Saccharide concentration prediction from proxy-sea surface microlayer samples analyzed via ATR-FTIR spectroscopy and quantitative machine learning

Abigail A. Enders, Nicole M. North, Jessica B. Clark, Heather C. Allen*

Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States

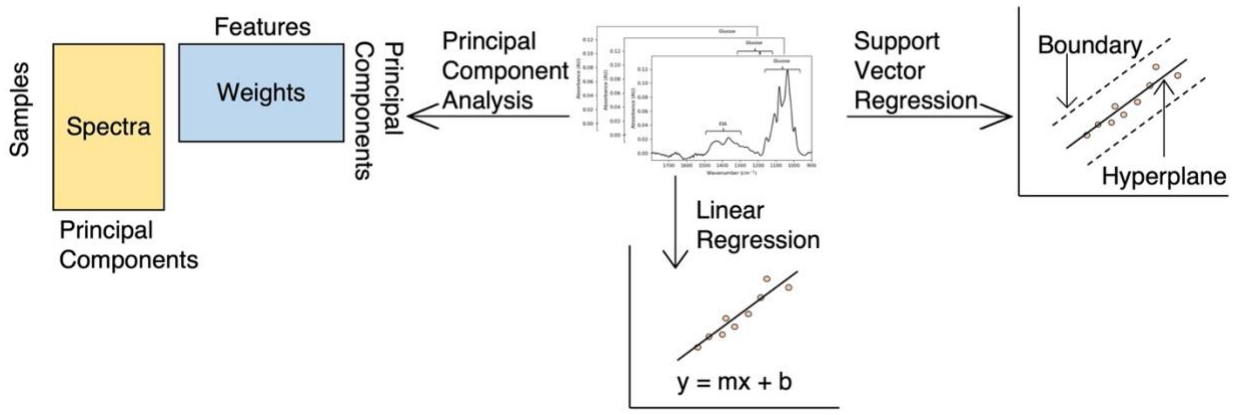
Corresponding author

* Heather C. Allen, allen@chemistry.ohio-state.edu

Abstract

The physical and chemical properties of the sea surface microlayer (SSML) are dynamic and complex. With an enrichment of organics from dissolved organic carbon (DOC) and many mechanisms for their release into the atmosphere, high-throughput analysis of SSML samples is necessary. Collection of more detailed information about the SSML would enable greater understanding of the release of ice nucleating and cloud condensation particles and provide critical feedback for climate models. The work presented herein details an investigation to determine the most accurate and precise machine learning (ML) model for analyzing SSML samples. Support vector regression (SVR) models predict the true saccharide concentration best and we evaluate unknown SSML samples using the model to predict the amount of carbohydrate present. Model predictions were 60-90 mM saccharide concentrations from SSML samples. Our work presents an application combining fast spectroscopic techniques with ML to analyze SSML chemistry more efficiently, without sacrificing accuracy and precision.

Graphical Abstract



Keywords
Machine learning, sea surface microlayer, ocean chemistry, spectroscopy, support vector regression

Introduction

The sea surface microlayer (SSML) is a multifaceted, deeply complex region of the ocean.[1–7] As the interface between the Earth's atmosphere and ocean, the SSML performs vital functions that affect climate[5,8–10] and ice formation.[4,11–13] Because of unique interfacial anisotropy,[14–17] the physical and chemical properties of the SSML are of interest for their divergence from bulk water behavior. Generally, the SSML is enriched with saccharides, lipids, and proteins that are all components of dissolved organic carbon (DOC).[18–22] Understanding the chemical composition of the SSML provides insight into the biological activity and productivity within the SSML and enables predictions of cloud condensation[23] or ice nucleation,[4] ultimately aiding climatological models.[24–27] 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,[5,28,29] 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 sugar 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.[30]

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,[31,32] novel drug discovery,[33,34] catalyst optimization,[35,36] and clean energy production[37,38] 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. Takamura and colleagues explored methods to identify donor biological sex from urine samples.[39] 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.[40] 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.[41] 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.[42] PCA is common in chemometrics;[43] examples in the literature include identifying trace elements in wheat,[44] analysis of time of flight-secondary ion mass spectra from organic monolayers,[45] detecting sparse compounds via FTIR spectra,[46] and identifying peak shape changes in chromatography.[47] 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.[48] 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[49] and lead[50] 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.[51] 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.[52] As another example, ATR-FTIR and SVR were employed by Chen et al. 2022 to predict bio-oil characteristics quickly.[53] Our review of the literature and ML methods indicates that the SVR model will perform best for predicting sugar 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 with a relatively simple training dataset. 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 sugar concentration from FTIR spectra of proxy-SSML samples using ML.

Methods

Training Solution Preparation, Data Collection, and Data Preprocessing

All chemicals were used as received and all solutions requiring water were prepared using ultrapure water (18 mΩ) from a MilliQ system. A 1M stock solution of glucose (Sigma Aldrich, ≥99.5% (GC)) in ultrapure water was prepared. A 5 mg/mL stock solution of ovalbumin (Sigma

Aldrich, 62-88%, agarose gel electrophoresis) in ultrapure water was prepared. The solution matrix was prepared 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 sugar have IR responses from 1800 to 900 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.[26,27,29] 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 from 4000 to 450 cm^{-1} with a resolution of 1 cm^{-1}. Spectra were converted to absorbance with a water background using the established relationship of $-\log(R/R_o)$. Background correction was done using a linear fit model for the baseline to correct for inconsistent baseline between 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 1).

