org/10.1021/acs.jpcllett.9b01781>.
- (83) Zhai, Y.; Caruso, A.; Gao, S.; Paesani, F. Active Learning of Many-Body Configuration Space: Application to the Cs<sup>+</sup>-Water MB-Nrg Potential Energy Function as a Case Study. *J. Chem. Phys.* **2020**, *152* (14). <https://doi.org/10.1063/5.0002162>.
- (84) Doblies, A.; Boll, B.; Fiedler, B. Prediction of Thermal Exposure and Mechanical Behavior of Epoxy Resin Using Artificial Neural Networks and Fourier Transform Infrared Spectroscopy. *Polymers* **2019**, *11* (2), 363. <https://doi.org/10.3390/polym11020363>.
- (85) Lasch, P.; Stämmler, M.; Zhang, M.; Baranska, M.; Bosch, A.; Majzner, K. FT-IR Hyperspectral Imaging and Artificial Neural Network Analysis for Identification

- of Pathogenic Bacteria. *Anal. Chem.* **2018**, *90* (15), 8896–8904. <https://doi.org/10.1021/acs.analchem.8b01024>.
- (86) Lasch, P.; Diem, M.; Hänsch, W.; Naumann, D. Artificial Neural Networks as Supervised Techniques for FT-IR Microspectroscopic Imaging. *J. Chemom.* **2006**, *20* (5), 209–220. <https://doi.org/10.1002/cem.993>.
- (87) Takamura, A.; Haramkova, 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>.
- (88) Fine, J. A.; Rajasekar, A. A.; Jethava, K. P.; Chopra, G. Spectral Deep Learning for Prediction and Prospective Validation of Functional Groups. *Chem. Sci.* **2020**, *11* (18), 4618–4630. <https://doi.org/10.1039/c9sc06240h>.
- (89) Linstrom, P. J.; Mallard, W. G. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*.
- (90) Stehman, S. V. Selecting and Interpreting Measures of Thematic Classification Accuracy. *Remote Sens. Environment* **1997**, *62*, 77–89.
- (91) Heath, C.; Myers, M.; Pejčić, B. The Effect of Pressure and Temperature on Mid-Infrared Sensing of Dissolved Hydrocarbons in Water. *Anal. Chem.* **2017**, *89* (24), 13391–13397. <https://doi.org/10.1021/acs.analchem.7b03623>.
- (92) Wade, L.; Simek, J. *Organic Chemistry*; Pearson, 2016.
- (93) NOAA. *What is the carbon cycle?* <https://oceanservice.noaa.gov/facts/carbon-cycle.html> (accessed 2021-12-15).
- (94) Taub, D. Effects of Rising Atmospheric Concentrations of Carbon Dioxide on Plants. *Nat. Educ. Knowl.* **2010**, *3* (21).
- (95) EPA. *Global greenhouse gas emissions data*. <https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data> (accessed 2021-12-15).
- (96) Friedlingstein, P.; O’Sullivan, M.; Jones, M. W.; Andrew, R. M.; Hauck, J.; Olsen, A.; Peters, G. P.; Peters, W.; Pongratz, J.; Sitch, S.; Le Quéré, C.; Canadell, J. G.; Ciais, P.; Jackson, R. B.; Alin, S.; Aragão, L. E. O. C.; Arneeth, A.; Arora, V.; Bates, N. R.; Becker, M.; Benoit-Cattin, A.; Bittig, H. C.; Bopp, L.; Bultan, S.; Chandra, N.; Chevallier, F.; Chini, L. P.; Evans, W.; Florentie, L.; Forster, P. M.; Gasser, T.; Gehlen, M.; Gilfillan, D.; Gkritzalis, T.; Gregor, L.; Gruber, N.; Harris, I.; Hartung, K.; Haverd, V.; Houghton, R. A.; Ilyina, T.; Jain, A. K.; Joetzjer, E.; Kadono, K.; Kato, E.; Kitidis, V.; Korsbakken, J. I.; Landschützer, P.; Lefèvre, N.; Lenton, A.; Lienert, S.; Liu, Z.; Lombardozzi, D.; Marland, G.; Metz, N.; Munro, D. R.; Nabel, J. E. M. S.; Nakaoka, S. I.; Niwa, Y.; O’Brien, K.; Ono, T.; Palmer, P. I.; Pierrot, D.; Poulter, B.; Resplandy, L.; Robertson, E.; Rödenbeck, C.; Schwinger, J.; Séférian, R.; Skjelvan, I.; Smith, A. J. P.; Sutton, A. J.; Tanhua, T.; Tans, P. P.; Tian, H.; Tilbrook, B.; Van Der Werf, G.; Vuichard, N.; Walker, A. P.; Wanninkhof, R.; Watson, A. J.; Willis, D.; Wiltshire, A. J.; Yuan, W.; Yue, X.; Zaehle, S. Global Carbon Budget 2020. *Earth Syst. Sci. Data* **2020**, *12* (4), 3269–3340.

- (97) Ritchie, G. *Atmospheric Chemistry*; WSPC (Europe), 2017.
- (98) NOAA. *Carbon cycle greenhouse gases*. <https://gml.noaa.gov/ccgg/index.html> (accessed 2020-12-15).
- (99) Jiang, L. Q.; Carter, B. R.; Feely, R. A.; Lauvset, S. K.; Olsen, A. Surface Ocean PH and Buffer Capacity: Past, Present and Future. *Sci. Rep.* **2019**, *9* (1), 1–11. <https://doi.org/10.1038/s41598-019-55039-4>.
- (100) Jarvis, N. L.; Garrett, W. D.; Scheiman, M. A.; Timmons, C. O. Surface Chemical Characterization of Surface-Active Material in Seawater. *Limnol. Oceanogr.* **1967**, *12* (1), 88–96. <https://doi.org/10.4319/lo.1967.12.1.0088>.
- (101) Garrett, W. D. The Organic Chemical Composition of the Ocean Surface. *Deep-Sea Res. Oceanogr. Abstr.* **1967**, *14* (2). [https://doi.org/10.1016/0011-7471\(67\)90007-1](https://doi.org/10.1016/0011-7471(67)90007-1).
- (102) Barger, W. R.; Garrett, W. D.; Mollo-Christensen, E. L.; Ruggles, K. W. Effects of an Artificial Sea Slick upon the Atmosphere and the Ocean. *J. Appl. Meteorol.* **1970**, *9*, 396–400.
- (103) Cochran, R. E.; Jayarathne, T.; Stone, E. A.; Grassian, V. H. Selectivity Across the Interface: A Test of Surface Activity in the Composition of Organic-Enriched Aerosols from Bubble Bursting. *J. Phys. Chem. Lett.* **2016**, *7* (9), 1692–1696. <https://doi.org/10.1021/acs.jpcclett.6b00489>.
- (104) Rogers, M. M.; Neal, J. F.; Saha, A.; Algarni, A. S.; Hill, T. C. J.; Allen, H. C. The Ocean's Elevator: Evolution of the Air–Seawater Interface during a Small-Scale Algal Bloom. *ACS Earth Space Chem.* **2020**, *4* (12), 2347–2357. <https://doi.org/10.1021/acsearthspacechem.0c00239>.
- (105) Garrett, W. D. Stabilization of Air Bubbles at the Air-Sea Interface by Surface-Active Material. *Deep-Sea Res. Oceanogr. Abstr.* **1967**, *14* (6), 661–672. [https://doi.org/10.1016/S0011-7471\(67\)80004-4](https://doi.org/10.1016/S0011-7471(67)80004-4).
- (106) Graham, D. E.; Phillips, M. C. Proteins at Liquid Interfaces. II. Adsorption Isotherms. *J. Colloid Interface Sci.* **1979**, *70* (3), 415–426. [https://doi.org/10.1016/0021-9797\(79\)90049-3](https://doi.org/10.1016/0021-9797(79)90049-3).
- (107) Neuman, R. D. Stearic Acid and Calcium Stearate Monolayer Collapse. *J. Colloid Interface Sci.* **1976**, *56* (3), 505–510. [https://doi.org/10.1016/0021-9797\(76\)90117-X](https://doi.org/10.1016/0021-9797(76)90117-X).
- (108) 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>.
- (109) Rossel, P. E.; Bienhold, C.; Hehemann, L.; Dittmar, T.; Boetius, A. Molecular Composition of Dissolved Organic Matter in Sediment Porewater of the Arctic Deep-Sea Observatory HAUSGARTEN (Fram Strait). *Front. Mar. Sci.* **2020**, *7*.
- (110) 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>.



- (111) ARENZ, R. F.; LEWIS, W. M.; SAUNDERS, J. F. Determination of Chlorophyll and Dissolved Organic Carbon from Reflectance Data for Colorado Reservoirs. *Int. J. Remote Sens.* **1996**, *17* (8), 1547–1565. <https://doi.org/10.1080/01431169608948723>.
- (112) 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](https://doi.org/10.1007/10683826_5).
- (113) Kasprzak, P.; Padisák, J.; Koschel, R.; Krienitz, L.; Gervais, F. Chlorophyll a Concentration across a Trophic Gradient of Lakes: An Estimator of Phytoplankton Biomass? *Limnologia* **2008**, *38* (3–4), 327–338. <https://doi.org/10.1016/j.limno.2008.07.002>.
- (114) Longhurst, A. *Ecological Geography of the Sea*; Academic Press, 2006.
- (115) 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>.
- (116) Gaines, G. L. The Thermodynamic Equation of State for Insoluble Monolayers. I. Uncharged Films. *J. Chem. Phys.* **1978**, *69* (2), 924–930. <https://doi.org/10.1063/1.436608>.
- (117) 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>.
- (118) 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>.
- (119) Duffy, M. E.; Adams, C. M.; Homolka, K. K.; Neibauer, J. A.; Mayer, L. M.; Keil, R. G. Degradation of Diatom Protein in Seawater: A Peptide-Level View. *Front. Mar. Sci.* **2022**, *8*, 757245. <https://doi.org/10.3389/fmars.2021.757245>.
- (120) He, Y.; Sun, C.; Li, W.; Yang, G.-P.; Ding, H. Degradation of Lipids in Seasonal Hypoxic Seawater under Different Oxygen Saturation. *J. Oceanol. Limnol.* **2018**, *36* (5), 1570–1585. <https://doi.org/10.1007/s00343-018-7110-0>.
- (121) Benner, R.; Amon, R. M. W. The Size-Reactivity Continuum of Major Bioelements in the Ocean. *Annu. Rev. Mar. Sci.* **2015**, *7* (1), 185–205. <https://doi.org/10.1146/annurev-marine-010213-135126>.
- (122) 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>.

- (123) 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>.
- (124) Pham, D. Q.; O'Brien, R.; Fraund, M.; Bonanno, D.; Laskina, O.; Beall, C.; Moore, K. A.; Forestieri, S.; Wang, X.; Lee, C.; Sultana, C.; Grassian, V.; Cappa, C. D.; Prather, K. A.; Moffet, R. C. Biological Impacts on Carbon Speciation and Morphology of Sea Spray Aerosol. *ACS Earth Space Chem.* **2017**, *1* (9), 551–561. <https://doi.org/10.1021/acsearthspacechem.7b00069>.
- (125) 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>.
- (126) Duong-Ly, K. C.; Gabelli, S. B. *Salting out of Proteins Using Ammonium Sulfate Precipitation*; Elsevier Inc., 2014; Vol. 541. <https://doi.org/10.1016/B978-0-12-420119-4.00007-0>.
- (127) Sears, D. F.; Schulman, J. H. Influence of Water Structures on the Surface Pressure, Surface Potential, and Area of Soap Monolayers of Lithium, Sodium, Potassium, and Calcium. *J. Phys. Chem.* **1964**, *68* (12), 3529–3534. <https://doi.org/10.1021/j100794a015>.
- (128) Zhang, T.; Brantley, S. L.; Verreault, D.; Dhankani, R.; Corcelli, S. A.; Allen, H. C. Effect of PH and Salt on Surface PK a of Phosphatidic Acid Monolayers. *Langmuir* **2018**, *34* (1), 530–539. <https://doi.org/10.1021/acs.langmuir.7b03579>.
- (129) Carter-Fenk, K. A.; Allen, H. C. Collapse Mechanisms of Nascent and Aged Sea Spray Aerosol Proxy Films. *Atmosphere* **2018**, *9* (12). <https://doi.org/10.3390/ATMOS9120503>.
- (130) 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>.
- (131) Pakulski, J. D.; Benner, R. Abundance and Distribution of Carbohydrates in the Ocean. *Limnol. Oceanogr.* **1994**, *39* (4), 930–940. <https://doi.org/10.4319/lo.1994.39.4.0930>.
- (132) Gibson, G.; Weijer, W.; Jeffery, N.; Wang, S. Relative Impact of Sea Ice and Temperature Changes on Arctic Marine Production. *J. Geophys. Res. Biogeosciences* **2020**, *125* (7). <https://doi.org/10.1029/2019JG005343>.
- (133) NOAA. *What are phytoplankton?*
- (134) Biddanda, B.; Benner, R. Carbon, Nitrogen, and Carbohydrate Fluxes during the Production of Particulate and Dissolved Organic Matter by Marine Phytoplankton. *Limnol. Oceanogr.* **1997**, *42* (3), 506–518.
- (135) Kinsey, J. D.; Corradino, G.; Ziervogel, K.; Schnetzer, A.; Osburn, C. L. Formation of Chromophoric Dissolved Organic Matter by Bacterial Degradation of Phytoplankton-Derived Aggregates. *Front. Mar. Sci.* **2018**, *4*.
- (136) Legendre, L.; Michaud, J. Chlorophyll a to Estimate the Particulate Organic Carbon Available as Food to Large Zooplankton in the Euphotic Zone of Oceans.

- J. Plankton Res.* **1999**, *21* (11), 2067–2083.  
<https://doi.org/10.1093/plankt/21.11.2067>.
- (137) Fasham, M. J. R.; Sarmiento, J. L.; Slater, R. D.; Ducklow, H. W.; Williams, R. Ecosystem Behavior at Bermuda Station “s” and Ocean Weather Station “India”: A General Circulation Model and Observational Analysis. *Glob. Biogeochem. Cycles* **1993**, *7* (2), 379–415.
- (138) Manning, N. F.; Wang, Y. C.; Long, C. M.; Bertani, I.; Sayers, M. J.; Bosse, K. R.; Shuchman, R. A.; Scavia, D. Extending the Forecast Model: Predicting Western Lake Erie Harmful Algal Blooms at Multiple Spatial Scales. *J. Gt. Lakes Res.* **2019**, *45* (3), 587–595. <https://doi.org/10.1016/j.jglr.2019.03.004>.
- (139) Engel, A.; Sperling, M.; Sun, C.; Grosse, J.; Friedrichs, G. Organic Matter in the Surface Microlayer: Insights from a Wind Wave Channel Experiment. *Front. Mar. Sci.* **2018**, *5* (JUN). <https://doi.org/10.3389/fmars.2018.00182>.
- (140) Wakeham, S. G.; Lee, C.; Hedges, J. I.; Hernes, P. J.; Peterson, M. J. Molecular Indicators of Diagenetic Status in Marine Organic Matter. *Geochim. Cosmochim. Acta* **1997**, *61* (24), 5363–5369. [https://doi.org/10.1016/S0016-7037\(97\)00312-8](https://doi.org/10.1016/S0016-7037(97)00312-8).
- (141) Sarmiento, J. L.; Slater, R. D.; Fasham, M. J. R.; Ducklow, H. W.; Toggweiler, J. R.; Evans, G. T. A Seasonal Three-Dimensional Ecosystem Model of Nitrogen Cycling in the North Atlantic Euphotic Zone. *Glob. Biogeochem. Cycles* **1993**, *7* (2), 417–450.
- (142) Rideal, E. K. *An Introduction to Surface Chemistry*; Cambridge, 1926.
- (143) Matsuzaki, Y.; Fujita, I. In Situ Estimates of Horizontal Turbulent Diffusivity at the Sea Surface for Oil Transport Simulation. *Mar. Pollut. Bull.* **2017**, *117* (1–2), 34–40. <https://doi.org/10.1016/j.marpolbul.2016.10.026>.
- (144) Friedlingstein, P.; O’Sullivan, M.; Jones, M.; Andrew, R.; Hauck, J.; Olsen, A.; Peters, G.; Peters, W.; Pongratz, J.; Sitch, S.; Le Quéré, C.; Canadell, J.; Ciais, P.; Jackson, R.; Alin, S.; Aragão, L.; Arneeth, A.; Arora, V.; Bates, N.; Becker, M.; Benoit-Cattin, A.; Bittig, H.; Bopp, L.; Bultan, S.; Chandra, N.; Chevallier, F.; Chini, L.; Evans, W.; Florentie, L.; Forster, P.; Gasser, T.; Gehlen, M.; Gilfillan, D.; Gkritzalis, T.; Gregor, L.; Gruber, N.; Harris, I.; Hartung, K.; Haverd, V.; Houghton, R.; Ilyina, T.; Jain, A.; Joetzjer, E.; Kadono, K.; Kato, E.; Kitidis, V.; Korsbakken, J. I.; Landschützer, P.; Lefèvre, N.; Lenton, A.; Lienert, S.; Liu, Z.; Lombardozzi, D.; Marland, G.; Metzl, N.; Munro, D.; Nabel, J.; Nakaoka, S.-I.; Niwa, Y.; O’Brien, K.; Ono, T.; Palmer, P.; Pierrot, D.; Poulter, B.; Resplandy, L.; Robertson, E.; Rödenbeck, C.; Schwinger, J.; Séférian, R.; Skjelvan, I.; Smith, A.; Sutton, A.; Tanhua, T.; Tans, P.; Tian, H.; Tilbrook, B.; van der Werf, G.; Vuichard, N.; Walker, A.; Wanninkhof, R.; Watson, A.; Willis, D.; Wiltshire, A.; Yuan, W.; Yue, X.; Zaehle, S. Global Carbon Budget 2021. *Earth Syst. Sci. Data Discuss.* **2021**, No. November, 1–191. <https://doi.org/10.5194/essd-2020-286>.
- (145) 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>.

- (146) 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>.
- (147) Peixoto, J.; Oort, A. *Physics of Climate*, 1st ed.; Springer US.
- (148) *Acidity across the interface from the ocean surface to sea spray aerosol* | *PNAS*. <https://www.pnas.org/doi/full/10.1073/pnas.2018397118> (accessed 2022-07-28).
- (149) Romankevich, E. A.; Ljutsarev, S. V. Dissolved Organic Carbon in the Ocean. *Mar. Chem.* **1990**, *30* (C), 161–178. [https://doi.org/10.1016/0304-4203\(90\)90068-N](https://doi.org/10.1016/0304-4203(90)90068-N).
- (150) 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>.
- (151) Zäncker, B.; Bracher, A.; Röttgers, R.; Engel, A. Variations of the Organic Matter Composition in the Sea Surface Microlayer: A Comparison between Open Ocean, Coastal, and Upwelling Sites Off the Peruvian Coast. *Front. Microbiol.* **2017**, *8*, 2369. <https://doi.org/10.3389/fmicb.2017.02369>.
- (152) Tanoue, E. Detection of Dissolved Protein Molecules in Oceanic Waters. *Mar. Chem.* **1995**, *51* (3), 239–252. [https://doi.org/10.1016/0304-4203\(95\)00061-5](https://doi.org/10.1016/0304-4203(95)00061-5).
- (153) Thornton, D. C. O.; Brooks, S. D.; Chen, J. Protein and Carbohydrate Exopolymer Particles in the Sea Surface Microlayer (SML). *Front. Mar. Sci.* **2016**, *3* (AUG), 1–14. <https://doi.org/10.3389/fmars.2016.00135>.
- (154) Schiffer, J. M.; Luo, M.; Dommer, A. C.; Thoron, G.; Pendergraft, M.; Santander, M. V.; Lucero, D.; Pecora De Barros, E.; Prather, K. A.; Grassian, V. H.; Amaro, R. E. Impacts of Lipase Enzyme on the Surface Properties of Marine Aerosols. *J. Phys. Chem. Lett.* **2018**, *9* (14), 3839–3849. <https://doi.org/10.1021/acs.jpcclett.8b01363>.
- (155) Luo, M.; Dommer, A. C.; Schiffer, J. M.; Rez, D. J.; Mitchell, A. R.; Amaro, R. E.; Grassian, V. H. Surfactant Charge Modulates Structure and Stability of Lipase-Embedded Monolayers at Marine-Relevant Aerosol Surfaces. *Langmuir* **2019**, *35* (27), 9050–9060. <https://doi.org/10.1021/acs.langmuir.9b00689>.
- (156) Leck, C.; Bigg, E. K. Biogenic Particles in the Surface Microlayer and Overlaying Atmosphere in the Central Arctic Ocean during Summer. *Tellus B Chem. Phys. Meteorol.* **2005**, *57* (4), 305–316. <https://doi.org/10.3402/tellusb.v57i4.16546>.
- (157) 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>.
- (158) Wiesenburg, D. A.; Little, B. J. Synopsis of the Chemical/Physical Properties of Seawater. *Ocean Phys. Eng.* **1987**, *12* (3–4), 127–165.

- (159) Neurath, H.; Bull, H. B. The Surface Activity of Proteins. *Chem. Rev.* **1938**, *23* (3), 391–435. <https://doi.org/10.1021/cr60076a001>.
- (160) Wierenga, P. A.; Egmond, M. R.; Voragen, A. G. J.; de Jongh, H. H. J. The Adsorption and Unfolding Kinetics Determines the Folding State of Proteins at the Air-Water Interface and Thereby the Equation of State. *J. Colloid Interface Sci.* **2006**, *299* (2), 850–857. <https://doi.org/10.1016/j.jcis.2006.03.016>.
- (161) D'Imprima, E.; Floris, D.; Joppe, M.; Sánchez, R.; Grininger, M.; Kühlbrandt, W. Protein Denaturation at the Air-Water Interface and How to Prevent It. *eLife* **2019**, *8*, e42747. <https://doi.org/10.7554/eLife.42747>.
- (162) Alamdari, S.; Roeters, S. J.; Golbek, T. W.; Schmüser, L.; Weidner, T.; Pfaendtner, J. Orientation and Conformation of Proteins at the Air–Water Interface Determined from Integrative Molecular Dynamics Simulations and Sum Frequency Generation Spectroscopy. *Langmuir* **2020**, *36* (40), 11855–11865. <https://doi.org/10.1021/acs.langmuir.0c01881>.
- (163) Xiao, Y.; Konermann, L. Protein Structural Dynamics at the Gas/Water Interface Examined by Hydrogen Exchange Mass Spectrometry. *Protein Sci.* **2015**, *24* (8), 1247–1256. <https://doi.org/10.1002/pro.2680>.
- (164) Farber, P. J.; Mittermaier, A. Side Chain Burial and Hydrophobic Core Packing in Protein Folding Transition States. *Protein Sci.* **2008**, *17* (4), 644–651. <https://doi.org/10.1110/ps.073105408>.
- (165) Argos, P.; Rossmann, M. G.; Grau, U. M.; Zuber, H.; Frank, G.; Tratschin, J. D. Thermal Stability and Protein Structure. *Biochemistry* **1979**, *18* (25), 5698–5703. <https://doi.org/10.1021/bi00592a028>.
- (166) 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>.
- (167) Morrisett, J. D.; Pownall, H. J.; Gotto, A. M. Bovine Serum Albumin. Study of the Fatty Acid and Steroid Binding Sites Using Spin Labeled Lipids. *J. Biol. Chem.* **1975**, *250* (7), 2487–2494. [https://doi.org/10.1016/s0021-9258\(19\)41627-x](https://doi.org/10.1016/s0021-9258(19)41627-x).
- (168) Gew, L. T.; Misran, M. Albumin-Fatty Acid Interactions at Monolayer Interface. *Nanoscale Res. Lett.* **2014**, *9* (1), 1–6. <https://doi.org/10.1186/1556-276X-9-218>.
- (169) 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>.
- (170) Neal, J. F.; Zhao, W.; Grooms, A. J.; Flood, A. H.; Allen, H. C. Arginine-Phosphate Recognition Enhanced in Phospholipid Monolayers at Aqueous Interfaces. *J. Phys. Chem. C* **2018**, *122* (46), 26362–26371. <https://doi.org/10.1021/acs.jpcc.8b03531>.