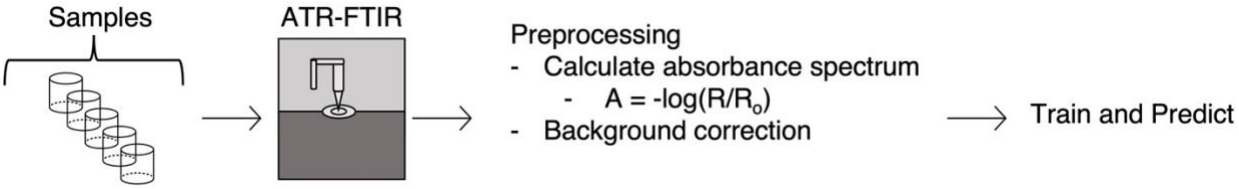


Figure 1. Schematic flow chart of data collection process to the ML pipeline.

Proxy-Sample and Real Sea Surface Water Preparation and Sampling

A stock proxy-solution was prepared to have 0.1 M sucrose, 0.1 M glucose, 0.5 mg/mL ESA, 3.323 mg/mL BSA, and 0.1 M 1-butanol. 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. Sea and river samples from Cocoa Beach, Florida were collected in January 2023. ATR-FTIR spectra were acquired for the 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.[54] Extracted DOC was analyzed via gas chromatography-mass spectrometry (GC-MS) to identify organic components (SI).

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. Principal component analysis (PCA) was used to elucidate any relationships between the data 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 approach 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 support vector regression (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 ε, threshold tolerance,

and C, regularization parameters to reach a minimization of mean squared error (MSE) (SI). 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 (SI). 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[1] 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.

Results and Discussion

The chemical complexity of the SSML is explored via ATR-FTIR spectroscopy and quantitative machine learning approaches to develop a simpler 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 we were not concerned about in the presently studied concentration ranges. Figure 2 is an examplespectrum 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 the SI.

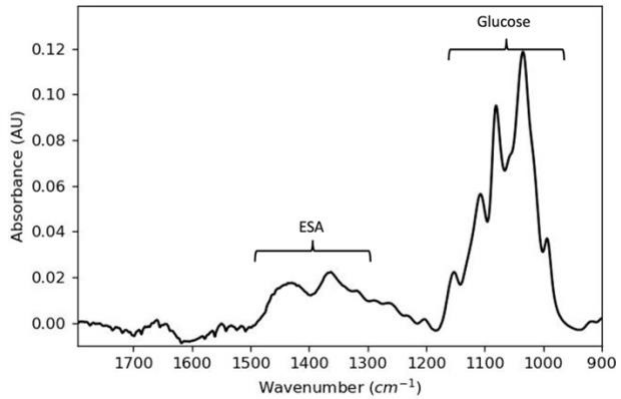


Figure 2. ATR-FTIR spectrum of 0.6 M glucose and 2 mg/mL egg serum albumin. The labels are provided to emphasize that the components do not compound on one another and are well resolved, despite being in a similar wavenumber region.

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 3). 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.

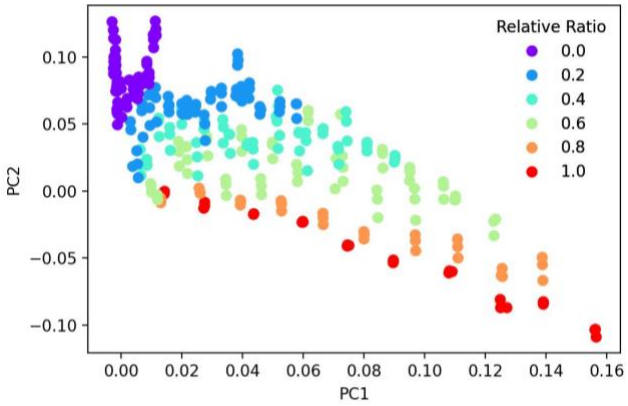


Figure 3. Principal components (PCs) one and two from the data dimensionality reduction performed using principal component analysis (PCA). The relative ratio is respective to glucose. Solutions with a relative ratio of '1' have no ESA. PC1 mainly captures glucose response and PC2 mainly captures ESA response.

Linear regression (LR) provides a mathematically simple fitting of the training data but does not accurately predict on more complex samples (Figure 4). 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 4, 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 cm^{-1}, the LR is performed using all wavenumbers from 1800 to 900 cm^{-1}, which eliminates feature selection biases.

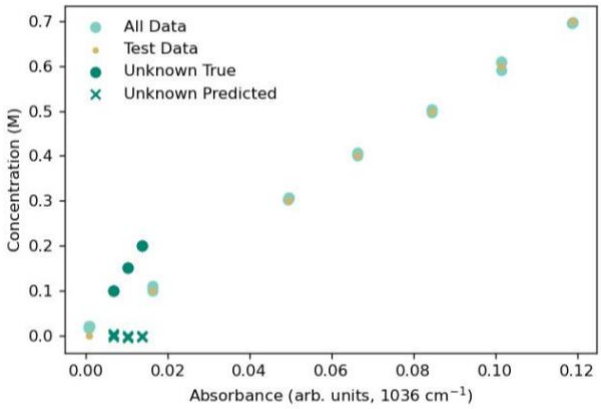


Figure 4. Linear regression (LR) fits the experimental training data well with a 100 % R^2 and no mean squared error. Proxy sample sugar concentrations are not correctly predicted, as shown with the teal 'X' demarcating the prediction.

In comparison to the LR, the support vector regression (SVR) fits the training data and closely predict the concentration of sugar in more complex solutions (Figure 5). 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 5.

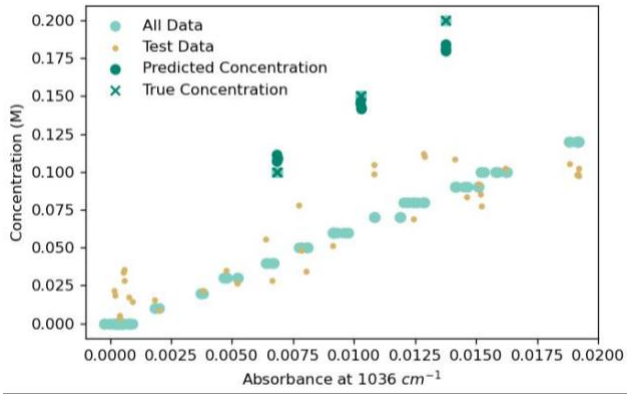


Figure 5. Support vector regression (SVR) results show that the test data accurately follows the training data. Predicted concentrations for the known complex samples are much closer to the true concentration. The training results in an R^2 of 97.1%.

The SVR and LR models were directly compared via the relative difference in the predicted versus true concentration of saccharides (Table 1). 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.

Table 1. Predicted sugar concentration (M) in more complex samples containing glucose, sucrose, ESA, bovine serum albumin (BSA), and 1-butanol were 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 saccharide (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

The LR and SVR model fit results are presented in Table 2 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.

Table 2. 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

We analyzed the 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 the SI. All the samples were predicted to have concentrations of saccharides in the mid to high mM range (70-100 mM) (Figure 6). Literature values range from 10-25 mM;[55] 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 1) 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 (SI). 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 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.

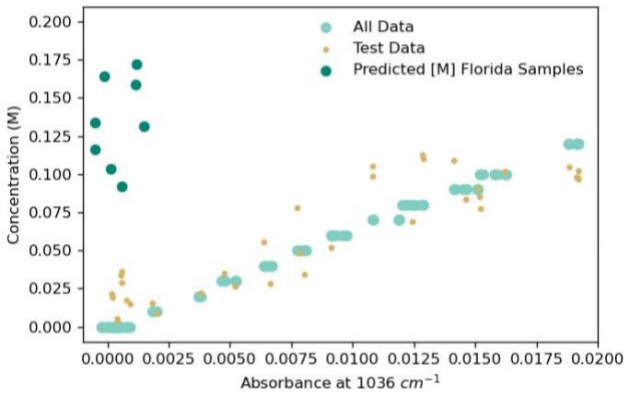


Figure 6. Support vector regression (SVR) model predictions on unknown field samples from Cocoa Beach, Florida. The predicted concentrations are closely aligned with the training and test data, although they are in the low absorbance range.

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 3). 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 (SI). 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 (SI).

Table 3. 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, nitrile	alcohol, alkyne, amide*, aromatic*, nitrile*, nitro*
Absent	acyl halide, alcohol, aldehyde, alkane, alkyl halide, alkyne, amine, aromatic, carboxylic acid, ether, ketone, nitro	acyl halide, aldehyde, alkane, alkene, alkyl halide, amine, carboxylic acid, ester, ether, ketone, methyl

Sensitivity, specificity, positive predictive value, and negative predictive value were calculated according to the definitions presented by Trevethan in 2017 (Table 4).[56] 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 TROPOS field 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.

Table 4. 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

Overall, the results from the CNN provide contextualization of the samples without the requirement of a lengthy extraction process to identify DOC (SI). 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.[55] 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.

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 aCNN 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.

Supplemental Information

Figure of all training spectra; FTIR and GC-MS spectra of Cocoa Beach, Florida field samples; figure of optimized ε and C for SVR; figure of optimized train-test split for LR and SVR

Acknowledgements

A.A.E. and H.C.A. acknowledge funding from the National Science Foundation through the Center for Aerosol Impacts on Chemistry of the Environment (CAICE) under Grant No. CHE-1801971. N.M.N. acknowledges funding support from NASA's Future Investigators of NASA Earth and Space Science Technology (FINESST) grant number 20-PLANET20-0067. J.B.C.

acknowledges funding support from the National Science Foundation through Grant No. CHE-2102313.

References

(1) Carlson, D. J. Dissolved Organic Materials in Surface Microlayers: Temporal and Spatial Variability and Relation to Sea State. *Limnol. Oceanogr.* **1983**, *28* (3), 415–431. https://doi.org/10.4319/lo.1983.28.3.0415.

(2) Cunliffe, M.; Engel, A.; Frka, S.; Gašparović, B.; Guitart, C.; Murrell, J. C.; Salter, M.; Stolle, C.; Upstill-Goddard, R.; Wurl, O. Sea Surface Microlayers: A Unified Physicochemical and Biological Perspective of the Air–Ocean Interface. *Prog. Oceanogr.* **2013**, *109*, 104–116. https://doi.org/10.1016/j.pocean.2012.08.004.