- (171) 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.jpcelett.2c00231>.
- (172) Hubberten, U.; Lara, R. J.; Kattner, G. Amino Acid Composition of Seawater and Dissolved Humic Substances in the Greenland Sea. *Mar. Chem.* **1994**, *45* (1), 121–128. [https://doi.org/10.1016/0304-4203\(94\)90096-5](https://doi.org/10.1016/0304-4203(94)90096-5).
- (173) Noskov, B. A.; Mikhailovskaya, A. A.; Lin, S.-Y.; Loglio, G.; Miller, R. Bovine Serum Albumin Unfolding at the Air/Water Interface as Studied by Dilational Surface Rheology. *Langmuir* **2010**, *26* (22), 17225–17231. <https://doi.org/10.1021/la103360h>.
- (174) Yuan, G.; Kienzle, P. A.; Satija, S. K. Salting Up and Salting Down of Bovine Serum Albumin Layers at the Air–Water Interface. *Langmuir* **2020**, *36* (50), 15240–15246. <https://doi.org/10.1021/acs.langmuir.0c02457>.
- (175) Pedraz, P.; Montes, F. J.; Cerro, R. L.; Díaz, M. E. Characterization of Langmuir Biofilms Built by the Biospecific Interaction of Arachidic Acid with Bovine Serum Albumin. *Thin Solid Films* **2012**, *525*, 121–131. <https://doi.org/10.1016/j.tsf.2012.10.055>.
- (176) Meinders, M. B. J.; Van den Bosch, G. G. M.; De Jongh, H. H. J. Adsorption Properties of Proteins at and near the Air/Water Interface from IRRAS Spectra of Protein Solutions. *Eur. Biophys. J.* **2001**, *30* (4), 256–267. <https://doi.org/10.1007/s002490000124>.
- (177) Martin, A. H.; Meinders, M. B. J.; Bos, M. A.; Stuart, M. A. C.; Van Vliet, T. Adsorption Properties and Conformational Aspects of Proteins at the Air-Water Interface Measured by Infra-Red Reflection Absorption Spectrometry. *Food Colloids Biopolym. Mater.* **2003**, *284*, 226–233.
- (178) 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>.
- (179) De Jongh, H. H. J.; Kusters, H. A.; Kudryashova, E.; Meinders, M. B. J.; Trofimova, D.; Wierenga, P. A. Protein Adsorption at Air-Water Interfaces: A Combination of Details. *Biopolymers* **2004**, *74* (1–2), 131–135. <https://doi.org/10.1002/bip.20036>.
- (180) Meinders, M. B. J.; De Jongh, H. H. J. Limited Conformational Change of  $\beta$ -Lactoglobulin When Adsorbed at the Air-Water Interface. *Biopolym. - Biospectroscopy Sect.* **2002**, *67* (4–5), 319–322. <https://doi.org/10.1002/bip.10115>.
- (181) Ulaganathan, V.; Retzlaff, I.; Won, J. Y.; Gochev, G.; Gunes, D. Z.; Gehin-Delval, C.; Leser, M.; Noskov, B. A.; Miller, R.  $\beta$ -Lactoglobulin Adsorption Layers at the Water/Air Surface: 2. Dilational Rheology: Effect of PH and Ionic Strength. *Colloids Surf. Physicochem. Eng. Asp.* **2017**, *521*, 167–176. <https://doi.org/10.1016/j.colsurfa.2016.08.064>.

- (182) 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>.
- (183) Kang, B.; Tang, H.; Zhao, Z.; Song, S. Hofmeister Series: Insights of Ion Specificity from Amphiphilic Assembly and Interface Property. *ACS Omega* **2020**, *5* (12), 6229–6239. <https://doi.org/10.1021/acsomega.0c00237>.
- (184) Zhou, H. X. Interactions of Macromolecules with Salt Ions: An Electrostatic Theory for the Hofmeister Effect. *Proteins Struct. Funct. Genet.* **2005**, *61* (1), 69–78. <https://doi.org/10.1002/prot.20500>.
- (185) Grooms, A. J.; Neal, J. F.; Ng, K. C.; Zhao, W.; Flood, A. H.; Allen, H. C. Thermodynamic Signatures of the Origin of *Anti* -Hofmeister Selectivity for Phosphate at Aqueous Interfaces. *J. Phys. Chem. A* **2020**, *124* (27), 5621–5630. <https://doi.org/10.1021/acs.jpca.0c02515>.
- (186) Hyde, A. M.; Zultanski, S. L.; Waldman, J. H.; Zhong, Y.-L.; Shevlin, M.; Peng, F. General Principles and Strategies for Salting-Out Informed by the Hofmeister Series. *Org. Process Res. Dev.* **2017**, *21* (9), 1355–1370. <https://doi.org/10.1021/acs.oprd.7b00197>.
- (187) Zhang, Y.; Cremer, P. S. Interactions between Macromolecules and Ions: The Hofmeister Series. *Curr. Opin. Chem. Biol.* **2006**, *10* (6), 658–663. <https://doi.org/10.1016/j.cbpa.2006.09.020>.
- (188) Noskov, B.; Mikhailovskaya, A. Adsorption Kinetics of Globular Proteins and Protein/Surfactant Complexes at the Liquid–Gas Interface. *Soft Matter* **2013**, *9* (39), 9392. <https://doi.org/10.1039/c3sm51357b>.
- (189) Burrows, S. M.; Easter, R. C.; Liu, X.; Ma, P.-L.; Wang, H.; Elliott, S. M.; Singh, B.; Zhang, K.; Rasch, P. J. OCEANFILMS (Organic Compounds from Ecosystems to Aerosols: Natural Films and Interfaces via Langmuir Molecular Surfactants) Sea Spray Organic Aerosol Emissions – Implementation in a Global Climate Model and Impacts on Clouds. *Atmospheric Chem. Phys.* **2022**, *22* (8), 5223–5251. <https://doi.org/10.5194/acp-22-5223-2022>.
- (190) Mouget, J. L.; Dakhama, A.; Lavoie, M. C.; de la Noüe, J. Algal Growth Enhancement by Bacteria: Is Consumption of Photosynthetic Oxygen Involved? *FEMS Microbiol. Ecol.* **1995**, *18* (1), 35–43. [https://doi.org/10.1016/0168-6496\(95\)00038-C](https://doi.org/10.1016/0168-6496(95)00038-C).
- (191) Rogers, M. M.; Neal, J. F.; Saha, A.; Algarni, A. S.; Hill, T. C. J.; Allen, H. C. The Ocean’s Elevator: Evolution of the Air-Seawater Interface during a Small-Scale Algal Bloom. *ACS Earth Space Chem.* **2020**, *4* (12), 2347–2357. <https://doi.org/10.1021/acsearthspacechem.0c00239>.
- (192) Cheng, Y. S.; McDonald, J. D.; Kracko, D.; Irvin, C. M.; Zhou, Y.; Pierce, R. H.; Henry, M. S.; Bourdelais, A.; Naar, J.; Baden, D. G. Concentration and Particle Size of Airborne Toxic Algae (Brevetoxin) Derived from Ocean Red Tide Events. *Environ. Sci. Technol.* **2005**, *39* (10), 3443–3449. <https://doi.org/10.1021/es048680j>.

- (193) Urquhart, E. A.; Schaeffer, B. A.; Stumpf, R. P.; Loftin, K. A.; Werdell, P. J. A Method for Examining Temporal Changes in Cyanobacterial Harmful Algal Bloom Spatial Extent Using Satellite Remote Sensing. *Harmful Algae* **2017**, *67*, 144–152. <https://doi.org/10.1016/j.hal.2017.06.001>.
- (194) Wang, X.; Sultana, C. M.; Trueblood, J.; Hill, T. C. J.; Malfatti, F.; Lee, C.; Laskina, O.; Moore, K. A.; Beall, C. M.; McCluskey, C. S.; Cornwell, G. C.; Zhou, Y.; Cox, J. L.; Pendergraft, M. A.; Santander, M. V.; Bertram, T. H.; Cappa, C. D.; Azam, F.; DeMott, P. J.; Grassian, V. H.; Prather, K. A. Microbial Control of Sea Spray Aerosol Composition: A Tale of Two Blooms. *ACS Cent. Sci.* **2015**, *1* (3), 124–131. <https://doi.org/10.1021/acscentsci.5b00148>.
- (195) Hogan, S. J. Some Effects of Surface Tension on Steep Water Waves. *J. Fluid Mech.* **1979**, *91* (1), 167–180. <https://doi.org/10.1017/S0022112079000094>.
- (196) Bian, H.; Feng, R.; Xu, Y.; Guo, Y.; Wang, H. Increased Interfacial Thickness of the NaF, NaCl and NaBr Salt Aqueous Solutions Probed with Non-Resonant Surface Second Harmonic Generation (SHG). *Phys. Chem. Chem. Phys.* **2008**, *10* (32), 4920. <https://doi.org/10.1039/b806362a>.
- (197) 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>.
- (198) Moore, J. K.; Doney, S. C.; Kleypas, J. A.; Glover, D. M.; Fung, I. Y. An Intermediate Complexity Marine Ecosystem Model for the Global Domain. *Deep-Sea Res. Part II Top. Stud. Oceanogr.* **2001**, *49* (1–3), 403–462. [https://doi.org/10.1016/S0967-0645\(01\)00108-4](https://doi.org/10.1016/S0967-0645(01)00108-4).
- (199) Abrosimova, K. V.; Shulenina, O. V.; Paston, S. V. FTIR Study of Secondary Structure of Bovine Serum Albumin and Ovalbumin. *J. Phys. Conf. Ser.* **2016**, *769* (1), 1. <https://doi.org/10.1088/1742-6596/769/1/012016>.
- (200) Grdadolnik, J.; Maréchal, Y. Bovine Serum Albumin Observed by Infrared Spectrometry. I. Methodology, Structural Investigation, and Water Uptake. *Biopolym.-Biospectroscopy* **2001**, *62* (1), 40–53.
- (201) Furlan, P. Y.; Scott, S. A.; Peaslee, M. H. FTIR-ATR Study of PH Effects on Egg Albumin Secondary Structure. *Spectrosc. Lett.* **2007**, *40* (3), 475–482. <https://doi.org/10.1080/00387010701295950>.
- (202) 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>.
- (203) Rabe, M.; Kerth, A.; Blume, A.; Garidel, P. Albumin Displacement at the Air–Water Interface by Tween (Polysorbate) Surfactants. *Eur. Biophys. J.* **2020**, *49* (7), 533–547. <https://doi.org/10.1007/s00249-020-01459-4>.
- (204) Wellen Rudd, B. A.; Vidalis, A. S.; Allen, H. C. Thermodynamic: Versus Non-Equilibrium Stability of Palmitic Acid Monolayers in Calcium-Enriched Sea Spray



- Aerosol Proxy Systems. *Phys. Chem. Chem. Phys.* **2018**, *20* (24), 16320–16332. <https://doi.org/10.1039/c8cp01188e>.
- (205) Kudryashova, E. V. Reversible Self-Association of Ovalbumin at Air-Water Interfaces and the Consequences for the Exerted Surface Pressure. *Protein Sci.* **2005**, *14* (2), 483–493. <https://doi.org/10.1110/ps.04771605>.
- (206) Meinders, M. B. J.; Van den Bosch, G. G. M.; De Jongh, H. H. J. Adsorption Properties of Proteins at and near the Air/Water Interface from IRRAS Spectra of Protein Solutions. *Eur. Biophys. J.* **2001**, *30* (4), 256–267. <https://doi.org/10.1007/s002490000124>.
- (207) Seki, T.; Chiang, K.-Y.; Yu, C.-C.; Yu, X.; Okuno, M.; Hunger, J.; Nagata, Y.; Bonn, M. The Bending Mode of Water: A Powerful Probe for Hydrogen Bond Structure of Aqueous Systems. *J. Phys. Chem. Lett.* **2020**, *11* (19), 8459–8469. <https://doi.org/10.1021/acs.jpcelett.0c01259>.
- (208) Lad, M. D.; Birembaut, F.; Matthew, J. M.; Frazier, R. A.; Green, R. J. The Adsorbed Conformation of Globular Proteins at the Air/Water Interface. *Phys. Chem. Chem. Phys.* **2006**, *8* (18), 2179–2186. <https://doi.org/10.1039/b515934b>.
- (209) Lad, M. D.; Birembaut, F.; Frazier, R. A.; Green, R. J. Protein – Lipid Interactions at the Air / Water Interface. *PCCP* **2005**, *7*, 3478–3485.
- (210) Enders, A. A.; Elliott, S. M.; Allen, H. C. Carbon on the Ocean Surface: Temporal and Geographical Investigation. *ACS Earth Space Chem.* **2023**, *7* (2), 360–369. <https://doi.org/10.1021/acsearthspacechem.2c00248>.
- (211) 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>.
- (212) 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>.
- (213) 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>.
- (214) 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>.
- (215) 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>.
- (216) 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>.
- (217) 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>.
- (218) 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>.
- (219) 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>.
- (220) 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>.
- (221) 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>.
- (222) 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>.
- (223) 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>.
- (224) 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 CO<sub>2</sub> and H<sub>2</sub> over TiO<sub>2</sub>-Supported Re and Zeolite Catalysts: Machine-Learning-Assisted Catalyst Optimization. *ACS Catal.* **2021**, *11* (9), 5829–5838. <https://doi.org/10.1021/acscatal.0c05661>.
- (225) 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.jpcclett.1c03526>.
- (226) 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>.
- (227) 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>.
- (228) 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>.
- (229) 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>.
- (230) 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>.
- (231) Liu, L.; Song, B.; Zhang, S.; Liu, X. A Novel Principal Component Analysis Method for the Reconstruction of Leaf Reflectance Spectra and Retrieval of Leaf Biochemical Contents. *Remote Sens.* **2017**, *9* (11), 1–24.  
<https://doi.org/10.3390/rs9111113>.
- (232) Škrbić, B.; Đurišić-Mladenović, N.; Cvejanov, J. Principal Component Analysis of Trace Elements in Serbian Wheat. *J. Agric. Food Chem.* **2005**, *53* (6), 2171–2175.  
<https://doi.org/10.1021/jf0402577>.
- (233) 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>.
- (234) Akinpelu, A. A.; Ali, Md. E.; Owolabi, T. O.; Johan, M. R.; Saidur, R.; Olatunji, S. O.; Chowdbury, Z. A Support Vector Regression Model for the Prediction of Total Polyaromatic Hydrocarbons in Soil: An Artificial Intelligent System for Mapping Environmental Pollution. *Neural Comput. Appl.* **2020**, *32* (18), 14899–14908.  
<https://doi.org/10.1007/s00521-020-04845-3>.
- (235) 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>.
- (236) 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>.
- (237) 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>.

- (238) 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>.
- (239) 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>.
- (240) Dittmar, T.; Koch, B.; Hertkorn, N.; Kattner, G. A Simple and Efficient Method for the Solid-Phase Extraction of Dissolved Organic Matter (SPE-DOM) from Seawater: SPE-DOM from Seawater. *Limnol. Oceanogr. Methods* **2008**, *6* (6), 230–235. <https://doi.org/10.4319/lom.2008.6.230>.
- (241) Trevethan, R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Front. Public Health* **2017**, *5*.
- (242) Fine, J. A.; Rajasekar, A. A.; Jethava, K. P.; Chopra, G. Spectral Deep Learning for Prediction and Prospective Validation of Functional Groups. *Chem. Sci.* **2020**, *11* (18), 4618–4630. <https://doi.org/10.1039/c9sc06240h>.

## Appendix A. Computational Methods Utilized in Chapter 3 for FTIR Functional Group Analysis

### A.1. Obtaining Spectra

*Retrieving the FTIR spectra:* A web scraping implementation was developed in Selenium to retrieve the FTIR spectra from the NIST Chemistry WebBook. Spectra were downloaded using the CAS number identifier in the jcamp-dx format and SMILES keys for each of the downloaded spectra were saved separately in a text file. With RDKit Python implementation, the functional groups were parsed from the spectrum's associated key.

### A.2. Spectral Processing

After downloading the FTIR data from NIT, the following processes are completed by running the preprocess\_subprocess.py script. To run this script after downloading from our GitHub, the following command line prompt should be used: "python preprocess\_subprocess.py". The script will complete eleven processes necessary to completing the training of the models.

#### A.2.1 Creating Directories on Computer

The functional group directories are created if not present in the working directory at the beginning of the subprocess script.

#### A.2.2. Removing Spectra not in Absorbance or Wavenumbers

Using the jcamp reader from [github.com/nzhagen/jcamp](https://github.com/nzhagen/jcamp), each downloaded FTIR spectrum is opened and read to confirm the x-axis of the trace is in wavenumber and the y-axis is in absorbance. Any spectrum that does not meet these conditions is removed from the directory. The subprocess calls on the script “`check_file_in_absorbance.py`”.

#### A.2.3. Convert from ‘jcamp-dx’ to ‘csv’

Spectra in absorbance and wavenumbers are converted from jcamp-dx to csv using the jcamp reader. The subprocess calls upon the “`jcamp_to_csv.py`” script and completes the conversion while maintaining the original jcamp file and creating a csv file.

#### A.2.4. Move ‘csv’ Files

Using the script “`move_file.py`”, the csv files are moved to a new directory.

#### A.2.5. Normalize ‘csv’ Files

Each spectrum is normalized with respect to the most intense peak in the spectrum using the “`normalize_csv.py`” script. After normalizing, a new csv file is saved to preserve the unnormalized spectrum.

#### A.2.6. Convert ‘csv’ to ‘jpg’ and move Spectra Images

With the script “`convert_to_jpg.py`”, the normalized csv files of the spectra are converted to jpg images of the spectra. The ML method used is chosen for the image-based

approach; the spectra are plotted as they would be for a chemist's analysis. Images are moved from the directory containing the csv files for separation in the following step.

#### A.2.7. Copy Spectra Images to Functional Group Directories

Using the SMARTS key, the functional groups for each spectrum were determined. A spectrum of a compound containing a functional group is copied to the directory for that functional group. For example, an alcohol-containing compound's spectra is copied into the alcohol containing directory. If a compound does not contain a functional group, the spectrum is copied to the not-containing directory.

#### A.2.7. Separation of Test Images and Setting Equivalent Examples per Class

Five spectra from each of the containing and not containing functional groups directories are moved to a directory within the functional group directory for testing of the models after training and validation. After, the directory with more spectral images is reduced to have an equal number of images as the smaller directory. Spectra are randomly deleted from the larger directory to achieve equal examples per class.

#### A.3. Model Training

Running "train\_ml\_subprocess.py" in the command line will train the models. Functional group models are trained using the subprocess script to call upon and train the modified Inception V3 network for each group. The training occurs over 20,000 steps with a learning rate of 0.01. Results are saved in the functional group's respective directory.

## A.4. Analysis of Models

Trained model graphs and logs are opened via Tensorboard. The accuracies and cross entropies are retrieved for analysis.

### A.4.1. Classify

Running “python classify\_subprocess.py” in the command line classifies the segmented test images that were not used for training. The results are saved in a csv file as an accessible format.

### A.4.2. Pearson’s Correlation Coefficient

The Pearson’s correlation coefficient is used to determine if there is a linear relationship between the number of spectra in a class and the accuracy and cross entropy in training or validation of the model. The coefficient was calculated using the SciPy “pearsonr” implementation.

### A.4.3. Plotting

All plots, pie charts, and confusion matrices are plotted using Matplotlib PyPlot implementations.



## Appendix B. Inception V3 Architecture and Optimization Functions Utilized in Chapter 3 for FTIR Analysis

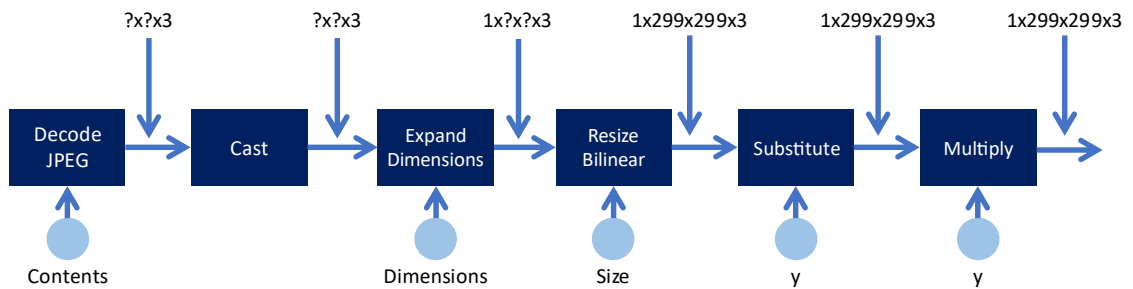


Figure 28. Overview of preprocessing steps included in model training. Dimensions of the output jpeg image file after each step.

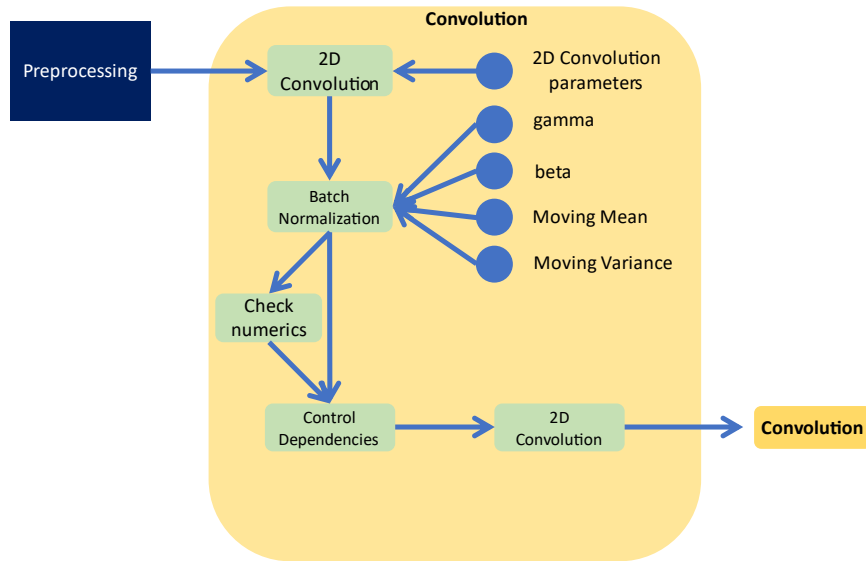


Figure 29. Summary of convolutional layer; gamma and beta are weights for the neuron nodes. Additionally, circles denote constant parameters (may be adjusted before or after but remain constant during the step with the arrow pointed at it.)