(3) Engel, A.; Bange, H. W.; Cunliffe, M.; Burrows, S. M.; Friedrichs, G.; Galgani, L.; Herrmann, H.; Hertkorn, N.; Johnson, M.; Liss, P. S.; Quinn, P. K.; Schartau, M.; Soloviev, A.; Stolle, C.; Upstill-Goddard, R. C.; van Pinxteren, M.; Zäncker, B. The Ocean's Vital Skin: Toward an Integrated Understanding of the Sea Surface Microlayer. *Front. Mar. Sci.* **2017**, *4* (MAY), 1–14. https://doi.org/10.3389/fmars.2017.00165.

(4) Chance, R. J.; Hamilton, J. F.; Carpenter, L. J.; Hackenberg, S. C.; Andrews, S. J.; Wilson, T. W. Water-Soluble Organic Composition of the Arctic Sea Surface Microlayer and Association with Ice Nucleation Ability. *Environ. Sci. Technol.* **2018**, *52* (4), 1817–1826. https://doi.org/10.1021/acs.est.7b04072.

(5) Cochran, R. E.; Laskina, O.; Trueblood, J. V.; Estillore, A. D.; Morris, H. S.; Jayarathne, T.; Sultana, C. M.; Lee, C.; Lin, P.; Laskin, J.; Laskin, A.; Dowling, J. A.; Qin, Z.; Cappa, C. D.; Bertram, T. H.; Tivanski, A. V.; Stone, E. A.; Prather, K. A.; Grassian, V. H. Molecular Diversity of Sea Spray Aerosol Particles: Impact of Ocean Biology on Particle Composition and Hygroscopicity. *Chem* **2017**, *2* (5), 655–667. https://doi.org/10.1016/j.chempr.2017.03.007.

(6) Ault, A. P.; Moffet, R. C.; Baltrusaitis, J.; Collins, D. B.; Ruppel, M. J.; Cuadra-Rodriguez, L. A.; Zhao, D.; Guasco, T. L.; Ebben, C. J.; Geiger, F. M.; Bertram, T. H.; Prather, K. A.; Grassian, V. H. Size-Dependent Changes in Sea Spray Aerosol Composition and Properties with Different Seawater Conditions. *Environ. Sci. Technol.* **2013**, *47* (11), 5603–5612. https://doi.org/10.1021/es400416g.

(7) Bertram, T. H.; Cochran, R. E.; Grassian, V. H.; Stone, E. A. Sea Spray Aerosol Chemical Composition: Elemental and Molecular Mimics for Laboratory Studies of Heterogeneous and Multiphase Reactions. *Chem. Soc. Rev.* **2018**, *47* (7), 2374–2400. https://doi.org/10.1039/c7cs00008a.

(8) Abraham, J. P.; Baringer, M.; Bindoff, N. L.; Boyer, T.; Cheng, L. J.; Church, J. A.; Conroy, J. L.; Domingues, C. M.; Fasullo, J. T.; Gilson, J.; Goni, G.; Good, S. A.; Gorman, J. M.; Gouretski, V.; Ishii, M.; Johnson, G. C.; Kizu, S.; Lyman, J. M.; Macdonald, A. M.; Minkowycz, W. J.; Moffitt, S. E.; Palmer, M. D.; Piola, A. R.; Reseghetti, F.; Schuckmann, K.; Trenberth, K. E.; Velicogna, I.; Willis, J. K. A Review of Global Ocean Temperature Observations: Implications for Ocean Heat Content Estimates and Climate Change. *Rev. Geophys.* **2013**, *51* (3), 450–483. https://doi.org/10.1002/rog.20022.

(9) Burrows, S. M.; Ogunro, O.; Frossard, A. A.; Russell, L. M.; Rasch, P. J.; Elliott, S. M. A Physically Based Framework for Modeling the Organic Fractionation of Sea Spray Aerosol

from Bubble Film Langmuir Equilibria. *Atmospheric Chem. Phys.* **2014**, *14* (24), 13601–13629. https://doi.org/10.5194/acp-14-13601-2014.

(10) Cheng, S.; Li, S.; Tsona, N. T.; George, C.; Du, L. Insights into the Headgroup and Chain Length Dependence of Surface Characteristics of Organic-Coated Sea Spray Aerosols. *ACS Earth Space Chem.* **2019**, *3* (4), 571–580. https://doi.org/10.1021/acsearthspacechem.8b00212.

(11) Wilson, T. W.; Ladino, L. A.; Alpert, P. A.; Breckels, M. N.; Brooks, I. M.; Browse, J.; Burrows, S. M.; Carslaw, K. S.; Huffman, J. A.; Judd, C.; Kilthau, W. P.; Mason, R. H.; McFiggans, G.; Miller, L. A.; Najera, J. J.; Polishchuk, E.; Rae, S.; Schiller, C. L.; Si, M.; Temprado, J. V.; Whale, T. F.; Wong, J. P. S.; Wurl, O.; Yakobi-Hancock, J. D.; Abbatt, J. P. D.; Aller, J. Y.; Bertram, A. K.; Knopf, D. A.; Murray, B. J. A Marine Biogenic Source of Atmospheric Ice-Nucleating Particles. *Nature* **2015**, *525* (7568), 234–238. https://doi.org/10.1038/nature14986.

(12) Ting Katty Huang, W.; Ickes, L.; Tegen, I.; Rinaldi, M.; Ceburnis, D.; Lohmann, U. Global Relevance of Marine Organic Aerosol as Ice Nucleating Particles. *Atmospheric Chem. Phys.* **2018**, *18* (15), 11423–11445. https://doi.org/10.5194/acp-18-11423-2018.

(13) DeMott, P. J.; Hill, T. C. J.; McCluskey, C. S.; Prather, K. A.; Collins, D. B.; Sullivan, R. C.; Ruppel, M. J.; Mason, R. H.; Irish, V. E.; Lee, T.; Hwang, C. Y.; Rhee, T. S.; Snider, J. R.; McMeeking, G. R.; Dhaniyala, S.; Lewis, E. R.; Wentzell, J. J. B.; Abbatt, J.; Lee, C.; Sultana, C. M.; Ault, A. P.; Axson, J. L.; Martinez, M. D.; Venero, I.; Santos-Figueroa, G.; Stokes, M. D.; Deane, G. B.; Mayol-Bracero, O. L.; Grassian, V. H.; Bertram, T. H.; Bertram, A. K.; Moffett, B. F.; Franc, G. D. Sea Spray Aerosol as a Unique Source of Ice Nucleating Particles. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (21), 5797–5803. https://doi.org/10.1073/pnas.1514034112.

(14) Carter-Fenk, K. A.; Dommer, A. C.; Fiamingo, M. E.; Kim, J.; Amaro, R. E.; Allen, H. C. Calcium Bridging Drives Polysaccharide Co-Adsorption to a Proxy Sea Surface Microlayer. *Phys. Chem. Chem. Phys.* **2021**, *23* (30), 16401–16416. https://doi.org/10.1039/d1cp01407b.

(15) Yao, X.; Liu, Q.; Wang, B.; Yu, J.; Aristov, M. M.; Shi, C.; Zhang, G. G. Z.; Yu, L. Anisotropic Molecular Organization at a Liquid/Vapor Interface Promotes Crystal Nucleation with Polymorph Selection. *J. Am. Chem. Soc.* **2022**, *144* (26), 11638–11645. https://doi.org/10.1021/jacs.2c02623.

(16) Neal, J. F.; Rogers, M. M.; Smeltzer, M. A.; Carter-Fenk, K. A.; Grooms, A. J.; Zerkle, M. M.; Allen, H. C. Sodium Drives Interfacial Equilibria for Semi-Soluble Phosphoric and Phosphonic Acids of Model Sea Spray Aerosol Surfaces. *ACS Earth Space Chem.* **2020**, *4* (9), 1549–1557. https://doi.org/10.1021/acsearthspacechem.0c00132.

(17) Vazquez De Vasquez, M. G.; Carter-Fenk, K. A.; McCaslin, L. M.; Beasley, E. E.; Clark, J. B.; Allen, H. C. Hydration and Hydrogen Bond Order of Octadecanoic Acid and Octadecanol Films on Water at 21 and 1°c. *J. Phys. Chem. A* **2021**, *125* (46), 10065–10078. https://doi.org/10.1021/acs.jpca.1c06101.

(18) Myklestad, S. M. Dissolved Organic Carbon from Phytoplankton. In *Marine Chemistry*; Wangersky, P. J., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2000; pp 111–148. https://doi.org/10.1007/10683826_5.

(19) Lønborg, C.; Carreira, C.; Jickells, T.; Álvarez-Salgado, X. A. Impacts of Global Change on Ocean Dissolved Organic Carbon (DOC) Cycling. *Front. Mar. Sci.* **2020**, *7* (June), 1–24. https://doi.org/10.3389/fmars.2020.00466.

(20) Jayarathne, T.; Sultana, C. M.; Lee, C.; Malfatti, F.; Cox, J. L.; Pendergraft, M. A.; Moore, K. A.; Azam, F.; Tivanski, A. V.; Cappa, C. D.; Bertram, T. H.; Grassian, V. H.; Prather, K. A.; Stone, E. A. Enrichment of Saccharides and Divalent Cations in Sea Spray Aerosol during Two Phytoplankton Blooms. *Environ. Sci. Technol.* **2016**, *50* (21), 11511–11520. https://doi.org/10.1021/acs.est.6b02988.

(21) Gericke, A.; Hühnerfuss, H. Investigation of Z- and E-Unsaturated Fatty Acids, Fatty Acid Esters, and Fatty Alcohols at the Air/Water Interface by Infrared Spectroscopy. *Langmuir* **1995**, *11* (1), 225–230. https://doi.org/10.1021/la00001a039.

(22) Li, Y.; Shrestha, M.; Luo, M.; Sit, I.; Song, M.; Grassian, V. H.; Xiong, W. Salting up of Proteins at the Air/Water Interface. *Langmuir* **2019**, *35* (43), 13815–13820. https://doi.org/10.1021/acs.langmuir.9b01901.

(23) 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.

(24) 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.

(25) 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**, No. May, 1–27. https://doi.org/10.5194/acp-2018-70.

(26) 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.

(27) 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). https://doi.org/10.3390/atmos9060216.

(28) 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.

(29) 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.

(30) 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.

(31) 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.