Stochastic gradient descent is given as

$$w_{k+1} = w_k - \alpha \nabla f(w_k)$$

Equation 47

Weights ( $w$ ) are adjusted in the direction of the negative gradient and decay,  $\alpha$ , has a commonly used value of 0.9.

Momentum is given as

$$z_{k+1} = \beta z_k + \nabla f(w_k)$$

Equation 48

where  $w$  is adjusted via

$$w_{k+1} = w_k - \alpha z_{k+1}$$

Equation 49

and weights are adjusted similarly to stochastic gradient descent and an additional component is added in the direction of the updated weight.

The dynamics can be written as

$$w_{k+1} = w_k - \alpha \nabla f(w_k) + \beta (w_k - w_{k-1})$$

Equation 50

where the momentum,  $\beta$ , has a commonly applied value of 0.9.

“RMSProp” uses the relationship

$$g_{k+1}^{-2} = \alpha g_k^{-2} + (1 - \alpha) g_k^2$$

Equation 51

where decay,  $\alpha$ , is set to 0.9. Momentum,  $\beta$ , is set to 0.9.

The described relationships can be written to give  $w$  as

$$w_{k+1} = \beta w_k + \frac{\eta}{\sqrt{g_{k+1}^{-2} + c}} \nabla f(w_k)$$

Equation 52

## Appendix C. Global carbon maps and additional figures from Chapter 4

### C.1. Maps of Surface Concentrations for Proteins and Lipids, Fractional Surface Coverage, and Non-normalized SSnL Carbon

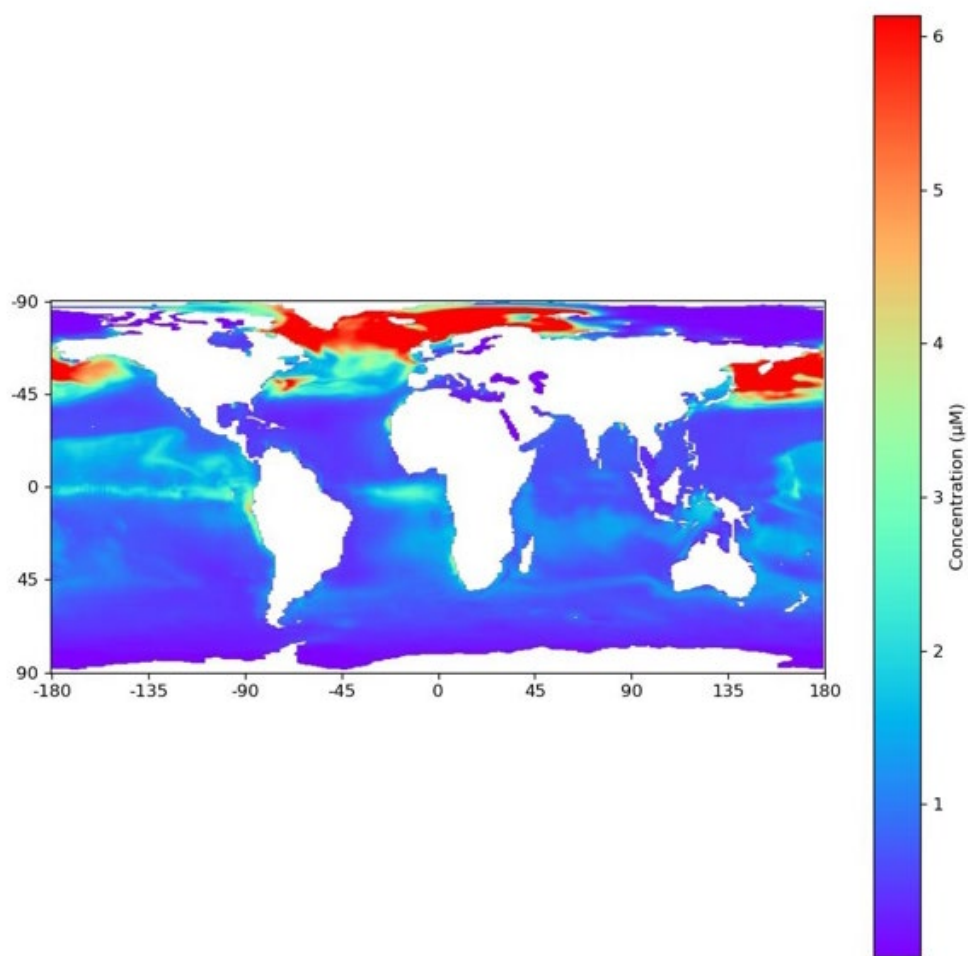


Figure 30. Modeled SSML lipid concentrations ( $\mu\text{M}$ ) for May 2005.

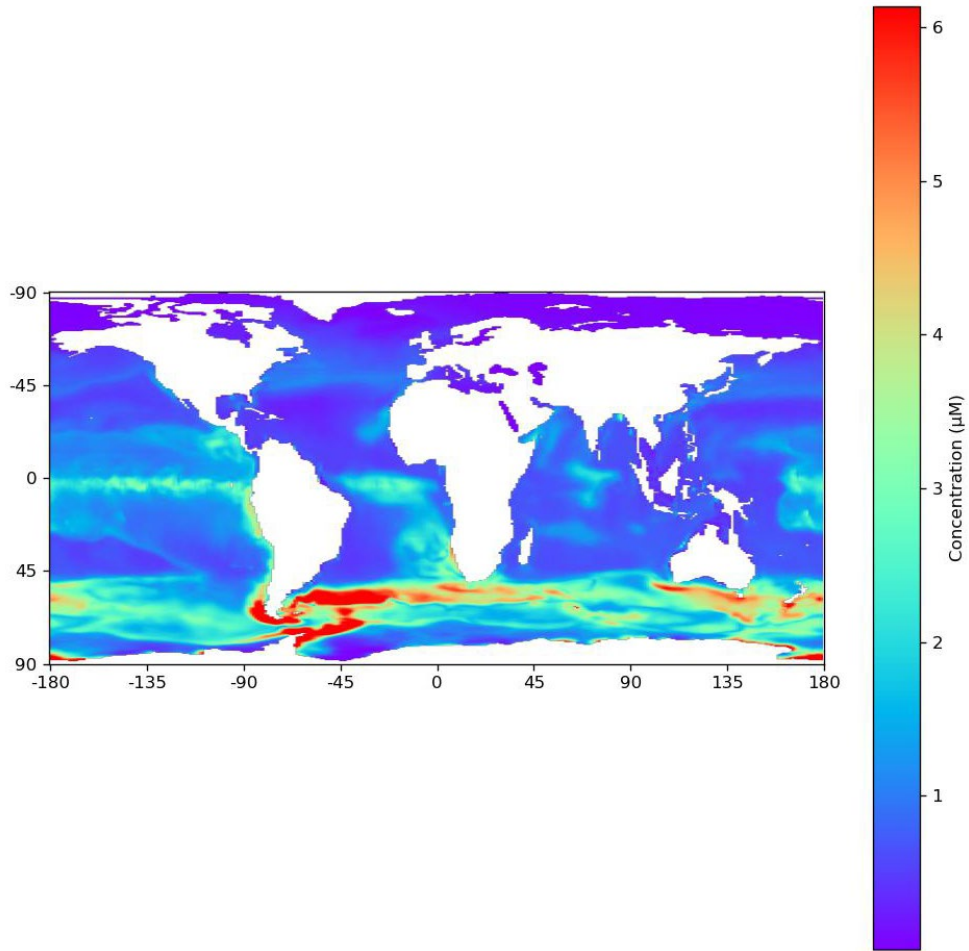


Figure 31. Modeled SSML lipid concentrations ( $\mu\text{M}$ ) for November 2005.

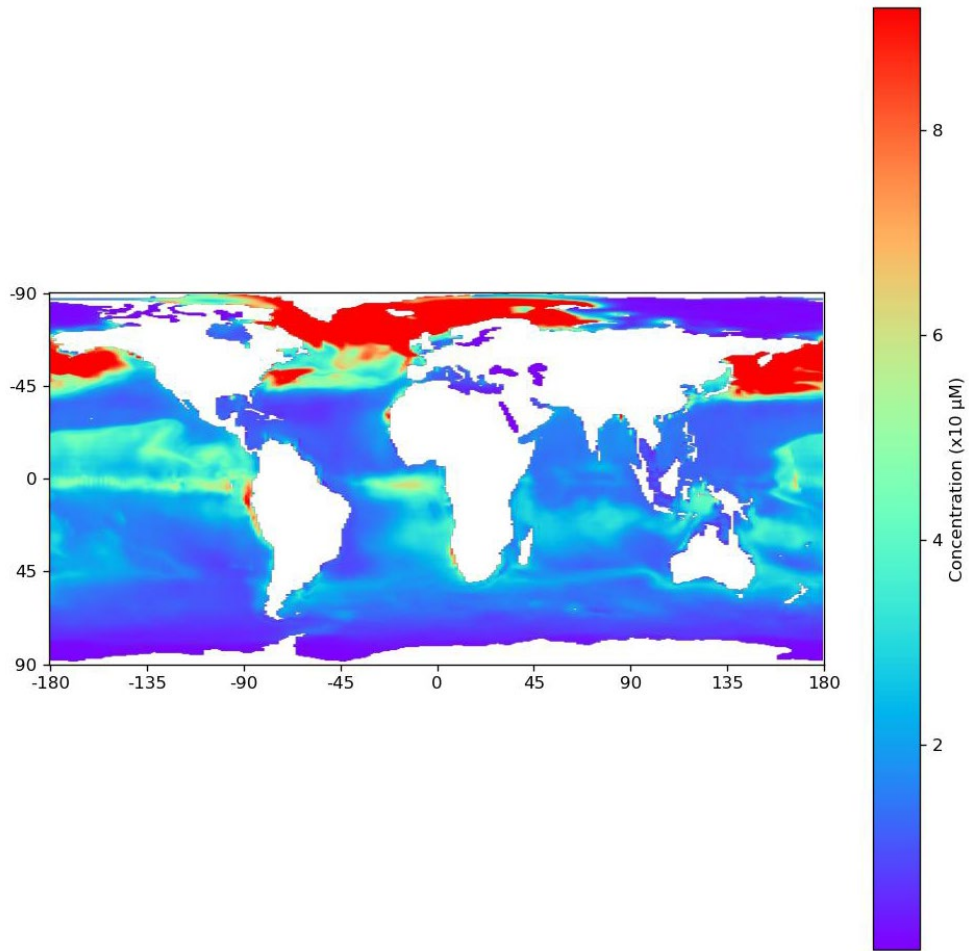


Figure 32. Modeled SSML protein concentrations (x 10  $\mu\text{M}$ ) for May 2005.

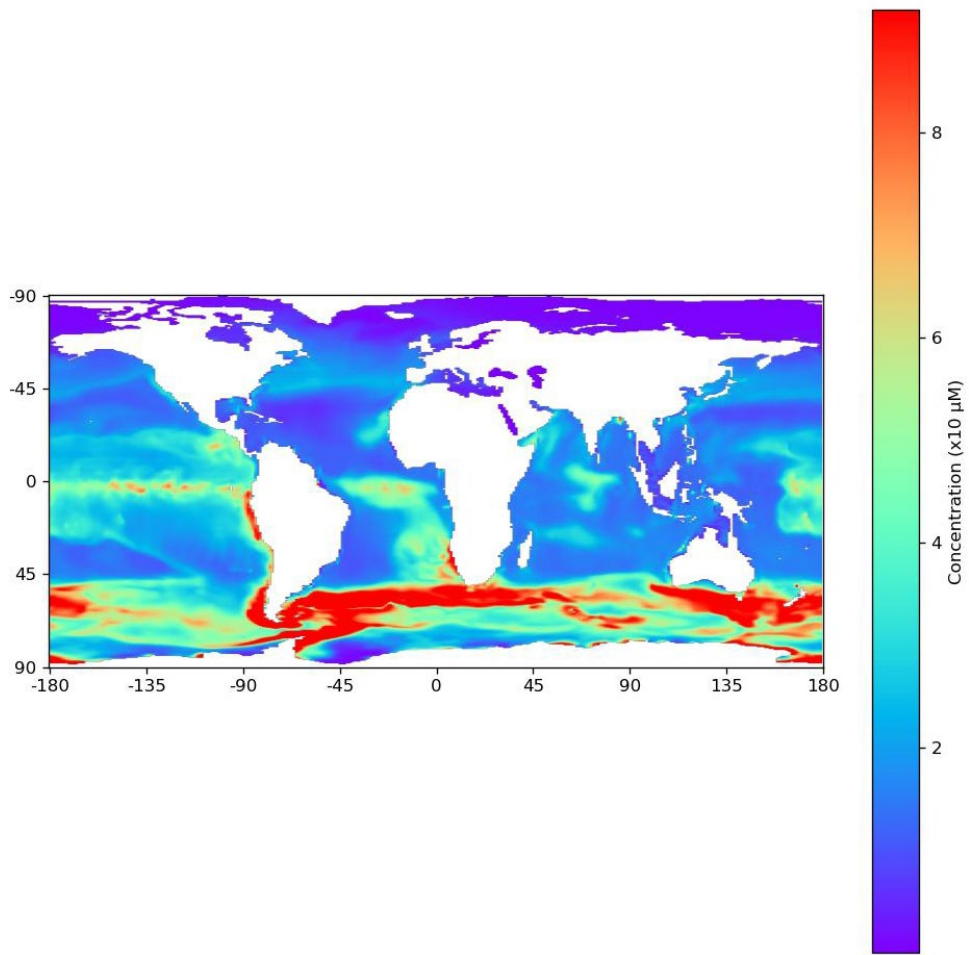


Figure 33. Modeled SSML protein concentrations (x 10  $\mu\text{M}$ ) for November 2005.

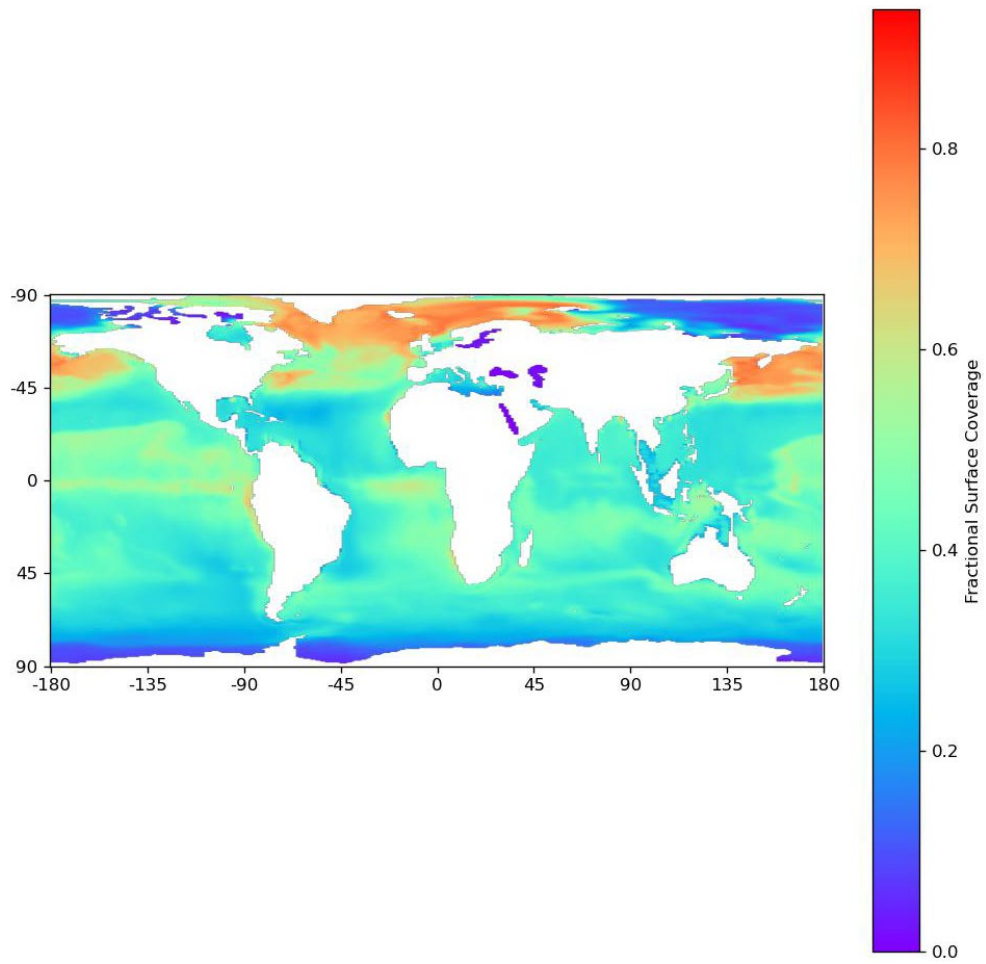


Figure 34. Modeled fractional surface coverage for May 2005.



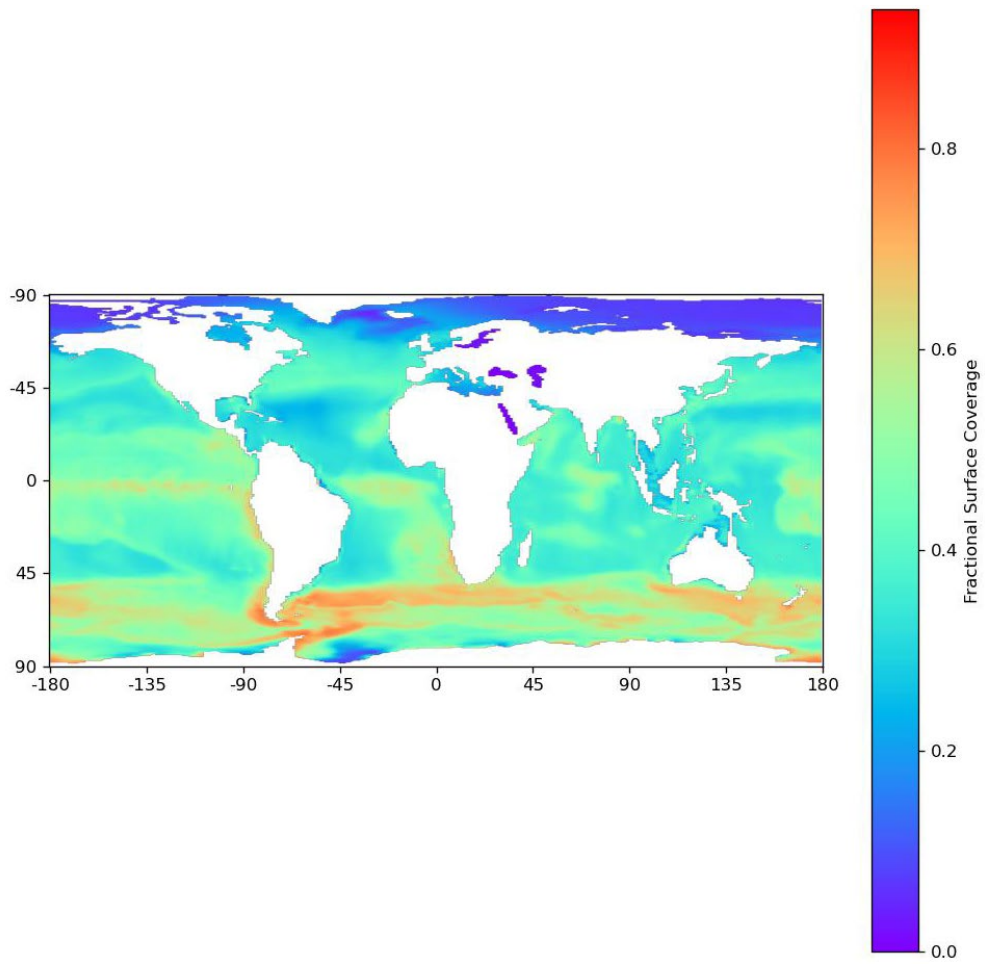


Figure 35. Modeled fractional surface coverage for November 2005.

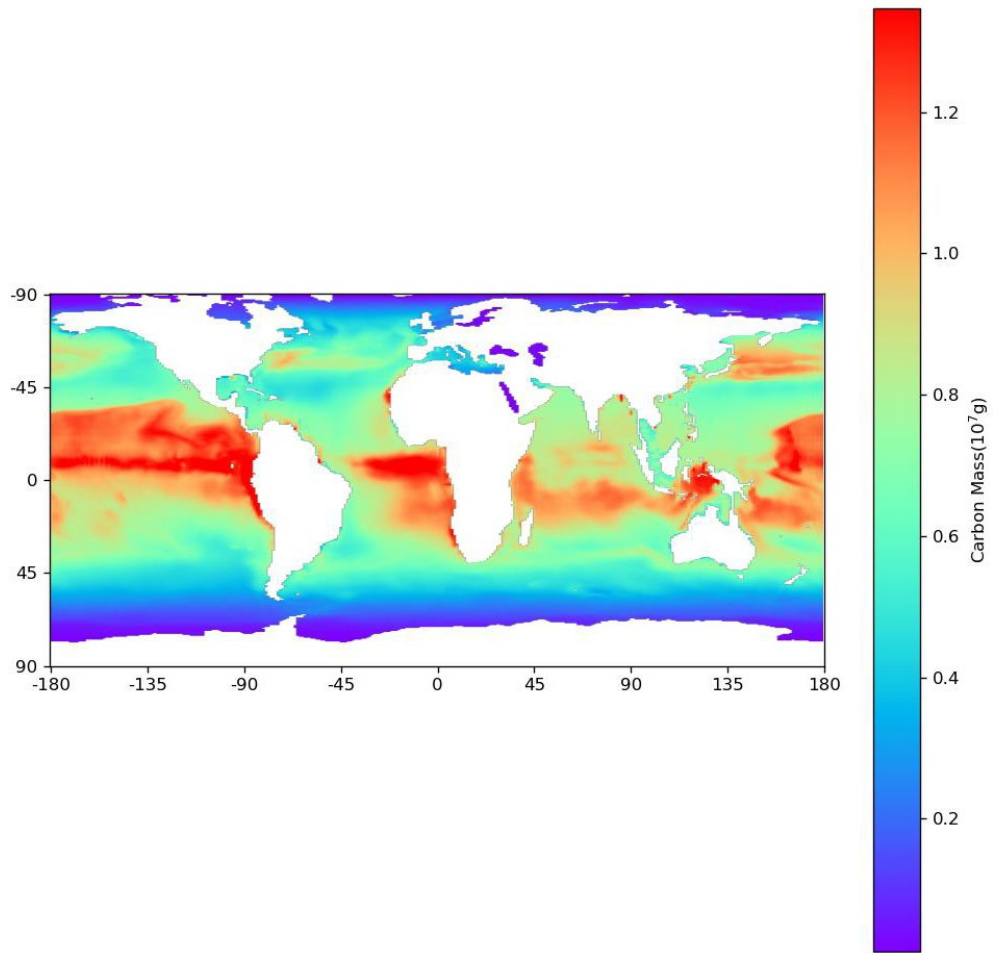


Figure 36. Modeled SSnL carbon mass ( $\times 10^7$  g) for May 2005.