(32) 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.
(33) 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.
(34) 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.
(35) 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.
(36) Ting, K. W.; Kamakura, H.; Poly, S. S.; Takao, M.; Siddiki, S. M. A. H.; Maeno, Z.; Matsushita, K.; Shimizu, K.; Toyao, T. Catalytic Methylation of M-Xylene, Toluene, and Benzene Using CO2 and H2 over TiO2-Supported Re and Zeolite Catalysts: Machine-Learning-Assisted Catalyst Optimization. *ACS Catal.* **2021**, *11* (9), 5829–5838. https://doi.org/10.1021/acscatal.0c05661.
(37) 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.jpclett.1c03526.
(38) 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.
(39) Takamura, A.; Halamkova, L.; Ozawa, T.; Lednev, I. K. Phenotype Profiling for Forensic Purposes: Determining Donor Sex Based on Fourier Transform Infrared Spectroscopy of Urine Traces. *Anal. Chem.* **2019**, *91* (9), 6288–6295. https://doi.org/10.1021/acs.analchem.9b01058.
(40) 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.
(41) 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.
(42) 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.
(43) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. 16.
(44) 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.
(45) 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.

(46) Hasegawa, T. Detection of Minute Chemical Species by Principal-Component Analysis. *Anal. Chem.* **1999**, *71* (15), 3085–3091. https://doi.org/10.1021/ac981430z.
(47) 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.
(48) 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.
(49) 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.
(50) Esbaugh, A. J.; Brix, K. V.; Mager, E. M.; De Schamphelaere, 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.
(51) 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.
(52) 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.
(53) 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.
(54) 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.
(55) 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.
(56) Trevethan, R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Front. Public Health* **2017**, *5*.

Supporting Information

Saccharide concentration prediction from proxy-sea surface microlayer samples analyzed via ATR-FTIR spectroscopy and quantitative machine learning

Abigail A. Enders†, Nicole M. North†, Jessica B. Clark†, and Heather C. Allen†*

†Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States

Corresponding author
* Heather C. Allen, allen@chemistry.ohio-state.edu

Table of Contents

All Python files can be accessed via the Allen Lab GitHub: https://github.com/Ohio-State-Allen-Lab/Sea-Surface-Microlayer-MachineLearning

Appendix A. Detailed explanation of machine learning specifications

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.[55] 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.

Appendix B. ATR-FTIR spectra of training and ocean samples and GC-MS of ocean samples

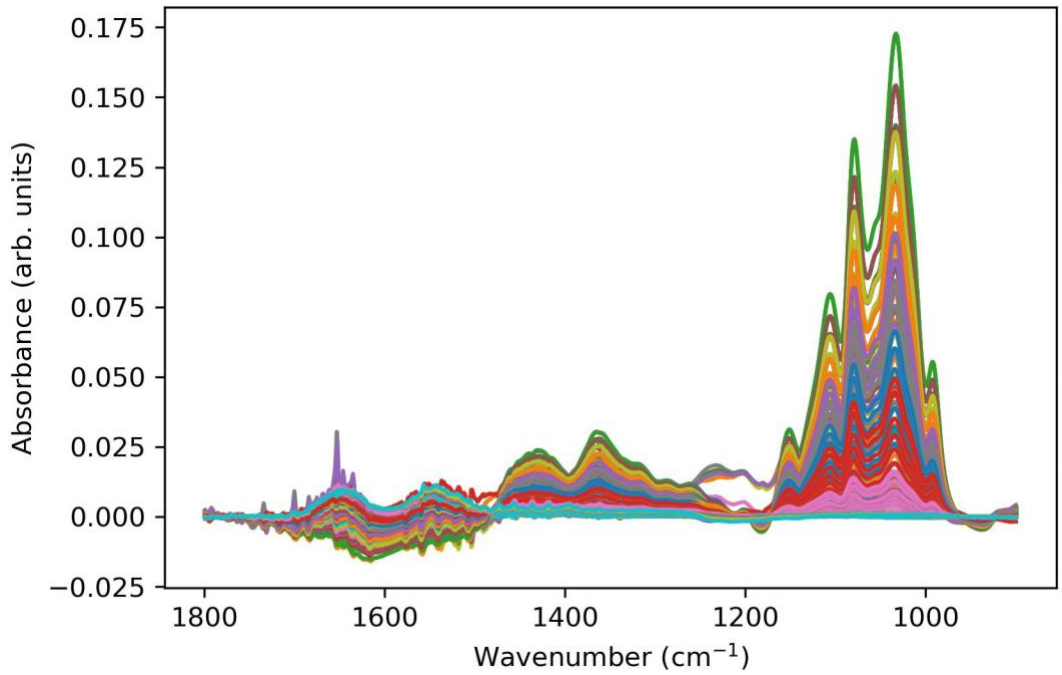


Figure S1. Composite spectra of all 100 samples used for training in each machine learning model.