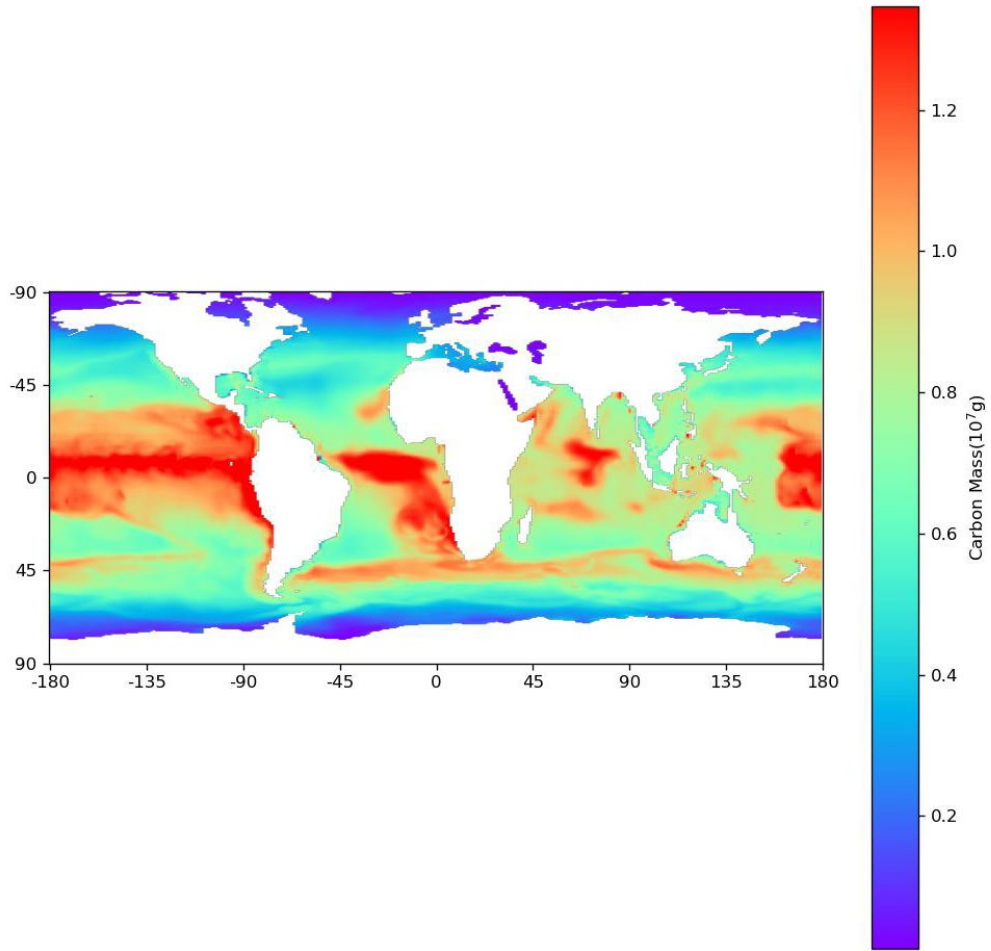


Figure 37. Modeled SSnL carbon mass ( $\times 10^7$  g) for November 2005.

## C.2. Longhurst Region Carbon Results

Table 11. Longhurst regional codes and the sum total carbon in the region for the months of May and November 2005.

<b>Longhurst Region</b>	<b>May</b>	<b>Nov</b>	<b>Difference</b>	<b>% Difference</b>
ALSK	6E+08	1E+09	-4E+08	-79%
ANTA	2E+10	1E+10	7E+09	39%
APLR	7E+09	4E+09	3E+09	46%
ARAB	4E+09	4E+09	3E+08	9%
ARCH	9E+09	9E+09	5E+08	5%
ARCT	3E+09	5E+09	-3E+09	-95%
AUSE	1E+09	1E+09	-2E+08	-21%
AUSW	2E+09	2E+09	-1E+08	-5%
BERS	4E+09	6E+09	-3E+09	-75%
BRAZ	1E+09	1E+09	-9E+06	-1%
CAMR	9E+08	1E+09	-1E+08	-11%
CARB	2E+09	3E+09	-5E+08	-22%
CCAL	2E+09	3E+09	-5E+08	-20%
CHIL	2E+09	2E+09	2E+08	9%
CHIN	3E+08	2E+08	5E+07	16%
CNRY	7E+08	8E+08	-7E+07	-9%
EAFR	1E+09	1E+09	-3E+08	-22%
ETRA	6E+09	6E+09	5E+08	8%
FKLD	1E+09	1E+09	2E+08	16%
GFST	3E+08	3E+08	-1E+07	-4%
GUIA	4E+08	5E+08	-8E+07	-21%
GUIN	9E+08	8E+08	3E+07	4%
INDE	1E+09	9E+08	2E+08	16%
INDW	1E+09	9E+08	9E+07	9%
ISSG	1E+10	1E+10	-6E+08	-5%
KURO	1E+09	1E+09	2E+08	15%
MEDI	2E+09	2E+09	-2E+08	-13%
MONS	1E+10	2E+10	-1E+09	-7%
NADR	4E+09	5E+09	-2E+09	-53%
NASE	4E+09	5E+09	-2E+08	-4%
NASW	2E+09	2E+09	-1E+08	-6%

NATR	4E+09	4E+09	-3E+07	-1%
NECS	1E+09	3E+09	-1E+09	-83%
NEWZ	1E+09	1E+09	2E+08	17%
NPPF	7E+09	7E+09	5E+08	7%
NPSW	1E+10	1E+10	2E+09	12%
NPTG	1E+10	2E+10	-1E+09	-10%
PEQD	2E+10	2E+10	-1E+09	-7%
PNEC	1E+10	1E+10	-1E+09	-10%
PSAE	3E+09	5E+09	-2E+09	-57%
PSAW	3E+09	3E+09	-4E+08	-15%
REDS	8E+08	6E+08	1E+08	15%
SANT	2E+10	2E+10	4E+09	20%
SARC	2E+09	3E+09	-1E+09	-80%
SATL	9E+09	9E+09	-3E+08	-3%
SPSG	4E+10	4E+10	2E+09	4%
SSTC	1E+10	1E+10	6E+07	1%
SUND	5E+09	5E+09	3E+08	5%
TASM	1E+09	1E+09	3E+07	2%
WARM	1E+10	1E+10	-3E+08	-2%
WTRA	5E+09	6E+09	-8E+08	-15%

### C.3. Normalized SSnL Carbon Maps for the Year of 2005

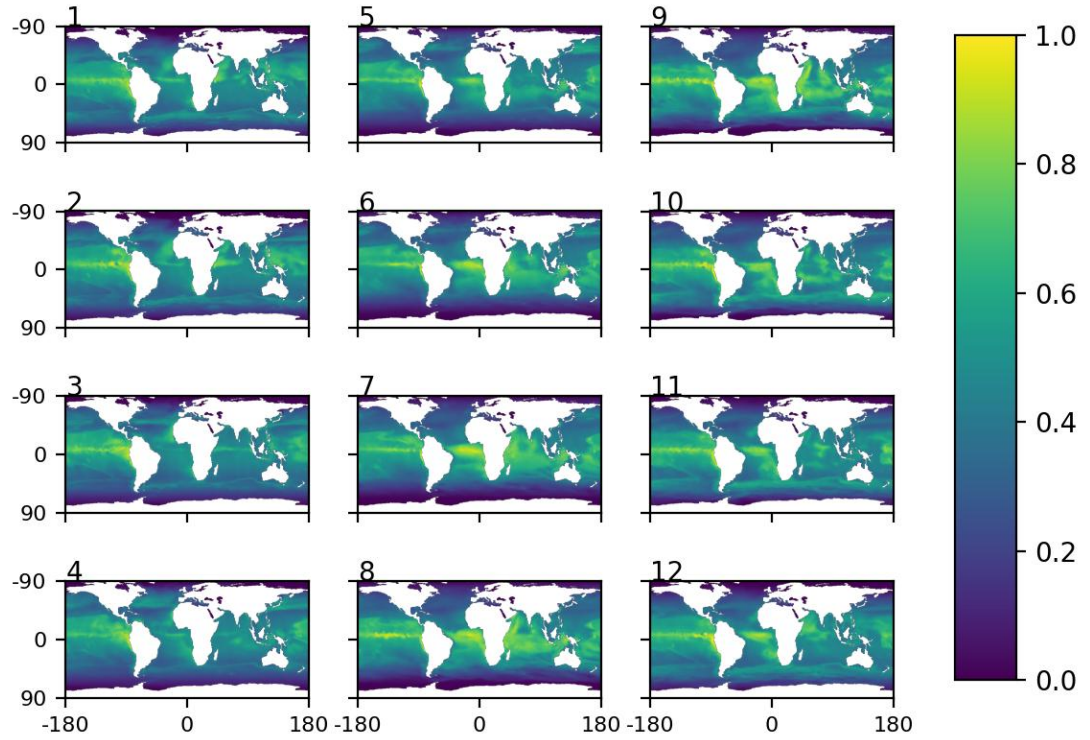


Figure 38. Modeled SSnL carbon normalized to the highest observed mass over all months for January ('1') through December ('12') in 2005 calculated from E3SM output.

**Appendix D. Attenuated Total Reflectance FTIR Absorbance Analysis and Details  
Regarding Pathlength Variability Discussed in Chapter 5**

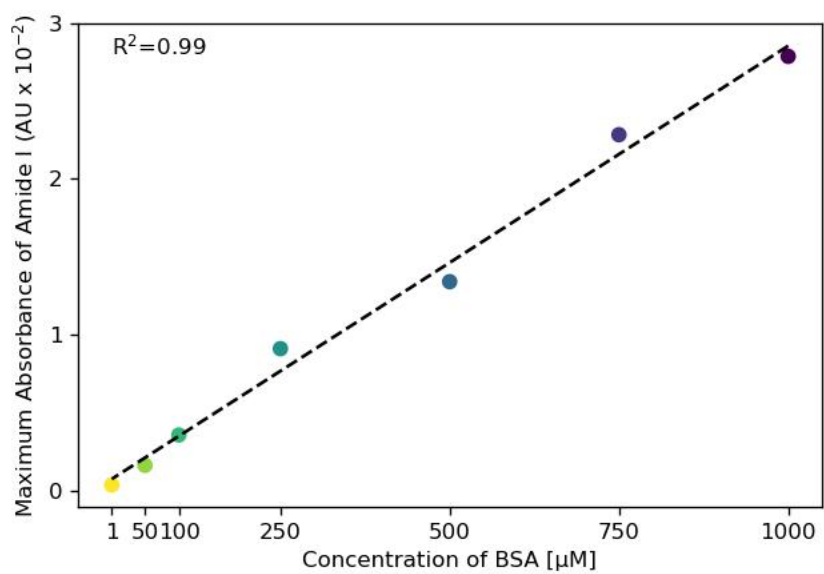


Figure 39. Maximum observed absorbance in amide I region (1653 cm<sup>-1</sup>) as a function of concentration. The dashed black line is a linear fit with an R<sup>2</sup> value of 0.99 (inset).

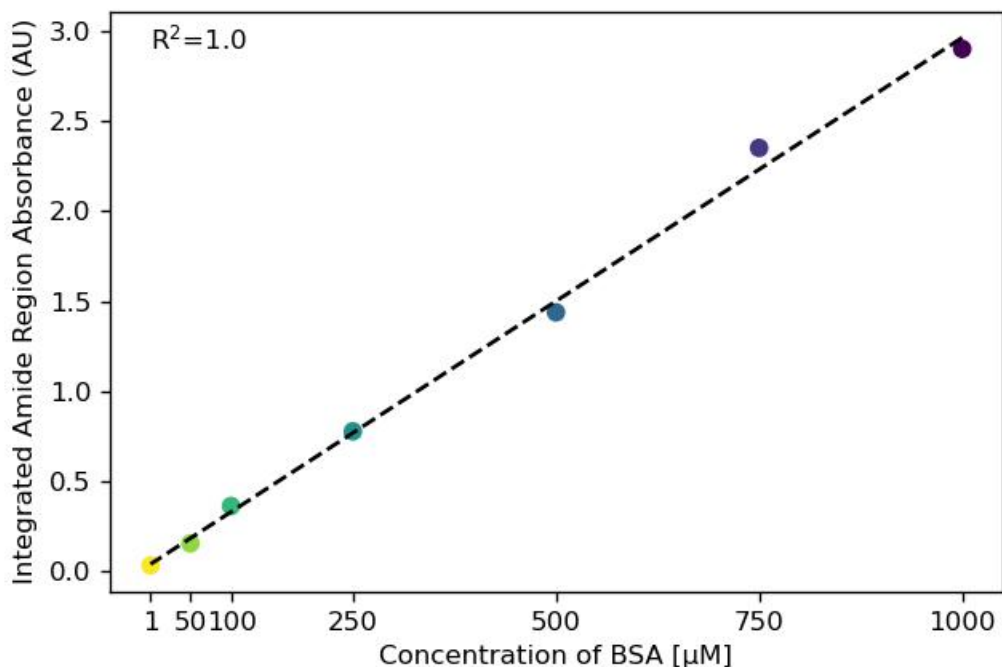


Figure 40. Integrated peak area for amide region is linear with increasing concentration.

Attenuated Total Reflectance (ATR) measurements are an accessible and simple sampling method to obtain FTIR spectra. However, the limitation of ATR includes an unknown and variable path length of the IR measurement. The pathlength in ATR is variable over all wavenumbers and changes based on the refractive index of the sample and crystal, or internal reflection element (IRE). The pathlength is also polarization dependent. Yet, our observed maximum absorbance in the amide region is at relatively low AU values (well below 1). The lack of strong absorption (well above 1 AU) should provide reliable and consistent pathlengths at a given wavenumber. Averett and colleagues describe in significant detail the computations necessary to determine the pathlength.<sup>34</sup>



For our experiment, we assert that the low absorptivity and linearity observed in both the maximum peak absorbance and integrated peak area as a function of concentration corresponds to a consistent pathlength. The data indicate Beer's Law is followed here and our bulk measurements via ATR are reasonably acceptable.

Bovine serum albumin (BSA) has a known molar extinction coefficient of  $43,824 \text{ cm}^{-1}\text{M}^{-1}$ . Given the extinction coefficient, we confirm that the pathlength is consistent using  $A = \epsilon bC$ . The pathlength at  $1653 \text{ cm}^{-1}$  is  $7.2 \pm 0.9 \text{ }\mu\text{m}$ . We determine the standard deviation as the difference in calculated pathlength at each concentration, excluding  $1 \text{ }\mu\text{M}$  because of the negligible observed absorbance at this low concentration. It is worth noting that this also confirms the assertions from Averett and colleagues that the penetration depth (determined by the IRE) is not a sufficient substitute for effective pathlength because the penetration depth of a single reflection diamond ATR IRE is about  $2 \text{ }\mu\text{m}$ .

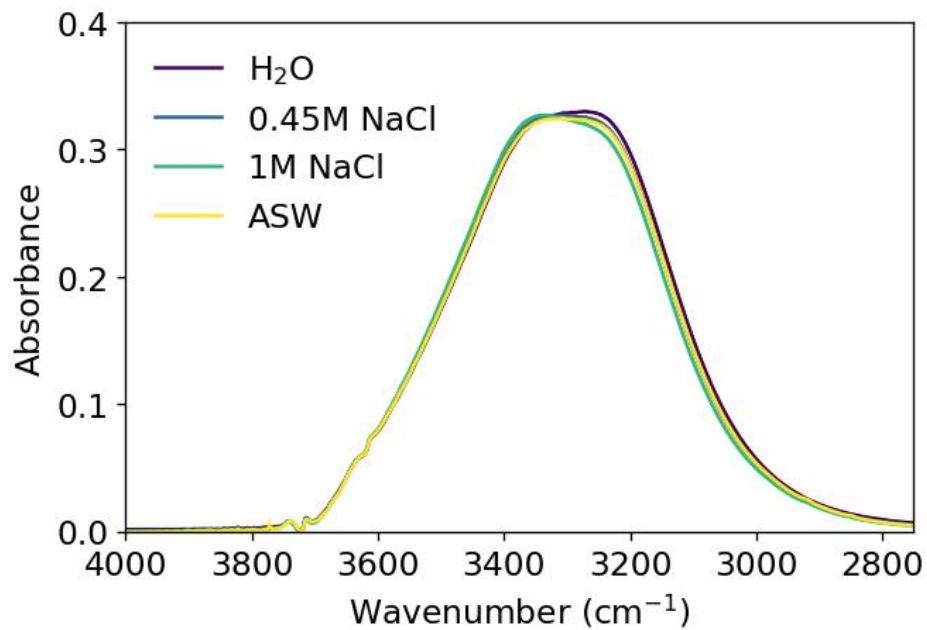


Figure 41. O-H stretching region for water, 0.45 M NaCl, 1 M NaCl, and artificial sea water (ASW). Standard deviation is shown and it is approximately the thickness of the line of the peak.

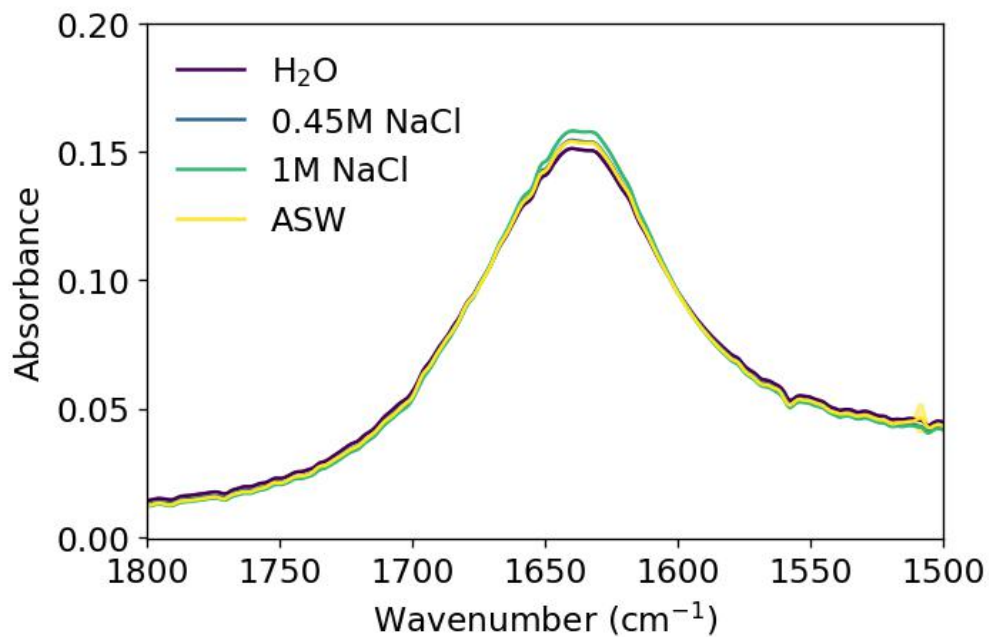


Figure 42. O-H bend of each solution. Standard deviation is shown and is approximately the thickness of the line.

**Appendix E. Bovine Serum Albumin Concentration Dependent IRRAS from  
Chapter 5**

Table 12. Ion/element concentrations for Instant Ocean from manufacturer.

<b>Ion</b>	<b>Instant Ocean (ppm)</b>
Cl <sup>-</sup>	19,290
Na <sup>+</sup>	10,780
SO <sub>4</sub> <sup>2-</sup>	2,660
Mg <sup>2+</sup>	1,320
K <sup>+</sup>	420
Ca <sup>2+</sup>	400
CO <sub>3</sub> <sup>2-</sup> /HCO <sub>3</sub> <sup>3-</sup>	200
Br <sup>-</sup>	56
Sr <sup>2+</sup>	8.8
B	5.6
F <sup>-</sup>	1.0
Li <sup>+</sup>	0.3
I <sup>-</sup>	0.24
Ba <sup>2+</sup>	Less than 0.04
Fe	Less than 0.04
Mn	Less than 0.025
Cr	Less than 0.015
Cu	Less than 0.015
Ni	Less than 0.015
Se	Less than 0.015
V	Less than 0.015
Zn	Less than 0.015
Mo	Less than 0.01
Al	Less than 0.006
Pb	Less than 0.005

<b>Ion (Cont.)</b>	<b>Instant Ocean (ppm) (Cont.)</b>
As	Less than 0.004
Cd	Less than 0.002
Nitrate	None
Phosphate	None

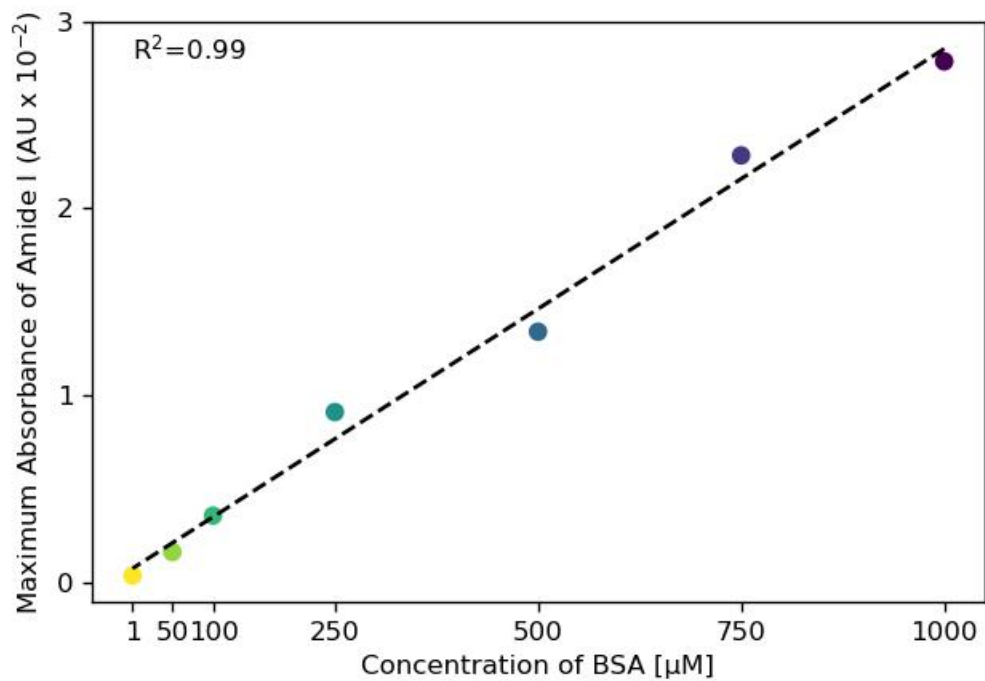


Figure 43. Maximum observed absorbance in amide I region (1653 cm<sup>-1</sup>) as a function of concentration. The dashed black line is a linear fit with an R<sup>2</sup> value of 0.99 (inset).