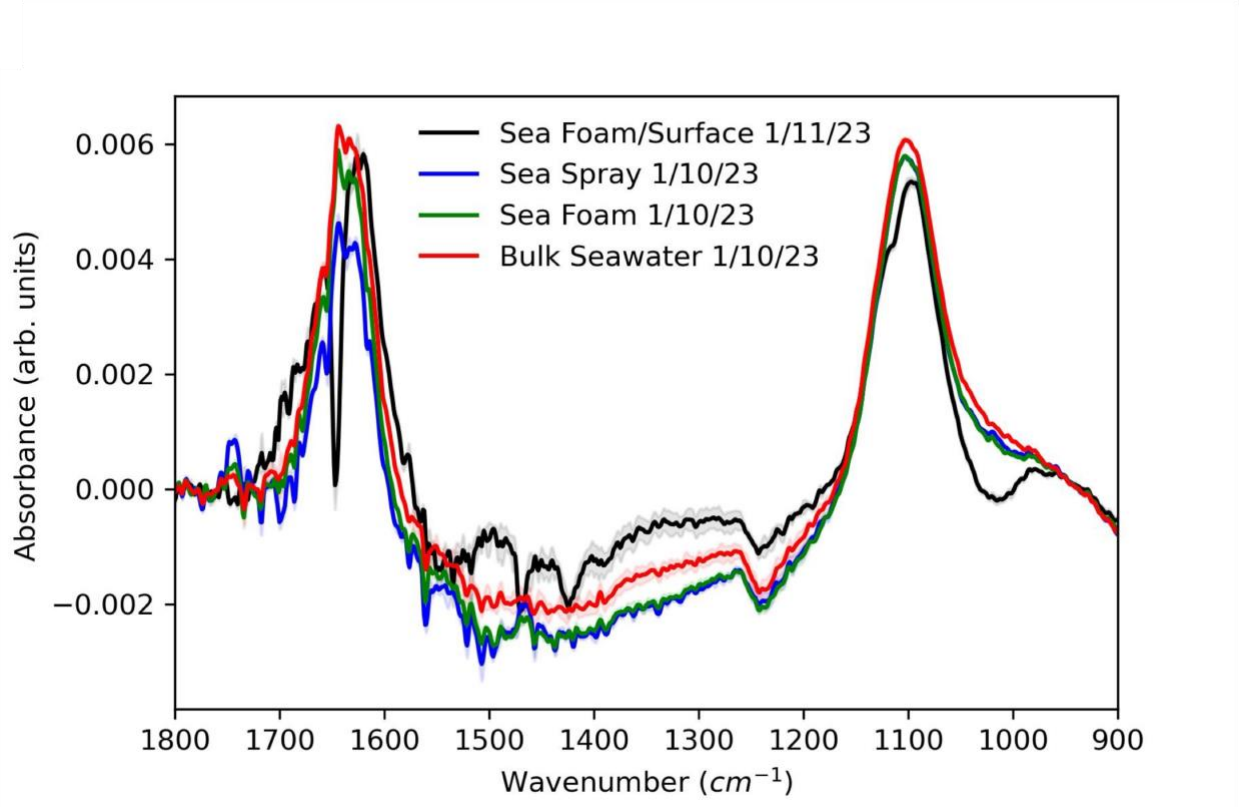


Figure S2. Average spectra of real ocean samples from Cocoa Beach, Florida. Standard deviation is shown but is approximately the thickness of the line.