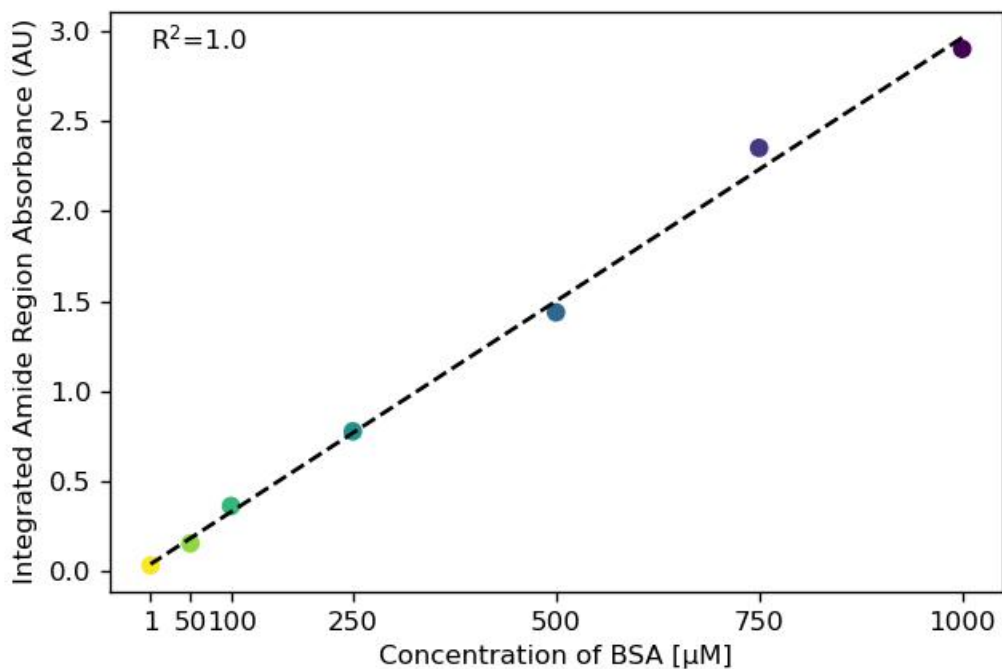


Figure 44. Integrated peak area for amide region is linear with increasing concentration.

Attenuated Total Reflectance (ATR) measurements are an accessible and simple sampling method to obtain FTIR spectra. However, the limitation of ATR includes an unknown and variable path length of the IR measurement. The pathlength in ATR is variable over all wavenumbers and changes based on the refractive index of the sample and crystal, or internal reflection element (IRE). The pathlength is also polarization dependent. Yet, our observed maximum absorbance in the amide region is at relatively low AU values (well below 1). The lack of strong absorption (well above 1 AU) should provide reliable and consistent pathlengths at a given wavenumber. Averett and colleagues describe in significant detail the computations necessary to determine the pathlength.<sup>34</sup>

For our experiment, we assert that the low absorptivity and linearity observed in both the maximum peak absorbance and integrated peak area as a function of concentration corresponds to a consistent pathlength. The data indicate Beer's Law is followed here and our bulk measurements via ATR are reasonably acceptable.

Bovine serum albumin (BSA) has a known molar extinction coefficient of  $43,824 \text{ cm}^{-1}\text{M}^{-1}$ . Given the extinction coefficient, we confirm that the pathlength is consistent using  $A = \epsilon bC$ . The pathlength at  $1653 \text{ cm}^{-1}$  is  $7.2 \pm 0.9 \text{ }\mu\text{m}$ . We determine the standard deviation as the difference in calculated pathlength at each concentration, excluding  $1 \text{ }\mu\text{M}$  because of the negligible observed absorbance at this low concentration. It is worth noting that this also confirms the assertions from Averett and colleagues that the penetration depth (determined by the IRE) is not a sufficient substitute for effective pathlength because the penetration depth of a single reflection diamond ATR IRE is about  $2 \text{ }\mu\text{m}$ .

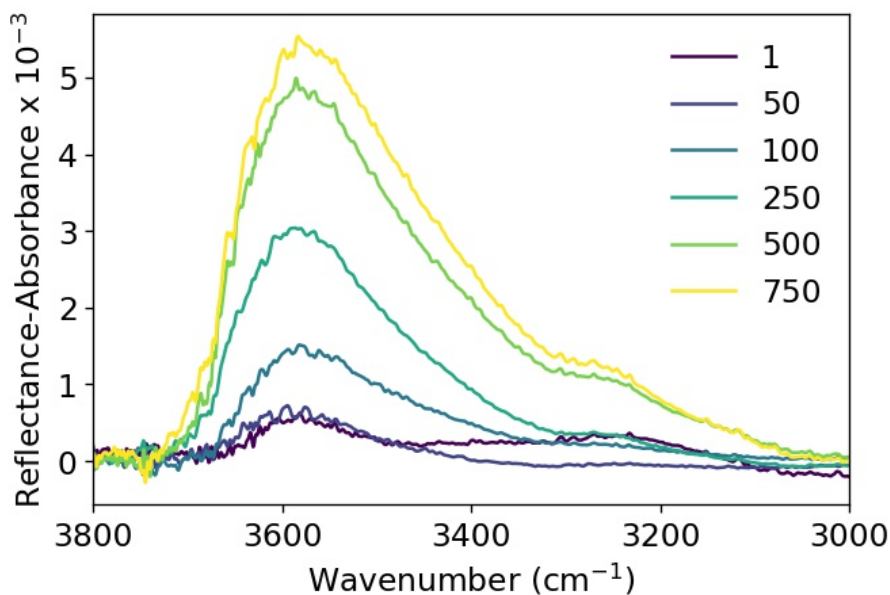


Figure 45. Background-corrected IRRAS showing O-H stretching region changes at variable BSA concentrations (given in  $\mu\text{M}$ ). Injections of 1 and 50  $\mu\text{M}$  solutions do not have a significant IR response as evidenced by the low intensity.

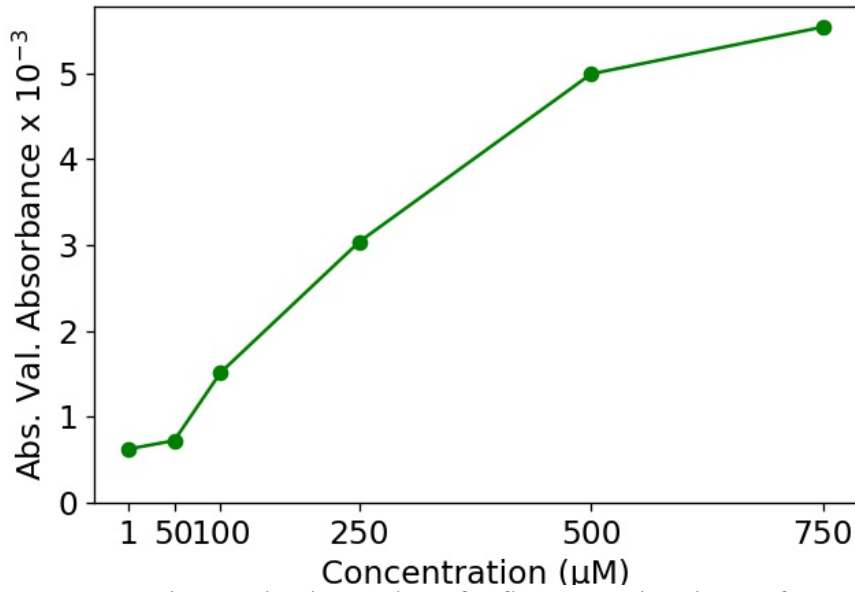


Figure 46. Maximum absolute value of reflectance-absorbance for O-H stretch at  $3585\text{ cm}^{-1}$ . While data for 1 and 50  $\mu\text{M}$  are presented, the conclusions that can be drawn from the observed IR response are limited because of the limit of detection.

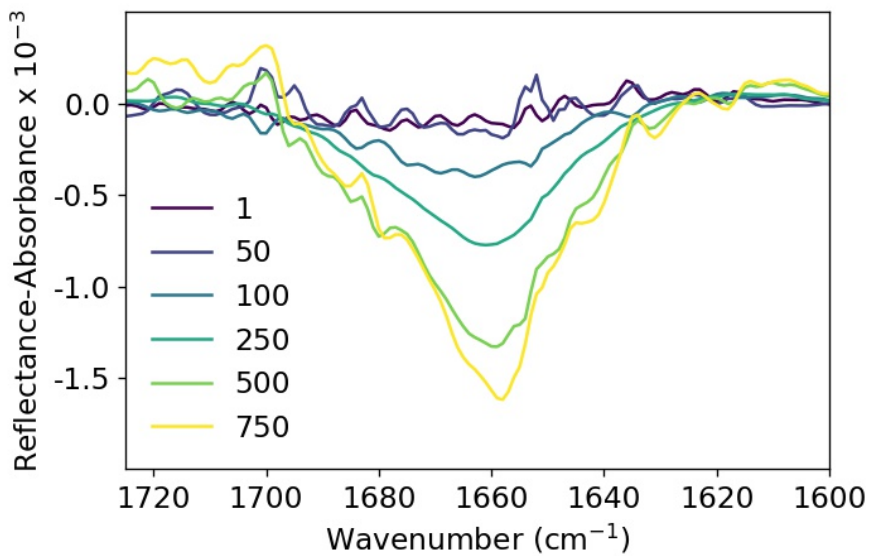




Figure 47. Background-corrected IRRAS showing amide I region changes at variable BSA concentrations (given in  $\mu\text{M}$ ). Injections of 1 and 50  $\mu\text{M}$  solutions do not have a significant IR response as evidenced by the low intensity.

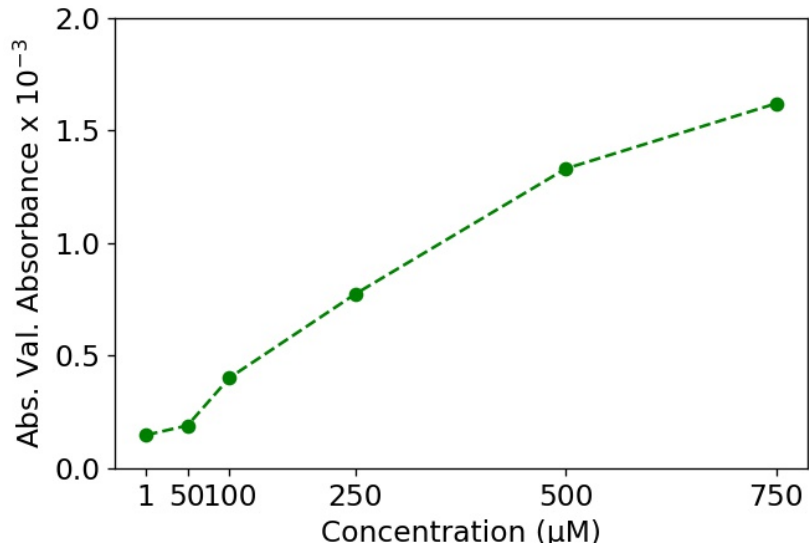


Figure 48. Maximum absolute value of reflectance-absorbance for amide I ( $\nu\text{C}=\text{O}$ ) at  $1640\text{ cm}^{-1}$ . While data for 1 and 50  $\mu\text{M}$  are presented, the conclusions that can be drawn from the observed IR response are limited because of the limit of detection.

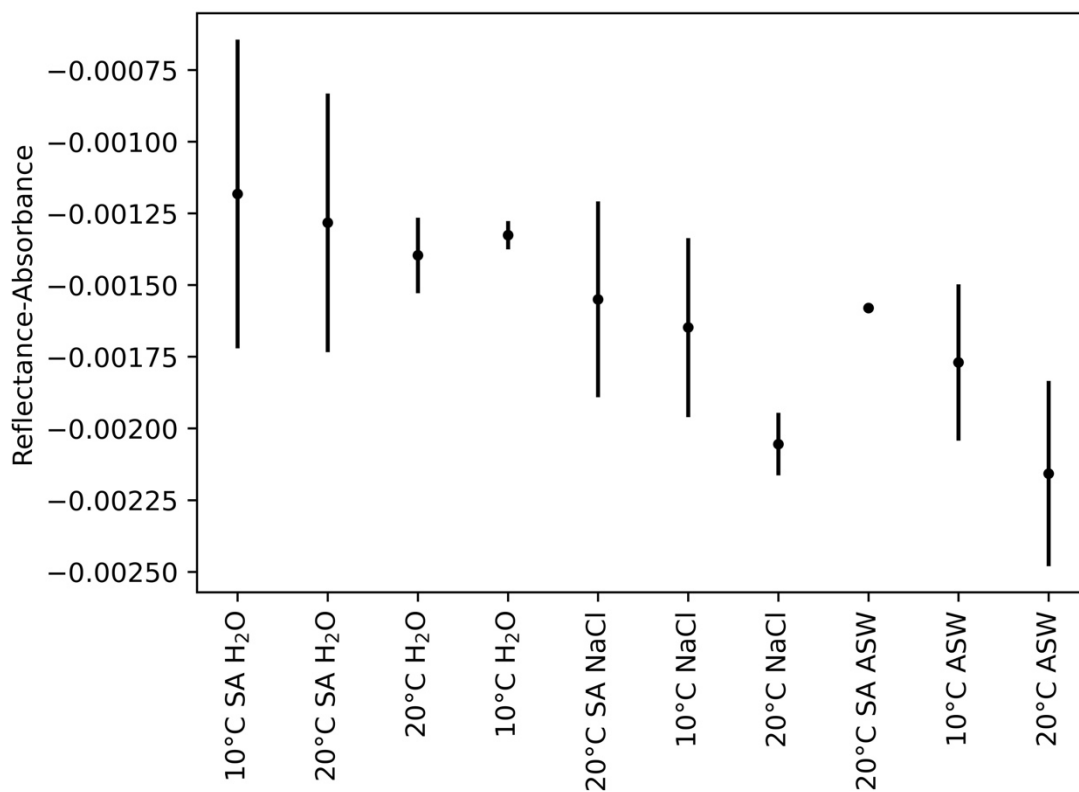


Figure 49. Standard deviation of RA for each solution at minimum peak intensity for amide I mode.

## **Appendix F. Explanation of Machine Learning Specifications, Plots of all Matrix Samples, Plots of Real Field Samples, and Optimization Results for SVR used in Chapter 6**

We utilize principal component analysis (PCA) as an unsupervised method and its prominence in chemistry applications. Linear regression (LR) and support vector regression (SVR) models are chosen for qualitative analysis. LR is a mathematically simple fit and relies on linear relationships of data, while SVR fits data to a chosen function and has tolerance boundaries.

While gas-phase spectra would not be expected to generate a model with predictive power for aqueous or liquid samples, there are examples in the literature where gas-phase, neat spectra training data produced models capable of accurately identifying components in liquid-phase, mixture spectra.<sup>242</sup> It is of interest to further evaluate if neat spectra can produce sufficient classification ML models because it would significantly reduce the amount of data needed for analyzing complex mixtures, such as those from the ocean's surface.

Table 13. Glucose concentration (M) of training samples.

	A	B	C	D	E	F	G	H	I	J
1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
2	0.09	0.18	0.27	0.36	0.45	0.54	0.63	0.72	0.81	0.9
3	0.08	0.16	0.24	0.32	0.4	0.48	0.56	0.64	0.72	0.8
4	0.07	0.14	0.21	0.28	0.35	0.42	0.49	0.56	0.63	0.7
5	0.06	0.12	0.18	0.24	0.3	0.36	0.42	0.48	0.54	0.6
6	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
7	0.04	0.08	0.12	0.16	0.2	0.24	0.28	0.32	0.36	0.4
8	0.03	0.06	0.09	0.12	0.15	0.18	0.21	0.24	0.27	0.3
9	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
10	0	0	0	0	0	0	0	0	0	0

Table 14. Egg serum albumin concentration (mg/mL) for training data solutions.

	A	B	C	D	E	F	G	H	I	J
1	0	0	0	0	0	0	0	0	0	0
2	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
3	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
4	0.15	0.3	0.45	0.6	0.75	0.9	1.05	1.2	1.35	1.5
5	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2
6	0.25	0.5	0.75	1	1.25	1.5	1.75	2	2.25	2.5
7	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3
8	0.35	0.7	1.05	1.4	1.75	2.1	2.45	2.8	3.15	3.5
9	0.45	0.9	1.35	1.8	2.25	2.7	3.15	3.6	4.05	4.5
10	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5

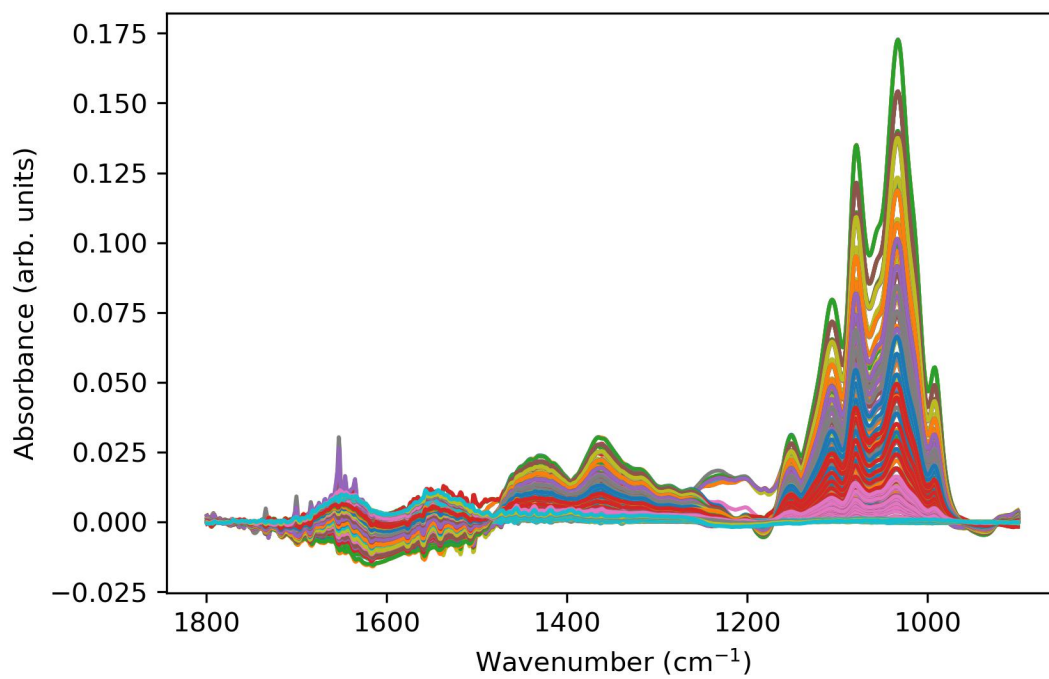


Figure 50. Composite spectra of all 100 samples used for training in each machine learning model.

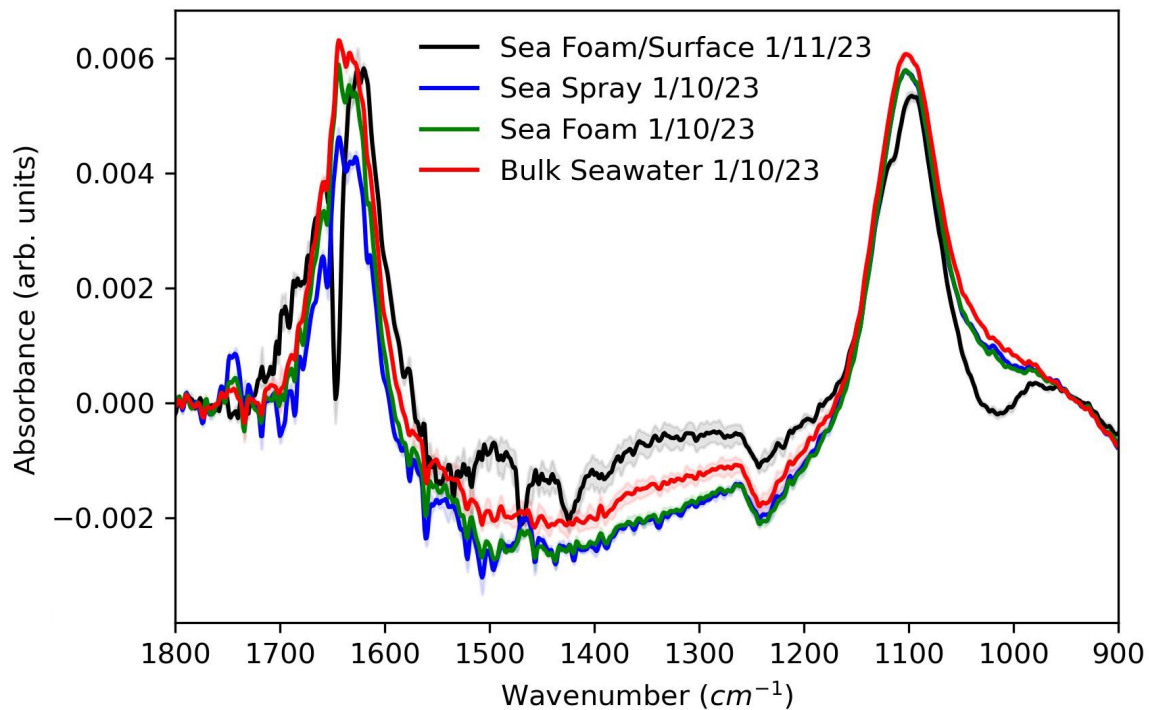


Figure 51. Average spectra of real ocean samples from Cocoa Beach, Florida. Standard deviation is shown but is approximately the thickness of the line.

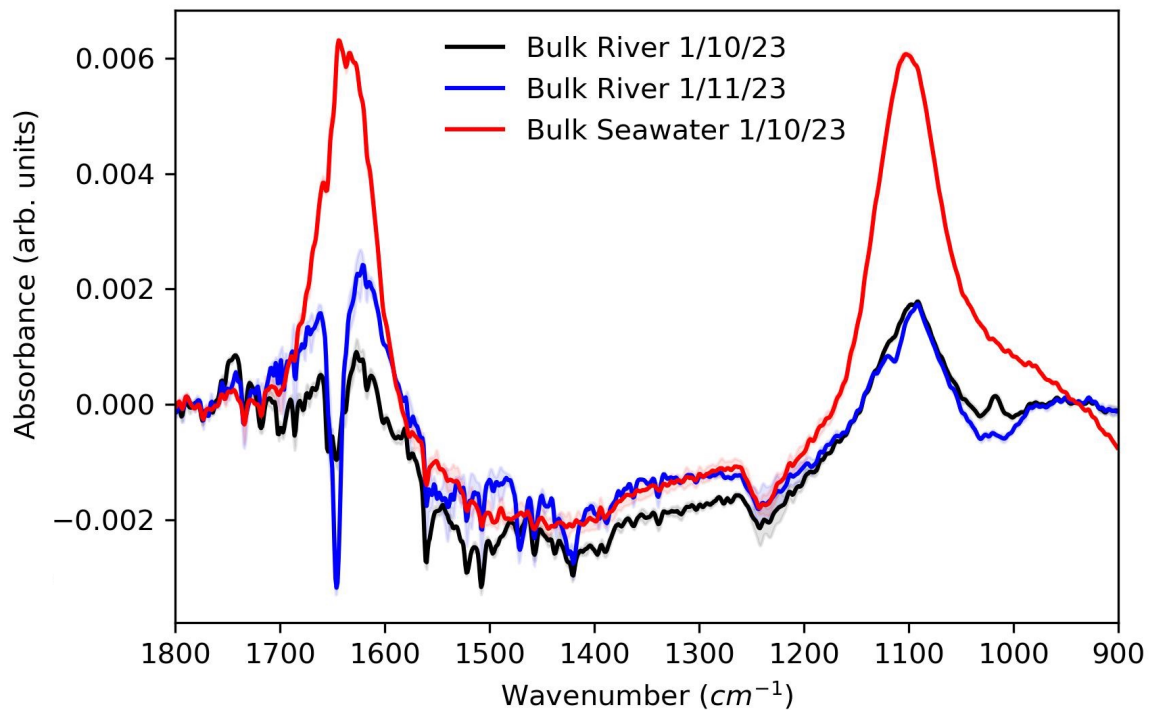
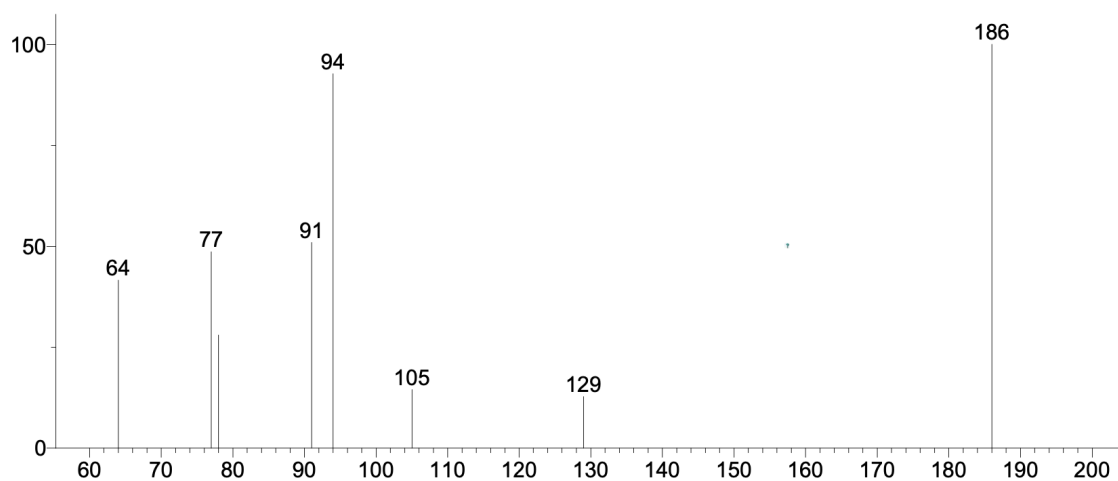


Figure 52. Average spectra of ocean and river samples from Cocoa Beach, Florida for comparison of sampling sites. Standard deviation is shown but is approximately the thickness of the line.

Unknown; InLib=-1558



(Text File) +EI Scan (rt: 1.035-1.279 min, 9 scans) C.D

Name: +EI Scan (rt: 1.035-1.279 min, 9 scans) C.D

MW: N/A ID#: 3502 DB: Text File

8 largest peaks:

186 999 | 94 926 | 91 509 | 77 486 | 64 416 | 78 279 | 105 145 | 129 128 |

8 m/z Values and Intensities:

64 416 | 77 486 | 78 279 | 91 509 | 94 926 | 105 145 | 129 128 | 186 999 |

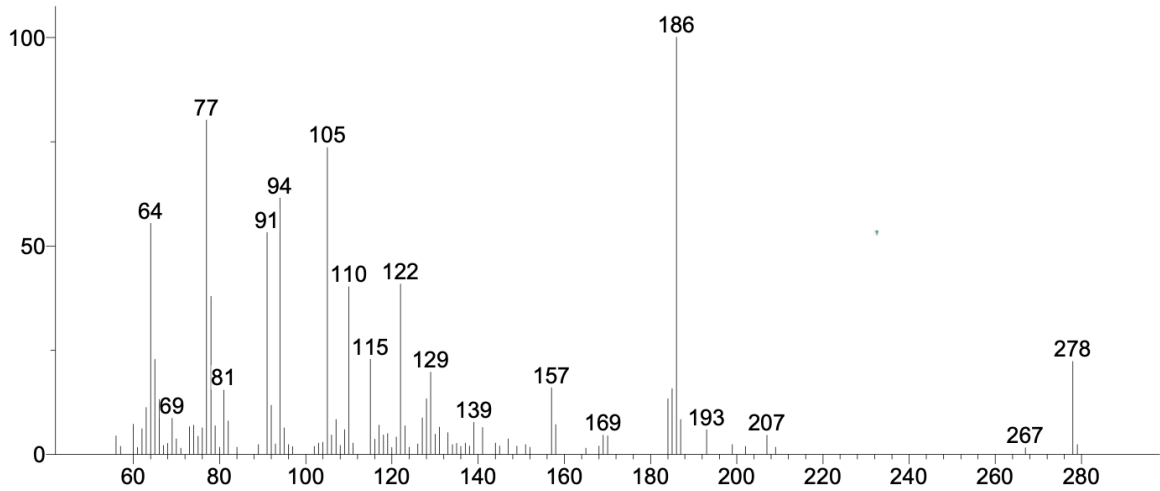
Synonyms:

no synonyms.

Figure 53. MS of GC retention for January 11, 2023, ocean surface sample from Cocoa Beach, Florida.



Unknown; InLib=-714



(Text File) +EI Scan (rt: 1.177-1.318 min, 24 scans) D.D

Name: +EI Scan (rt: 1.177-1.318 min, 24 scans) D.D

MW: N/A ID#: 3504 DB: Text File

10 largest peaks:

186 999 | 77 801 | 105 735 | 94 614 | 64 553 | 91 532 | 122 408 | 110 402 | 78 378 | 115 228 |

91 m/z Values and Intensities:

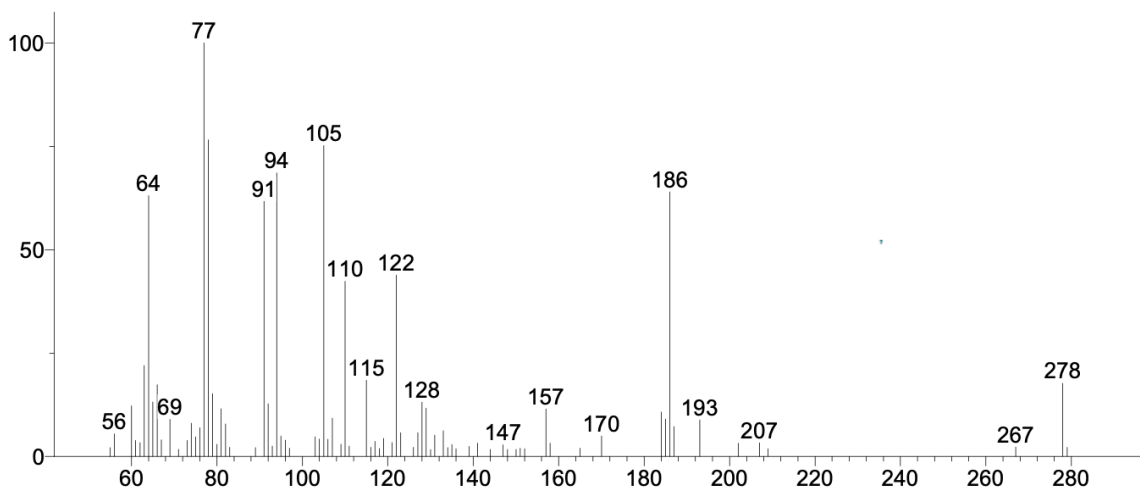
56 44 | 57 19 | 60 72 | 61 16 | 62 61 | 63 112 | 64 553 | 65 227 | 66 131 | 67 21 |  
68 26 | 69 87 | 70 37 | 71 15 | 73 66 | 74 70 | 75 43 | 76 63 | 77 801 | 78 378 |  
79 68 | 80 17 | 81 154 | 82 80 | 84 17 | 89 23 | 91 532 | 92 117 | 93 25 | 94 614 |  
95 63 | 96 23 | 97 18 | 102 19 | 103 27 | 104 29 | 105 735 | 106 46 | 107 83 | 108 21 |  
109 59 | 110 402 | 111 27 | 115 228 | 116 36 | 117 70 | 118 46 | 119 50 | 120 16 | 121 41 |  
122 408 | 123 68 | 124 17 | 126 25 | 127 87 | 128 133 | 129 197 | 130 48 | 131 65 | 133 52 |  
134 23 | 135 26 | 136 19 | 137 27 | 138 20 | 139 77 | 141 64 | 144 27 | 145 20 | 147 37 |  
149 20 | 151 23 | 152 17 | 157 160 | 158 71 | 165 15 | 168 20 | 169 46 | 170 44 | 184 133 |  
185 157 | 186 999 | 187 83 | 193 59 | 199 23 | 202 19 | 207 46 | 209 17 | 267 16 | 278 223 |  
279 23 |

Synonyms:

no synonyms.

Figure 54. MS of GC retention from January 11, 2023, river surface sample from the Banana River in Cocoa Beach, Florida.

Unknown; InLib=-1647



(Text File) +EI Scan (rt: 1.143-1.335 min, 24 scans) F.D

Name: +EI Scan (rt: 1.143-1.335 min, 24 scans) F.D

MW: N/A ID#: 3510 DB: Text File

10 largest peaks:

77 999 | 78 764 | 105 751 | 94 685 | 186 639 | 64 630 | 91 616 | 122 438 | 110 423 | 63 219 |

80 m/z Values and Intensities:

55 21 | 56 55 | 60 122 | 61 38 | 62 33 | 63 219 | 64 630 | 65 131 | 66 173 | 67 40 |  
69 90 | 71 17 | 73 38 | 74 80 | 75 47 | 76 69 | 77 999 | 78 764 | 79 151 | 80 29 |  
81 115 | 82 78 | 83 22 | 89 21 | 91 616 | 92 127 | 93 25 | 94 685 | 95 49 | 96 39 |  
97 20 | 103 47 | 104 42 | 105 751 | 106 41 | 107 92 | 109 29 | 110 423 | 111 25 | 115 184 |  
116 22 | 117 36 | 118 19 | 119 43 | 121 33 | 122 438 | 123 57 | 126 22 | 127 57 | 128 131 |  
129 116 | 130 16 | 131 51 | 133 62 | 134 21 | 135 28 | 136 18 | 139 24 | 141 32 | 144 17 |  
147 28 | 148 17 | 150 17 | 151 20 | 152 18 | 157 115 | 158 32 | 165 20 | 170 50 | 184 107 |  
185 90 | 186 639 | 187 72 | 193 88 | 202 32 | 207 33 | 209 18 | 267 23 | 278 177 | 279 22 |

Synonyms:

no synonyms.

Figure 55. MS of bulk surface water sample from Banana River in Cocoa Beach, Florida on January 10, 2023.

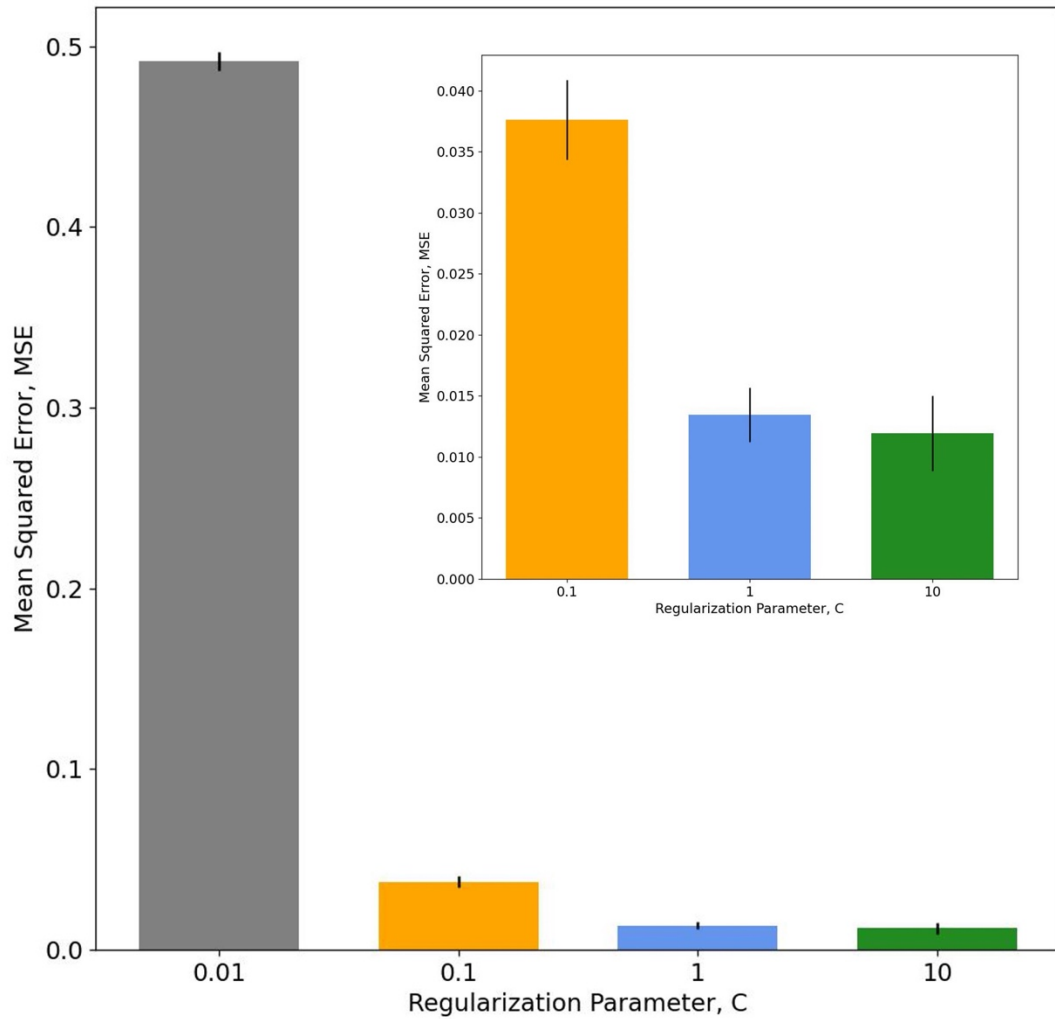


Figure 56. Optimization of regularization parameter C for the support vector regression (SVR). Variability shown is that of changing  $\epsilon$ , the tolerance limit, which varies little compared to the optimization of C.

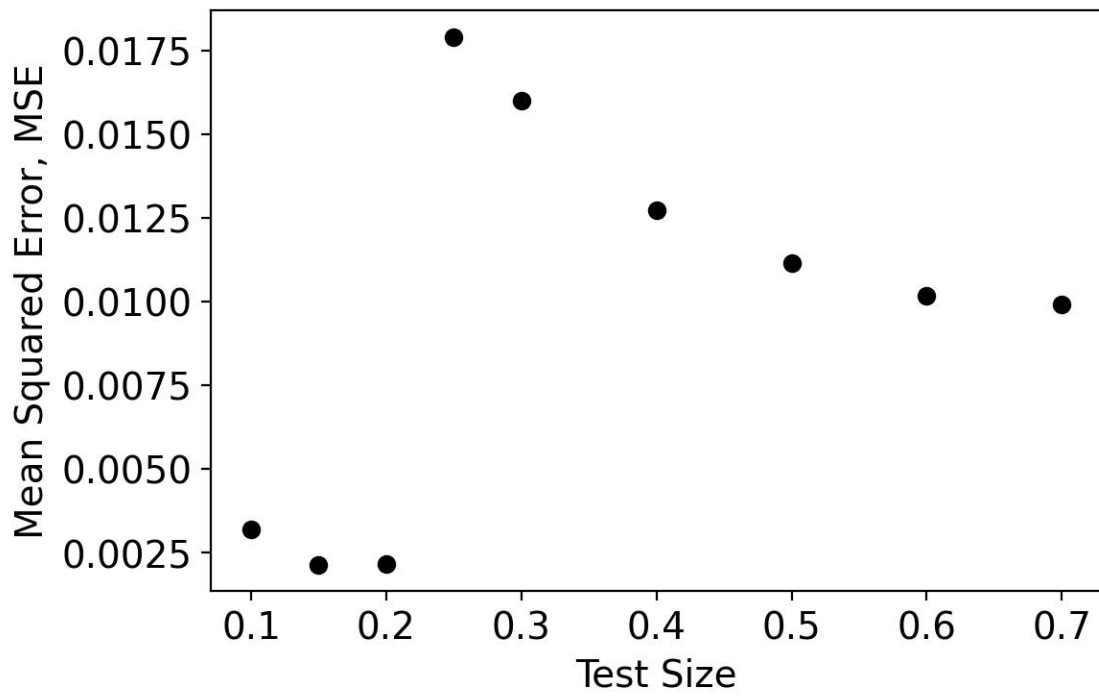


Figure 57. Optimization of train-test size split for the SVR. Minimization of MSE is prioritized for model performance. An 80/20 split minimizes MSE and has literature precedence.

## **Appendix G. Chemometric Investigation via Factor Analysis of Phosphate Raman Spectra to Elucidate Phosphate Monomer and Oligomer Spectral Components**

The work discussed in this appendix is my contribution to collaborative research with the following people, in alphabetical order. Affiliations are given for the associated institution that each person was at during the collaborative work.

Heather Allen<sup>1</sup>, Shelby Brantley<sup>2</sup>, Steven Corcelli<sup>2</sup>, James Dobscha<sup>3</sup>, Abigail Enders<sup>1</sup>, Amar Flood<sup>3</sup>, Douglas Vander Griend<sup>4</sup>, Jennifer Neal<sup>1</sup>, Liwei Yan<sup>1</sup>

1) Ohio State University, 2) University of Notre Dame, 3) Indiana University-Bloomington, 4) Calvin University

Polarized Raman spectra of aqueous  $\text{H}_2\text{PO}_4^-$  solutions from 0.01 to 4 M were analyzed via factor analysis to deconvolute the contributions of monomer, dimer, trimer, and tetramer phosphate structures. Using Python programming language and preinstalled packages, the spectra were decomposed into five components: four phosphate signatures and one noise component. The code is reproduced, in full, below. Figure 49 shows a subset of the phosphate Raman spectra. The four contributing factors are presented in Figure 50. To confirm the factor analysis was reasonable in its decomposition, reconstruction of a few sample concentrations was completed (Figure 51).

\*\*\*\*\*

Created on Thu Jul 9 10:04:30 2020

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html>

@author: AbbieEnders

\*\*\*\*\*

```
# import packages
import pandas as pd
from sklearn.decomposition import NMF
import os
# location of datafile
path = '[INSERT PATH HERE]' #path to directory containing datafile
filename = '[INSERT CSV NAME HERE].csv' #filename
newfilename = '[INSERT NEW FILE NAME HERE].csv'
os.chdir(path) # change directory
df = pd.read_csv(filename, header = 0) # read in the csv
# if there is no
wl = df['Wavelength'] # remove, convert wavenumber column to dataframe series
del df['Wavelength'] # remove wavenumbers from "training" data
# model we are using: NMF = non-negative matrix factorization
model = NMF(n_components=5, init = 'nndsvda', solver = 'cd', alpha = 1.) # init. method
model_fit = model.fit_transform(df) # fit model and transform using "training" data
model_fit_df = pd.DataFrame(data=model_fit[0:,0:]) # convert model to dataframe to write to csv
model_fit_df.insert(0, 'Wavelength', wl) # reinsert wl column
model_fit_df.to_csv(newfilename) # write to a new csv file
```

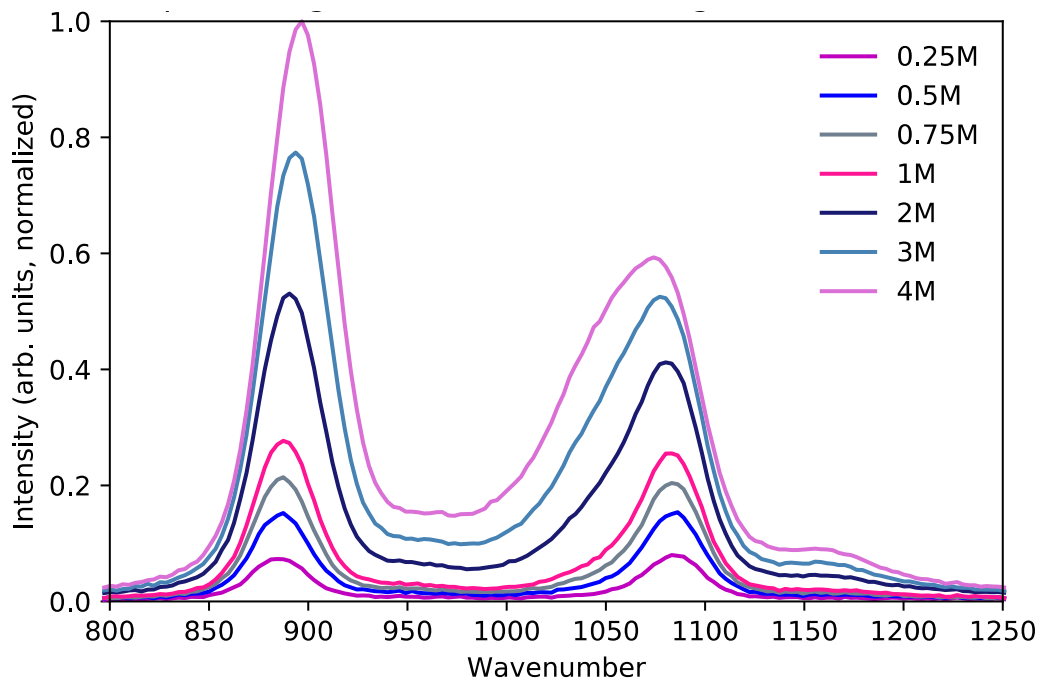


Figure 58. Select range of concentrations of phosphate ion Raman spectra.

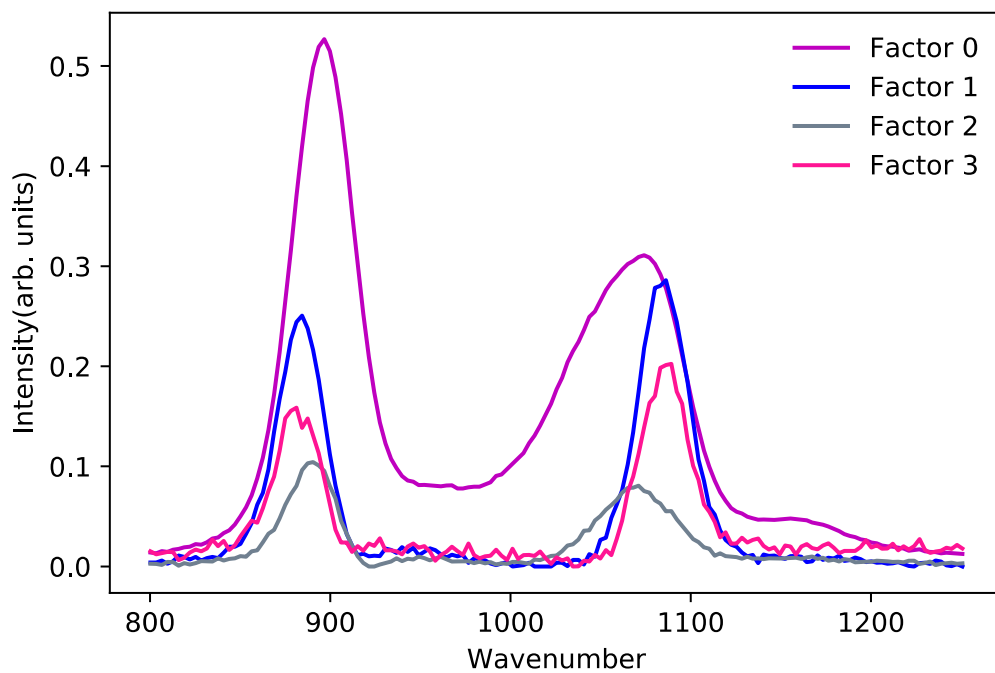


Figure 59. Resultant factor spectra after dimensionality reduction via factor analysis.

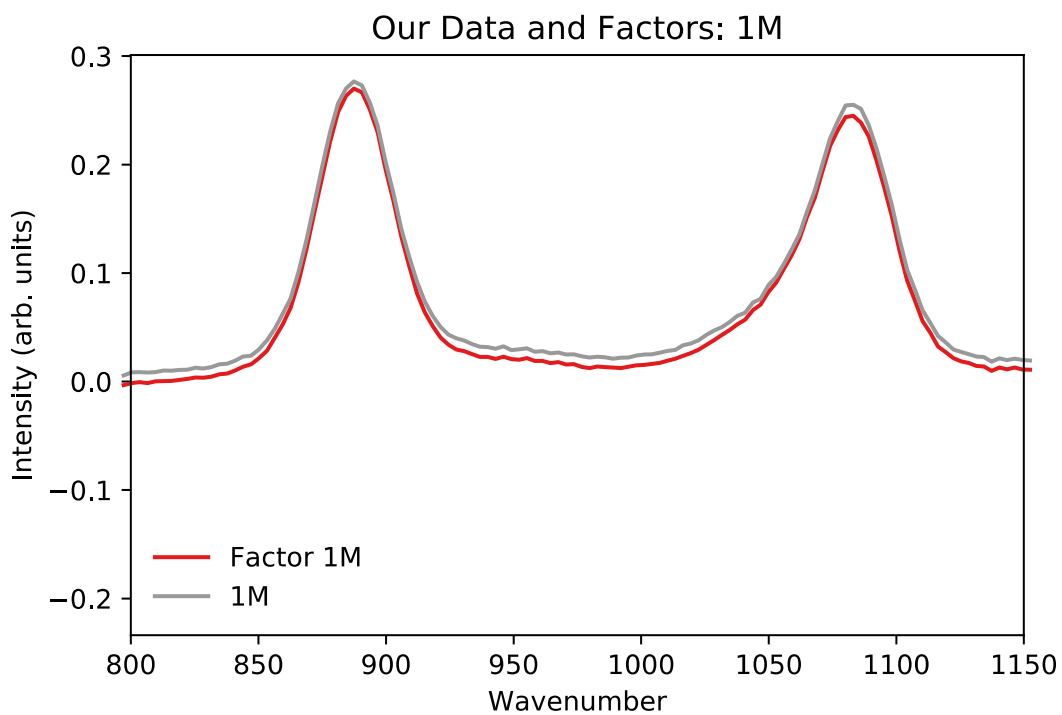


Figure 60. Reconstructed spectrum for 1M phosphate from factors and original 1M spectrum. The factors are reasonably similar.



## Appendix H. General Python Codes and GitHub Resources

Python codes are described with use and relevant output, where appropriate.

### H.1. Support Vector Regression

```
"""
Created on Wed Sep 14 17:53:00 2022
@author: AbbieEnders
"""
import matplotlib.pyplot as plt
import pandas as pd
import os
import numpy as np
from sklearn.svm import SVR
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error

# functions
# create list function

def createList(r1, r2, i):
    return np.arange(r1, r2+1, i)

def list_of_zeros(l):
    return [0] * l

path = r'[Insert Path to Directory Here]' # path to data files
unknown = '[Insert Unknown Sample File Name Here].csv' # unknown spectra for prediction
filename = '[Insert Training Data File Name Here].csv' # spectral data for training
conc = '[Insert File With Known Training Data Concentrations Here].csv' # glucose conc

figname = '[Insert Plot Name].svg'

os.chdir(path)
# read datafiles
df = pd.read_csv(filename, index_col=False)
```

```

concs = pd.read_csv(conc, index_col=False)
unse = pd.read_csv(unknown, index_col=False)
index_vals = createList(1,900,1)
unse = unse.set_index(index_vals)
conc_unknown = list_of_zeros(8)
conc_unknown = pd.DataFrame(conc_unknown)

#make list of concentrations (triplicate measurements), if not triplicate, do not repeat this
conc_list = list(concs.iloc[:, 0])
conc_l2 = []
c = 0
for i in conc_list:
    while c < 3:
        conc_l2.append(i)
        c += 1
    c = 0
conc = pd.DataFrame(conc_l2)

#set up spectra data
fdata = df.iloc[1:, 1:]
fdata = pd.concat([fdata,unse],1)
fdata = fdata.iloc[:900,:]

# scale data
sc_X = StandardScaler()
sc_y = StandardScaler()
plt.savefig('all_data.jpg',dpi=300)

# transpose data so samples are wavenumbers
X = np.transpose(fdata)
# use scaler on x,y data
X = sc_X.fit_transform(X.values.astype(float))
y = sc_y.fit_transform(conc.values.astype(float))

#remove 'unknown' data now
x_unknown = X[300,:]
X = X[:300,:]
y_unknown = y[300,:]
y = y[:300,:]

# train test data split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
# initialize regressor & fit
regressor = SVR(kernel='rbf')
regressor.fit(X_train,y_train)

```

```

score = regressor.score(X_test,y_test) # get R2 value for fit

y_pred = regressor.predict(X_test) # predict on new data
y_un_pred = regressor.predict(x_unknown) # predict on new data

mse = mean_squared_error(y_test, y_pred)# mean squared error
print(mse)
# transpose data back into og values for easier understanding (like it will give us a useful conc.)
X_test_trans = pd.DataFrame(np.transpose(X_test)) # transpose X data from test split
y_pred = y_pred.reshape(-1,1) # reshape y test data
y_pred_new = sc_y.inverse_transform(y_pred) # transform predicted y vals
y_un_pred = y_un_pred.reshape(-1,1) # reshape y unknown data
y_un_pred = sc_y.inverse_transform(y_un_pred) # reshape predicted y data from unknowns
x_test = sc_X.inverse_transform(X_test)
x_test = pd.DataFrame(np.transpose(x_test))
x_unknown_test = sc_X.inverse_transform(x_unknown)
x_unknown_test = pd.DataFrame(np.transpose(x_unknown_test))

# plotting to see how well fitting does, examine prediction
plt.scatter(fddata.iloc[765,:300],conc,c='#d8b365',marker='.',label='All Data')
plt.scatter(x_test.iloc[765:],y_pred_new,c='#80cdc1', marker = '.', label='Test Data')
plt.scatter(x_unknown_test.iloc[765:],y_un_pred,c='#018571',label='Predicted Concentration')
plt.legend(frameon=False)
plt.xlabel('Absorbance at 1036 $cm^{-1}$')
plt.ylim(-0.01,0.21)
plt.xlim(-0.001,0.02)
plt.ylabel('Concentration (M)')
plt.savefig(figurename, dpi=120)

```

Output: One figure with training data, line of best fit, and predicted values from unknown spectra plotted

## H.2. Work-Up of IRRAS Data

''''

```

@Author: AbbieEnders
# Work-up IRRAS Data

```

```

**Dependencies**

```

- \* data needs to be in a ascll format
- \* change the path to your path on your computer (see cell with "Bring in data" as header)
- \* need to be connected to your local run time

```

Imports

```

```

"""

# Commented out IPython magic to ensure Python compatibility.
import pandas as pd
import os
import numpy as np
from lmfit.models import LinearModel

def get_index(subset, shift):

    return min(range(len(subset)), key=lambda i: abs(subset[i]-shift))

#define the peak integration method

def fit_background(x, y, basestart1, peakstart1, peakend1, baseend1, plotnumber):

    #creates peak integration function to be called whenever needed from loop
    # return peakintegration
    # convert the peak shifts into indices using fn above and code below

    basestart = get_index(x, basestart1)
    peakstart = get_index(x, peakstart1)
    peakend = get_index(x, peakend1)
    baseend = get_index(x, baseend1)

    #seperate the peaks from the baseline to get the area where we think we have baseline
    xbase = np.zeros((basestart-peakstart)+(peakend-baseend))

    # can print if need to reference how many data points are in the area, not required
    xbase[0:(basestart-peakstart)] = x[peakstart:basestart]
    xbase[(basestart-peakstart):(basestart-peakstart) +
        peakend-baseend] = x[baseend:peakend]

    #get the y part of the data for the fit
    ybase = np.zeros((basestart-peakstart)+(peakend-baseend))
    #again, can print if a reference is needed
    ybase[0:(basestart-peakstart)] = y[peakstart:basestart]
    ybase[(basestart-peakstart):(basestart-peakstart) +
        peakend-baseend] = y[baseend:peakend]

    #create the model of the background (currently assuming it is linear, but can adjust based on
    needs, should proceed well enough with linear)
    prodbackground = LinearModel(prefix="prodback_")

```

```

pars = prodbackground.guess(ybase, xbase)
#put the model together from the different components coded thus far
model = prodbackground
#fit the model with the input file data
out = model.fit(ybase, pars, x=xbase)
xsubset = x[baseend:basestart]
# generates new x,y data so the background correlation doesn't affect plot
ysubset = y[baseend:basestart] - (out.params["prodback_slope"].value *
                                x[baseend:basestart]+out.params["prodback_intercept"].value)

return prodbackground, xsubset, ysubset

```

""IRRAS data workup walk-through:

1. calculate absorbance
2. average spectra
3. calculate standard deviation
4. plot average with standard deviation shading

Data is brought in

""

""filenames""

```

# REQUIRED USER INPUT, input your data file names in asc
water_1 = '.asc'
water_2 = '.asc'
water_3 = '.asc'
data_1 = '.asc'
data_2 = '.asc'
data_3 = '.asc'
newfile = '.csv'
path = r'[Insert Path Name Here]'
os.chdir(path)
# bring in data
water_1 = pd.DataFrame(np.genfromtxt(path+'/'+water_1, skip_header=25))
water_2 = pd.DataFrame(np.genfromtxt(path+'/'+water_2, skip_header=25))
water_3 = pd.DataFrame(np.genfromtxt(path+'/'+water_3, skip_header=25))
data_1 = pd.DataFrame(np.genfromtxt(path+'/'+data_1, skip_header=25))
data_2 = pd.DataFrame(np.genfromtxt(path+'/'+data_2, skip_header=25))
data_3 = pd.DataFrame(np.genfromtxt(path+'/'+data_3, skip_header=25))

```

""Data workup (as described above)""

```

# calculate absorbance
spectrum_1 = -np.log10(data_1.iloc[:, 1]/water_1.iloc[:, 1])
spectrum_2 = -np.log10(data_2.iloc[:, 1]/water_2.iloc[:, 1])
spectrum_3 = -np.log10(data_3.iloc[:, 1]/water_3.iloc[:, 1])

# background fit peaks of interest (H2O peak, etc)
# base start, peak start, peak end, base end
bsa_peak1 = [900, 980, 1750, 1800]

bp1s1_fit, bp1x1, bp1y1 = fit_background(
    water_1.iloc[:, 0], spectrum_1, bsa_peak1[0], bsa_peak1[1], bsa_peak1[2], bsa_peak1[3], 1) #
BSA peak 1, spectrum 1
bp1s2_fit, bp1x2, bp1y2 = fit_background(
    water_1.iloc[:, 0], spectrum_2, bsa_peak1[0], bsa_peak1[1], bsa_peak1[2], bsa_peak1[3], 1) #
BSA peak 1, spectrum 1
bp1s3_fit, bp1x3, bp1y3 = fit_background(
    water_1.iloc[:, 0], spectrum_3, bsa_peak1[0], bsa_peak1[1], bsa_peak1[2], bsa_peak1[3], 1) #
BSA peak 1, spectrum 1

frames = [pd.Series(spectrum_1), pd.Series(spectrum_2), pd.Series(spectrum_3),
    pd.Series(bp1y1), pd.Series(bp1y2), pd.Series(bp1y3)]
data = pd.concat(frames, axis=1)
data['wn'] = water_1.iloc[:, 0]
data.to_csv(path+'/' + newfile)

```

Output: data fit within specified region saved to one file with original data

### H.3. Calculating SSnL Carbon Using Chlorophyll and Zooplankton data from E3SM Model

```

"""
Created on Tue Nov 23 13:27:19 2021

@author: AbbieEnders
"""
# adjusting the
import pandas as pd
import os
import math
import numpy
import re
import glob
# rotating matrix
def rotate_180(array, M, N, out):

```

```

for i in range(M):
  for j in range(N):
    out[i, N-1-j] = array[M-1-i, j]

# equation
# Ci = GCz(Cp/(Ck,inges + Cp))(1-a)(pi%)(ti)(o)
# variables
# Ci = carbon of ith species
# G = 1/day zooplanktonic growth rate
# Cz = uM carbon zooplankton concentration
# Cp = uM carbon planktonic carbon atom concentration
# Ck,inges = uM carbon half saturation for ingestion
# a = assimilation efficiency
# pi% = percentage of the ith macromolecule content in a typical planktonic cell
# ti = day lifetime of the ith macromolecule, total restricted to 2
# o = coating of the surface based on partial adsorption
# ChIA = remotely sensed by NASA MODIS
# Cmr = chlorophyll mass ratio (multiply ChIA by 50 to get Cp)

# variables defined
a = 0.75
G = 1
#Cz = 0.5
Ckinges = 7
tprot = 10
tlip = 2
pprot = 0.6
plip = 0.2
Cmr = 50
C_ratio = 0.5
CpRef = 10 # carbon protein reference uM
CIRef = 0.5 # carbon lipid reference uM
np = 0.5
nl = 1
ap = 1
al = 1

mol_mass_carbon = 12.01 # g/mol
ocean_surf_area = 3.60580510*10**14 #m^2

earth_surf_area = 5.10082000*10**14 # m^2
num_of_instances = 180 *360 #1 steps in lat and lon
a_pixel = earth_surf_area/num_of_instances
surfprot = 0.002*a_pixel # grams
surfliplip = 0.0025*a_pixel # grams

```

```

# relate num of instances over ocean surface area

# new dataframe
# math
# if ChlA = 99999 do nothing
path = '[Insert Path to Data]'
os.chdir(path)
all_files_chl = glob.glob(path + "/*chl.csv")
#print(all_files_chl)
all_files_zoo = glob.glob(path + "/*zoo.csv")
#lat_list = list(range(-90, 91,0.5))
lat_list = numpy.arange(-90,91,0.5).tolist()
#lat_list.append(list(range(90,-1)))
#print(lat_list)
os.chdir(path)
for filename in all_files_chl:

    data = pd.read_csv(filename, index_col=0)
    Cp_temp = Cmr*data
    Cz = re.sub('chl','zoo',filename)
    zoo = pd.read_csv(Cz, index_col = 0)
    zoo = zoo/1000
    C_prot_temp = G*zoo*(Cp_temp/(Ckinges + Cp_temp))*(1-a)*(pprot)*(tprot)
    C_lip_temp = G*zoo*(Cp_temp/(Ckinges + Cp_temp))*(1-a)*(plip)*(tlip)
    C_sum_temp = C_prot_temp + C_lip_temp

    theta_prot = (((1/CpRef)**np)*(ap*C_prot_temp**np))/(1
+((((1/CpRef)**np)*(ap*C_prot_temp**np))+(((1/ClRef)**nl)*(al*C_lip_temp**nl))))
    theta_lip = (((1/ClRef)**nl)*(al*C_lip_temp**nl))/(1
+((((1/CpRef)**np)*(ap*C_prot_temp**np))+(((1/ClRef)**nl)*(al*C_lip_temp**nl))))
    sums = 0
    counter = 0
    my_dict = {}
    for i in lat_list:
        i=float(i)
        C_total = theta_lip.iloc[counter,:]*surflip*math.cos(numpy.deg2rad(abs(i))) +
theta_prot.iloc[counter,:]*surfprot*math.cos(numpy.deg2rad(abs(i)))
        my_dict[i]=(abs(C_total))
        sums = sums + C_total.to_numpy().sum()
        counter += 1
    savefile = re.sub('chl','carbon_w_zoo',filename)
    C_total = pd.DataFrame.from_dict(my_dict, orient = 'index')

```

Output: carbon on the ocean surface in csv file



## H.4. Principal Component Analysis

''''

Created on Mon Feb 28 15:05:24 2022

@author: AbbieEnders

''''

```
import numpy as np
import pandas as pd
import os
import matplotlib.pyplot as plt
from sklearn.decomposition import FactorAnalysis, PCA
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score, explained_variance_score
from sklearn.model_selection import train_test_split
from factor_analyzer import FactorAnalyzer
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
from factor_analyzer.factor_analyzer import calculate_kmo

path = r'[Insert Path to Directory with Data]'

os.chdir(path)
fa_data_file = '[Insert File Name].csv'
glucosedata = '[Insert Concentration File Name].csv'
datasavefile = '[Insert Name For Save File].csv'
plotsavefile = '[Insert Name For Plot File].svg'
# get data
factor_data = pd.read_csv(fa_data_file, index_col=False) # rename a copy of dataframe to work with
glucose_columns = pd.read_csv(glucosedata, index_col=None)
glucose_columns = list(glucose_columns['list']) # make sure you have a column named list
#PCA
pca_model = PCA(n_components= 4)
pca_fitx = pca_model.fit(factor_data).transform(factor_data)
pca = pca_model.fit(factor_data)
pca_fit = pca.transform(factor_data)
pca_fit_df = pd.DataFrame(data = pca_fit[0:,0:])
pca_loadings = pd.DataFrame(pca.components_, columns = glucose_columns)
pca_loadings.to_csv(datasavefile)
# plot
df = pd.read_csv(datasavefile, index_col = None)
plt.scatter(pca_loadings.iloc[0,0:],pca_loadings.iloc[1,0:], color = 'b')#,label = 'Real')
plt.ylabel('PC2')
```

```
plt.xlabel('PC1')
plt.legend(frameon=False)
plt.savefig(plotsavefile,dpi = 120)
```

Output: Loadings, or components, of the PCA saved to a file and a figure with the first two principal components compared to each other to determine their relationship

## H.5. Support Vector Machine

```
"""
```

Created on Mon Feb 28 15:05:24 2022

@author: AbbieEnders

```
"""
```

```
import pandas as pd
import os
import matplotlib.pyplot as plt
from sklearn import svm
import numpy as np
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.inspection import DecisionBoundaryDisplay

path = r'[Insert Path to Directory for Data]'
#newfilename = 'absorbance.csv'
os.chdir(path)
fa_data_file = '[Insert Data File Name].csv'
factor_data = pd.read_csv(fa_data_file,index_col=0) # rename a copy of dataframe to work with
factor_dataT = factor_data.transpose()
glucose_columns = pd.read_csv('[Insert Concentration Data File Name].csv',index_col=None)
y = list(glucose_columns['list'])

X = factor_dataT.iloc[:, :2]

# fit the model, don't regularize for illustration purposes
clf = svm.SVC(kernel="linear", C=100)
clf.fit(X, y)

plt.scatter(X.iloc[:, 0], X.iloc[:, 1], c=y, s=30, cmap=plt.cm.Paired)
plt.axis([0.2, .23, -.5, .5])
```

```

# plot the decision function
ax = plt.gca()
DecisionBoundaryDisplay.from_estimator(
    clf,
    X,
    plot_method="contour",
    colors="k",
    levels=[-1, 0, 1],
    alpha=0.5,
    linestyle=["--", "-", "--"],
    ax=ax,
)
# plot support vectors
ax.scatter(
    clf.support_vectors_[:, 0],
    clf.support_vectors_[:, 1],
    s=100,
    linewidth=1,
    facecolors="none",
    edgecolors="k",
)
plt.show()

```

Output: Plot of SVM results

## H.6. FTIR Spectrum Calculations Based on Angle of Incident Light

\*\*\*\*\*

Created on Tue Apr 20 16:02:46 2021

@author: AbbieEnders

\*\*\*\*\*

```

# imports
import math
import numpy as np
import csv
import os
import matplotlib.pyplot as plt
import pandas as pd

```

```

# variables

```

```

path = r'[Insert Path to Directory Here]'
os.chdir(path)
filename = '[Insert File with Imaginary Refractive Index].csv'
file_H2O = '[Insert File Real Imaginary H2O Index Data].csv'
df = pd.read_csv(filename, header = 0)
df_H2O = pd.read_csv(file_H2O, header = 0)
data_dict = {}
nfn = '[New File Name].csv' #filename

def createList(r1, r2, i):
    return np.arange(r1, r2+1, i)

angle0 = createList(45, 55, 1) # list for angle of incidence
n2 = df_H2O['n']
k2 = df_H2O['k']
k = df['k']
wl = df_H2O['wavelength']
wavenumber = df['Wavenumber']
i = 0
d = 2.5 #nm #depth of monolayer/surface
y = []

while i < len(wavenumber):
    y.append((1/wavenumber[i])*10**7)
    i += 1

for angle in angle0:
    data_list = []
    counter = 0 #start with first thing (0)
    for wavelength in y:
        kappa = k2[counter]/n2[counter]
        k_monolayer = k[counter]
        n = complex(1.42545, k_monolayer)
        l = (n**2 - n2[counter]**2)*d
        # Equation 14 Reflection Absorbance for s polarized light
        RAs_top = (16*math.pi*kappa*math.cos(angle)*l)
        RAs_bottom = (wavelength*((n2[counter]**2)-1)**2)
        RAs = RAs_top/RAs_bottom
        data_list.append(RAs)
        counter += 1 # counter = counter + 1
    data_dict[angle] = data_list

X = [x.imag for x in data_dict[angle]]
plt.plot(wavenumber, X)
plt.text(0,1,angle)

```

```

plt.title('Calculated Reflectance Absorbance at Given Angle')
plt.xlabel('Wavenumber ( $\text{cm}^{-1}$ )')
plt.ylabel('Reflectance-Absorbance of s-Polarized Light')
plt.savefig(str(angle)+'fig.svg')
plt.show()

```

```

with open(nfn, 'w') as csvfile:
    writer = csv.writer(csvfile)
    writer.writerow(data_dict.keys())
    writer.writerows(zip(*data_dict.values(), y))

```

Output: Figures with calculated spectrum and a datafile to use these spectra

## H.7. Linear Regression Model

```

"""
Created on Wed Sep 14 17:53:00 2022
@author: AbbieEnders
"""
# imports
import matplotlib.pyplot as plt
import pandas as pd
import os
import numpy as np
from sklearn import linear_model
from sklearn.metrics import mean_squared_error, r2_score

# functions
# create list function

def createList(r1, r2, i):
    return np.arange(r1, r2+1, i)

path = r'[Insert Path to Directory With Data]'
unknown = '[File for Unknown Data].csv' # unknown data file
filename = '[File for Training Data].csv'
conc = '[File with concentration values].csv' # glucose conc
savefigfile = '[File for figure after model fit].svg'
os.chdir(path)
df = pd.read_csv(filename, index_col=False)

```

```

concs = pd.read_csv(conc, index_col=False)
unse = pd.read_csv(unknown, index_col=False)
unse = unse.iloc[:, :9]
conc_unknown = [0.200203929,0.200203929,0.200203929,
                0.150152947,0.150152947,0.150152947,
                0.100101964,0.100101964,0.100101964]
conc_unknown = pd.DataFrame(conc_unknown)

#make list of concentrations (triplicate measurements)
conc_list = list(concs.iloc[:, 0])
conc_l2 = []
c = 0
for i in conc_list:
    while c < 3:
        conc_l2.append(i)
        c += 1
    c = 0
conc = pd.DataFrame(conc_l2)

#set up spectra data
fdata = df.iloc[1:, 1:]
# transpose df
fdata = fdata.transpose()
# train test splits of x y data
X_train = fdata[:-20]
X_test = fdata[-20:]
y_train = conc_l2[:-20]
y_test = conc_l2[-20:]

#create linear regressor
lreg = linear_model.LinearRegression()
# Fit to training data
lreg.fit(X_train, y_train)
# predict on test data
y_pred = lreg.predict(X_test)

# predict on "new" data
unknown = unse.transpose()
y_pred_un = lreg.predict(unknown)

# get coef of model
#print('coef: \n', lreg.coef_)
# print mean sq. err
print('mean sq. err: %.2f' % mean_squared_error(y_test, y_pred))

```

```

print('coef of determ: %.2f %r2_score(y_test,y_pred))

plt.scatter(X_test.iloc[:,765], y_test, color="#d8b365", marker = '.',label='Experimental Conc.')
plt.scatter(X_test.iloc[:,765], y_pred, color="#80cdc1", marker = '.', label = 'Predicted Conc.')
plt.scatter(unknown.iloc[:,765],conc_unknown, color = '#018571', marker='o',label = 'Unknown
True')
plt.scatter(unknown.iloc[:,765],y_pred_un, color='#018571',marker='x', label = 'Unknown
Predicted')

plt.xlabel ('Absorbance (arb. units, 1036 cm-1)')
plt.ylabel ('Concentration (M)')
plt.legend(frameon=False)

diff = conc_unknown - pd.DataFrame(y_pred_un)
diffper = 100*(diff/conc_unknown)
plt.savefig(savefigfile,dpi=120)

```

Output: figure of model fit with concentration versus absorbance

## H.8. Subprocess Script for Preprocessing NIST FTIR Spectra

```

"""
Created on Sun Nov 8 09:46:55 2020
@author: AbbieEnders
"""
#IMPORTS
import subprocess
import os

# here we will call on each of our processes
# Step X: Run X
# subprocess.run(['python', 'filename.py'], shell=True)
#VARIABLES
path = 'jcamp_files'
bad_path_1 = 'cond_not_met'
bad_path_2 = 'in_trans'
path_to_csv = 'unnormalized_csv'
path_to_dest = 'csv'
path_to_images = 'images'
ext_1 = '/*.csv'
ext_2 = '/*_n.csv'

```

```

ext_3 = '/*.jpg'
top_dir = ['nitrile','ketone','ether','ester','carboxylic_acid','aromatic',
          'amine','amide','alkyne','alkane','alkene','alcohol',
          'nitro','alkyl_halide','acyl_halide','methyl','aldehyde']
func_groups = ['nitrile','ketone','ether','ester','carboxylic_acid','aromatic',
              'amine','amide','alkyne','alkane','alkene','alcohol',
              'nitro','alkyl_halide','acyl_halide','methyl','aldehyde',
              'not_nitrile','not_ketone','not_ether','not_ester','not_carboxylic_acid','not_aromatic',
              'not_amine','not_amide','not_alkyne','not_alkane','not_alkene','not_alcohol',
              'not_nitro','not_alkyl_halide','not_acyl_halide','not_methyl','not_aldehyde']

# Step #1: Create directories if they don't exist
for i in top_dir:
    if not os.path.exists(top_dir):
        os.mkdir(os.path.join(top_dir, top_dir))
        not_dir = 'not_'+top_dir
        os.mkdir(os.path.join(top_dir, not_dir))

# Step #2: Move any files that are not in absorbance\wavenumbers
subprocess.run(['python', 'check_file_in_absorbance.py', path, bad_path_1, bad_path_2])

# Step #3: convert from jcampdx to csv
subprocess.run(['python', 'jcamp_to_csv.py', path])

# Step #4: move csv to their own folder
subprocess.run(['python', 'move_file.py', path, path_to_csv, ext_1])

# Step #5: normalize each spectrum
subprocess.run(['python', 'normalize_csv.py', path_to_csv, path_to_dest])

# Step #6: move normalized spectrum
subprocess.run(['python', 'move_file.py', path, path_to_csv, ext_2])

# Step #7: turn each csv file into a jpg image
subprocess.run(['python', 'convert_to_jpg.py', path_to_dest])

# Step #8: move jpg images
subprocess.run(['python', 'move_file.py', path_to_dest, path_to_images, ext_3])

```



```

# Step #9: copy files to the folder for functional groups
for d in top_dir:
    dst_for_images = d
    #for fg in func_groups:
    listname = d+'.csv'
    dst_for_images = os.path.join(d, d)
    subprocess.call(['python','copy_file_to_ndir.py', path_to_images, listname, dst_for_images],
shell=True)
    # then do the same but for "not_X"
    listname = 'not_'+d+'.csv'
    dst_for_images = os.path.join(d, 'not_'+d)
    subprocess.call(['python','copy_file_to_ndir.py', path_to_images, listname, dst_for_images],
shell=True)

# Step #10: "even" out folders and remove files for validation
for d in top_dir:
    dst_for_images1 = os.path.join(os.getcwd(), d, d)
    n = 'not_'+d
    dst_for_images2 = os.path.join(os.getcwd(), d, n)
    val = 'test_images'
    dst_for_val = os.path.join(os.getcwd(), d, val)
    # dst_for_val = d+'\test_images'
    subprocess.call(['python','random_number_files.py', dst_for_images1, dst_for_images2,
dst_for_val], shell=True)

# Step 11: create functional group directory in each functional group's directory to move photo
directories to
for direc in top_dir:
    d = direc
    n = 'not_'+d
    subprocess.call(['python', 'make_functional_group_directory.py', 'functional_group', d, n])

```

Output: preprocessed spectra in separated directories for training and testing

## H.9. Check that Spectrum is in Units of Absorbance

\*\*\*\*

Created on Sun Nov 8 17:19:16 2020

@author: AbbieEnders

\*\*\*\*

```

import jcamp
import os
import glob
import shutil
import sys
# This is where your data is coming from and going to
# The following lines will find all of the files of a given type in the path's folder

path = os.path.join(os.getcwd(), sys.argv[1])
destination = os.path.join(os.getcwd(), sys.argv[2])
final_dest = os.path.join(os.getcwd(), sys.argv[3])
# you are probably not moving the py file around, so just change directory to look/touch in
correct folder
# Here are the files that fit your criterion that are within the path file

extension = 'jdx'
all_files = glob.glob(path + "*" + extension)
# read jcampdx file and check absorbance and wavenumbers units
# return false if not in micrometers (wavenumbers does not equal micrometers = True)
# return false if not in absorbance (y-units in absorbance = True)

for file in all_files:
    data = jcamp.JCAMP_reader(file)
    wavenumbers = data.get('x_units', r'N/A').lower() != 'micrometers'
    absorbance = data.get('yunits', r'N/A').lower() == 'absorbance'
    # move file if wavenumbers is in micrometers

    if wavenumbers == False:
        print('bad apple')
        shutil.move(file, destination)

    # move file if not in absorbance
    if absorbance == False:
        shutil.move(file, final_dest)

```

Output: moves any spectrum files that are not in absorbance mode

## H.10. Copy a Given File to a New Directory

\*\*\*\*

```
Created on Wed Aug 26 09:09:56 2020
@author: AbbieEnders
****

# move files based on the functional group present
from shutil import copyfile
import csv
import re
import glob
import os
import sys
#path to spectra csv files

path = os.path.join(os.getcwd(), sys.argv[1])
path_to_list = os.path.join(os.getcwd(), sys.argv[2])
dst = os.path.join(os.getcwd(), sys.argv[3])

# read list from file with names of files containing or not containing a functional group
with open(path_to_list, newline="") as f:
    reader = csv.reader(f)
    filenames = list(reader)
os.chdir(path)
extension = '.jpg' # extension of the file you are searching for

#results = glob.glob(path + extension)
#print(results)
os.chdir(path)
results = glob.glob('*.{0}'.format(extension))
print(results)
#Here are the files that fit your criterion that are within the path file

for file in filenames:
    file = str(file)
    file = re.sub('[\]', '', file)
    file = re.sub('\\"', '', file)
    file = re.sub('\|', '', file)
    file = file+'.jpg'
```

```

if file in results:
    print('file in list')
    src = os.path.join(path, file)
    dest = os.path.join(dst, file)
    print(dest)
    copyfile(src, dest)

```

Output: specified files are moved to a new directory

## H.11. Convert File From 'csv' to 'jpeg' Format

\*\*\*\*

Created on Tue Aug 25 09:24:53 2020

@author: AbbieEnders

\*\*\*\*

#create a jpg of each spectrum

```

import pandas as pd
import matplotlib.pyplot as plt
import os
import glob
import re
import sys
#path to spectra csv files
path = os.path.join(os.getcwd(), sys.argv[1])

extension = 'csv' # extension of the file you are searching for
os.chdir(path) # change the working directory so you can access this from anywhere on your
computer
result = glob.glob('*.{}'.format(extension))
#Here are the files that fit your criterion that are within the path file

for filename in result:
    df = pd.read_csv(filename, index_col=False, header=0)
    filename = re.sub('_n.csv', "", filename)
    filename=filename+".jpg"
    fig = df.plot(df.columns[0], df.columns[1], color='black', legend=None)
    fig.set_xlabel(None)
    fig.set_xlim(4000, 600)

```

```
plt.savefig(filename)
plt.close()
```

Output: specified csv files converted to jpeg files

## H.12. Convert Specified File From 'jcampdx' to 'csv' File

''''

```

Created on Mon Jun 15 17:53:29 2020
@author: AbbieEnders
Using jcamp.py from GITHUB
#####
#####
#####NOTES#####
#####
#####
#####
RUN IN COMMAND LINE: 1 ) python -m pip install git+https://github.com/nzhagen/jcamp
numpy version: pip install numpy==1.19.0
(install git, pip if you don't have it)
2) python jcamp_to_csv.py
#####
#####
''''

# read in jcampdx file to dict and write list to csv file
import jcamp
import os
import glob
import csv
import re
import sys

#This is where your data is coming from and going to

path = os.path.join(os.getcwd(), sys.argv[1])
#The following lines will find all of the files of a given type in the path's folder
extension = 'jdx'
os.chdir(path) #you are probably not moving the py file around, so just change directory to
look\touch in correct folder
#Here are the files that fit your criterion that are within the path file

```

```

result = glob.glob('*.{}.format(extension))
all_files = glob.glob(path + "*"'.jdx")
for file in all_files:
    data = jcamp.JCAMP_reader(file)
    nfn = re.sub('.jdx','.csv',file) #nfn = new filename
    with open(nfn, 'w', newline = ") as f:
        writer = csv.writer(f, delimiter = ',')
        writer.writerow(('x','y'))
        writer.writerows(zip(data['x'], data['y']))

    if not f.closed:
        f.close()

```

Output: specified jcampdx files are converted to csv files

### H.13. Make Directory Given Keywords for Naming

```

"""
Created on Mon Nov 16 19:41:52 2020
@author: AbbieEnders
"""

import os
import sys
import shutil
# directory to create inside the original directory
target_dir = os.path.join(os.getcwd(), sys.argv[2], sys.argv[1])
move_dir1 = os.path.join(os.getcwd(), sys.argv[2], sys.argv[3])
move_dir2 = os.path.join(os.getcwd(), sys.argv[2], sys.argv[2])

os.mkdir(target_dir)
shutil.move(move_dir1, target_dir)
shutil.move(move_dir2, target_dir)

```

Output: creates directories given input details

### H.14. Move File to Different Directory

```

"""
Created on Fri Nov 13 12:54:59 2020
@author: AbbieEnders
"""

```

```

# move csv files to new folder

import shutil
import os
import sys
import glob
extension = '.csv' # file format
og = os.path.join(os.getcwd(), sys.argv[1]) # original
new = os.path.join(os.getcwd(), sys.argv[2]) # destination
os.chdir(og) # change directory
# get all of the files in the folder that match the extension
results = glob.glob('*.{0}'.format(extension))
for file in results:
    og = os.path.join(og, file)
    new = os.path.join(new, file)
    shutil.move(og, new) # move

```

Output: moves files based on specifications

## H.15. Normalize Data

''''

```

Created on Mon Jun 15 19:27:00 2020
@author: AbbieEnders
Normalize csv data and save as a new csv file, preserving og data
''''

import pandas as pd
import numpy as np
import os
import glob
import csv
import re
import sys
import shutil

#This is where your data is coming from and going to
path = os.path.join(os.getcwd(), sys.argv[1])
dst = os.path.join(os.getcwd(), sys.argv[2])
#The following lines will find all of the files of a given type in the path's folder
extension = 'csv'
os.chdir(path)

```

```

result = glob.glob('*.{}.format(extension))
#Here are the files that fit your criterion that are within the path file
li = []
for filename in result:
    #read in file
    df = pd.read_csv(filename, index_col=False, header=0)
    #rename file
    filename = re.sub('.csv','_n.csv',filename)
    #normalize with respect to max y value in file
    df['y']= (df['y']/np.amax(df['y']))
    li.append(df['y'])
    #write normalized files to a new csv
    with open(filename, 'w', newline='') as file:
        writer = csv.writer(file, delimiter = ',')
        writer.writerow(('cm-1', 'l'))
        writer.writerows(zip(df['x'],df['y']) )
    if not file.closed:
        file.close()
    src = os.path.join(os.getcwd(), filename)
    dst = os.path.join(os.getcwd(), filename)
    shutil.move(src, dst)

```

Output: normalized data for given directory and filetype

## H.16. Plot Confusion Matrix

```

"""

```

```

Created on Wed Nov 18 15:48:46 2020

```

```

@author: AbbieEnders

```

```

"""

```

```

c = 'present'

```

```

b = 'not present'

```

```

# do you want the user to state a title for the graph? If not remove line 11

```

```

# title = 'Confusion Matrix for Carboxylica Identification'

```

```

import numpy as np

```

```

import matplotlib.pyplot as plt

```

```

from matplotlib import cm

```



```

conf_arr = [[5,0],
            [0,5]]

norm_conf = []
for i in conf_arr:
    a = 0
    tmp_arr = []
    a = sum(i, 0)
    for j in i:
        tmp_arr.append(float(j)/float(a))
    norm_conf.append(tmp_arr)

fig = plt.figure()
plt.clf()
ax = fig.add_subplot(111)
ax.set_aspect(1)
res = ax.imshow(np.array(conf_arr), cmap=cm.summer,
                interpolation='nearest', vmin=0, vmax=5)

for x in range(2):
    for y in range(2):
        ax.annotate("{:.0f}".format(conf_arr[x][y]), xy=(y, x),
                    horizontalalignment='center',
                    verticalalignment='center',
                    color='black',
                    size='12')

cb = fig.colorbar(res)
#plt.title(title)
plt.xlabel('Predicted Group')
plt.ylabel('Actual Group')
plt.xticks(range(2), [c,b])
plt.yticks(range(2), [c,b])
plt.savefig('perfect_10000_0_01.svg', format='svg')
plt.show()

```

**Output:** saved confusion matrix for provided results

## H.17. Remove Random Files Until Directories are Equal in Files

```
""""
Created on Sun Nov 8 14:14:59 2020
@author: AbbieEnders
""""
'''
#####
####
#####TO RUN IN THE CONSOLE
#####
#####
####
python random_number_files.py [path_to_directory_1] [path_to_directory_2]
'''

# imports
import os
import sys
import random
import shutil

# randomly delete files in folder that has excess images
dir_1 = os.path.join(os.getcwd(), sys.argv[1]) #this should be POSITIVE\CONTAINING
functional group for naming purposes
dir_2 = os.path.join(os.getcwd(), sys.argv[2]) #this should be NEGATIVE\NOT CONTAINING
functional group for naming purposes
dest_dir = os.path.join(os.getcwd(), sys.argv[3])

len_dir_one = len(os.listdir(dir_1))
len_dir_two = len(os.listdir(dir_2))
num_of_validation_files = 5
i = 0
# first we will randomly move five files each to be used as validation files

while i < num_of_validation_files:
    print('!m here') #is this meaningful?
    f_d1 = os.path.join(dir_1, random.choice(os.listdir(dir_1))) # file for directory 1
    f_d2 = os.path.join(dir_2, random.choice(os.listdir(dir_2))) # file for directory 2
```

```

shutil.move(f_d1, dest_dir)
shutil.move(f_d2, dest_dir)
i += 1

if len_dir_one > len_dir_two:
    # get difference of files in the two directories and remove random files to
    # get the directories to equal lengths
    dif = len_dir_one - len_dir_two
    for file in random.sample(os.listdir(dir_1),dif):
        os.remove(os.path.join(dir_1,file))
elif len_dir_one < len_dir_two:
    # get difference of files in the two directories and remove random files to
    # get the directories to equal lengths
    dif = len_dir_two - len_dir_one
    for file in random.sample(os.listdir(dir_2),dif):
        os.remove(os.path.join(dir_2, file))
else:
    exit()

```

Output: given two directories, equivalent number of files will be present in each

## H.18. Train Model to Predict Functional Group Subprocess Code

```

"""
Created on Mon Nov 16 20:14:05 2020
@author: AbbieEnders
"""

# run the machine learning for each compound
import os
import shutil

cwd = os.getcwd()
top_dir = ['nitrile', 'ketone', 'ether',
           'amine', 'amide', 'alkyne', 'alkane', 'alkene', 'alcohol',
           'nitro', 'alkyl_halide', 'acyl_halide']
for d in top_dir:
    #shutil.copytree(os.path.join(cwd,'scripts'), os.path.join(cwd,d,'scripts'))
    os.chdir(os.path.join(cwd, d))
    os.system('python scripts/retrain.py --image_dir functional_group --output_graph 10000-
0.001_graph.pb'+

```

```

' --output_labels 10000-0.001_labels.txt --summaries_dir 10000-0.001_sum --
how_many_training_steps'+
' 10000 --learning_rate 0.001 --architecture inception_v3')

```

Output: trained models from retrain python code

## H.19. Classify Unknown Images in Batches

```

"""

```

```

Created on Tue Nov 17 08:37:12 2020

```

```

@author: AbbieEnders

```

```

"""

```

```

import tensorflow as tf, sys

```

```

import csv

```

```

from os import walk

```

```

import numpy as np

```

```

image_dir = sys.argv[1]

```

```

output_file = sys.argv[2]

```

```

#DO YOU WANT THESE PATHS TO BE ARGUEMENTS TOO?

```

```

graph_path = '20000-0.01_graph.pb'

```

```

labels_path = '20000-0.01_labels.txt'

```

```

#Create list of files in given directory

```

```

image_list = []

```

```

for (dirpath, dirnames, filenames) in walk(image_dir):

```

```

    image_list.extend(filenames)

```

```

    break

```

```

#Open output (.csv) file to be written to

```

```

csv_header = ['Image Name', 'Containing Fn Group', 'Not Containing Fn Group']

```

```

with open(output_file, 'w') as csvFile:

```

```

    writer = csv.writer(csvFile)

```

```

    writer.writerow(csv_header)

```

```

    csvFile.close()

```

```

for image_path in image_list:

```

```

    # Read in the image_data

```

```

    image_name = image_path # save image name for column

```

```

    image_path = image_dir + '/' + image_path #TEST: build correct img path

```

```

image_data = tf.gfile.FastGFile(image_path, 'rb').read()

# Loads label file, strips off carriage return
label_lines = [line.rstrip() for line
                in tf.gfile.GFile(labels_path)]

# Unpersists graph from file
with tf.gfile.FastGFile(graph_path, 'rb') as f:
    graph_def = tf.GraphDef()
    graph_def.ParseFromString(f.read())
    _ = tf.import_graph_def(graph_def, name="")

# Feed the image_data as input to the graph and get first prediction
with tf.Session() as sess:
    softmax_tensor = sess.graph.get_tensor_by_name('final_result:0')
    predictions = sess.run(softmax_tensor,
                           {'DecodeJpeg/contents:0': image_data})
# Sort to show labels of first prediction in order of confidence. Later, sorted by image
name.
    top_k = predictions[0].argsort()[-len(predictions[0]):::-1]
    score_list = [] # Clear list
    score_list[0].append(image_name)

    for node_id in np.sort(top_k):
        human_string = label_lines[int(node_id)]
        score = predictions[0][int(node_id)]
        score = format(score, '.5f') #Format score
        print('TEST-----')
        score_list[0].append(score)
        # print('%s (score = %.5f)' % (human_string, score))

    with open(output_file, 'a') as csvFile:
        writer = csv.writer(csvFile)
        writer.writerows(score_list)
        csvFile.close()

    image_path = ''

# Ensure output file is properly closed
if not csvFile.closed:
    csvFile.close()

```

Output: predicted class for the unknown images

## H.20. GitHub Resources

<https://github.com/Ohio-State-Allen-Lab>

<https://github.com/AbbieEnders/AbbieEnders>

## **Appendix I. Sea Surface and Bulk Sampling of Atlantic Ocean and Banana River in Florida in January 2023**

Work done on this trip was done alongside Jessica B. Clark and Nicole M. North.

### **I.1. Precleaning the Glass Sample Vessels**

The glass sample vials were cleaned via dishwasher with a surfactant solution followed by another rinse in the dishwasher with only water. Both times the dishwasher was run on the “Sanitize” mode. When removed from the dishwasher the glass sample vessels were only handled with nitrile gloves.

### **I.2. SSML/ Surface Sampling**

The surface water was sampled using a method established by Harvey and Burzell.<sup>239</sup> A glass slide (MilliporeSigma, unframed, H × W × D 200 mm × 260 mm × 4 mm), held by a clip, was submerged vertically into the water and withdrawn quickly (~1 second). Film was transferred from the glass to the glass storage vessel using a repurposed squeegee. Repeated sampling enabled collection of sufficient volume.

### **I.3. Sea Foam Sampling**

Sea foam was collected by placing the glass slide on the surface of the water and “picking” up the foam. Using a squeegee, the foam was transferred to a glass storage vessel.

### **I.4. Bulk Ocean Sampling**

The jars were rinsed 10 times in the sample water. The lid was then used to cover the glass vessel as it was submerged under the surface approximately 6 inches, the lid was then removed to collect a bulk sea sample without sampling the SSML.



Figure 61. Photo of Abbie rinsing glass vessel in accordance with protocol for collecting bulk samples.





Figure 62. Picture of Abbie (light blue) with glass slide on Banana River, assisted by Nicole (gray long sleeve) holds the kayak steady and Jess (gray short sleeve) operates the squeegee and glass storage vessel.




Figure 63. Jess squeegees the glass slide, held by Abbie, after it was dipped in the Banana River during surface sampling.



Figure 64. Nicole (left) and Jess (right) collect sea foam/surface in Atlantic ocean by placing slide on surface squeegeeing off the water into glass storage vessel.

## Appendix J. Permissions

Chapter 3: Reproduced in part with permission from 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. Copyright 2021 American Chemical Society.

Home Help Live Chat Sign in Create Account

**Functional Group Identification for FTIR Spectra Using Image-Based Machine Learning Models**

Author: Abigail A. Enders, Nicole M. North, Chase M. Fensore, et al

Publication: Analytical Chemistry

Publisher: American Chemical Society

Date: Jul 1, 2021

Copyright © 2021, American Chemical Society

**PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE**

This type of permission/license, instead of the standard Terms and Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from {COMPLETE REFERENCE CITATION}. Copyright {YEAR} American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your RightsLink request. No additional uses are granted (such as derivative works or other editions). For any uses, please submit a new request.

If credit is given to another source for the material you requested from RightsLink, permission must be obtained from that source.

[BACK](#) [CLOSE WINDOW](#)