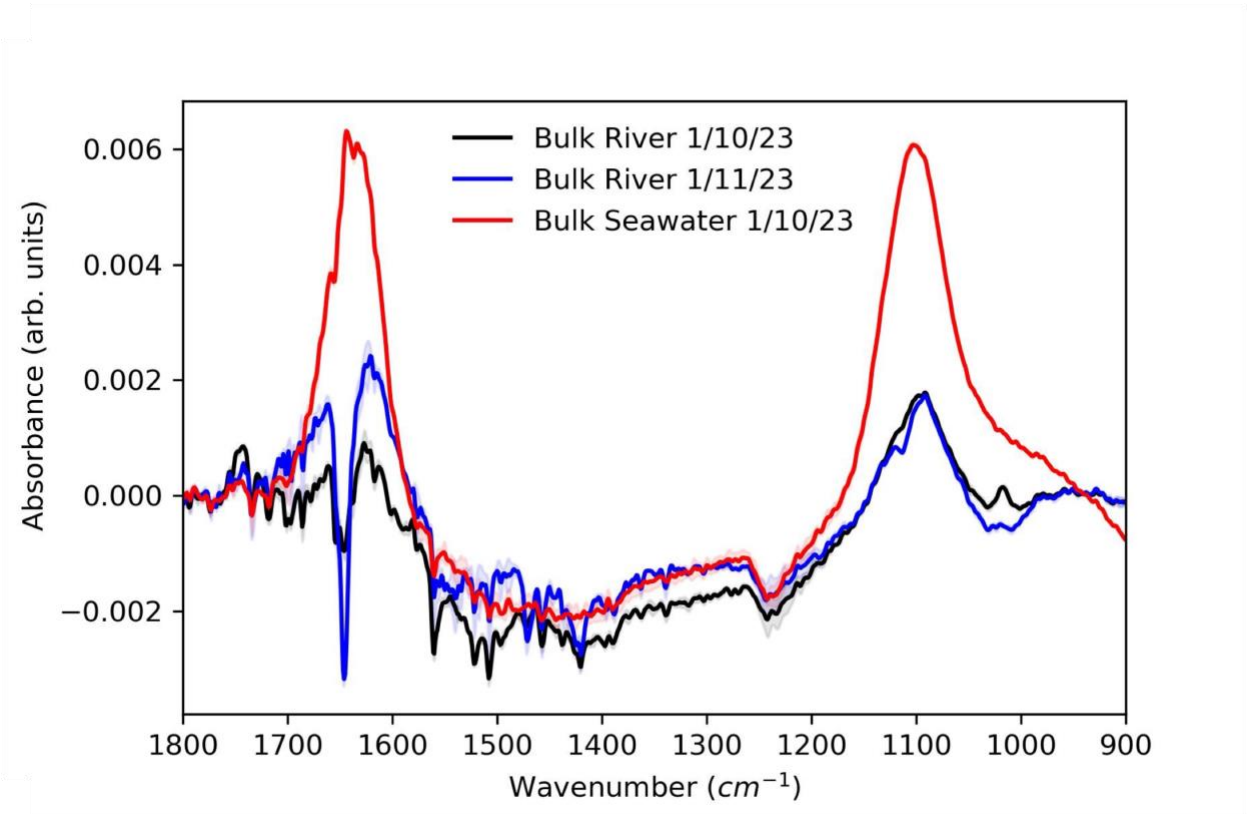
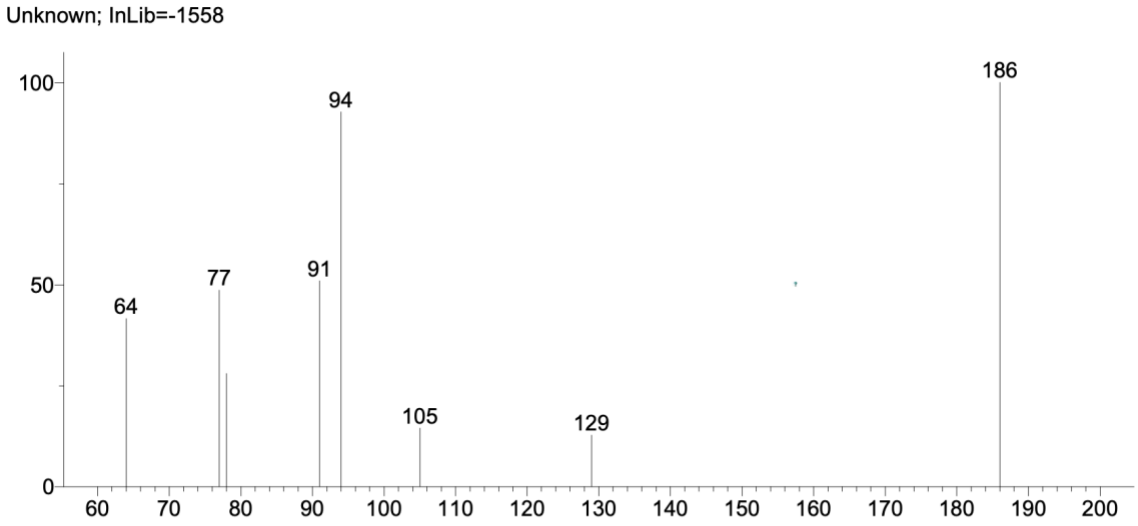


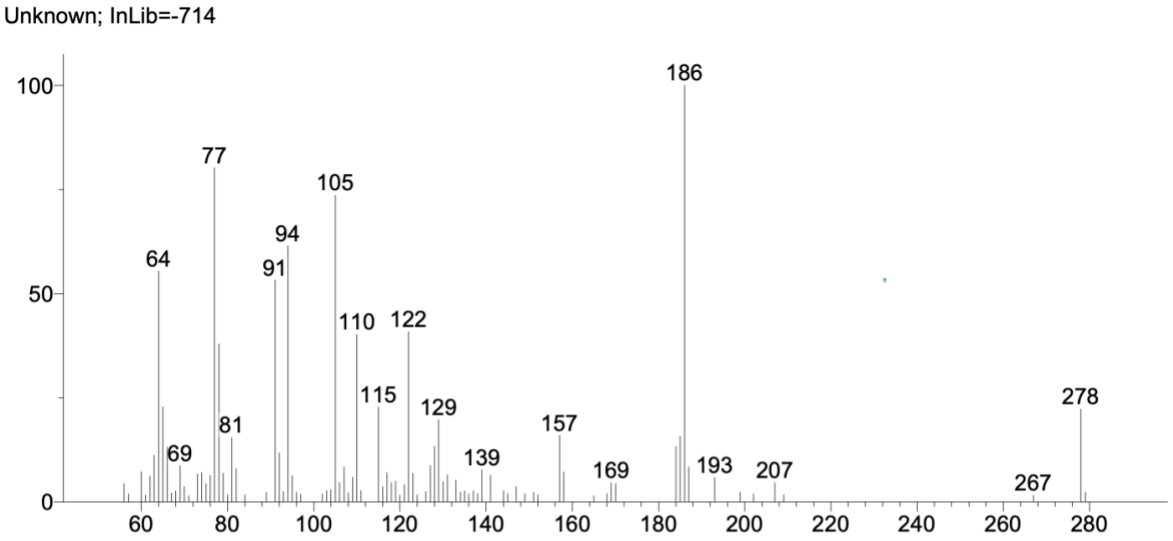
Figure S3. Average spectra of ocean and river samples from Cocoa Beach, Florida for comparison of sampling sites. Standard deviation is shown but is approximately the thickness of the line.



(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 S4. MS of GC retention for January 11, 2023, ocean surface sample from Cocoa Beach, Florida.



(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:

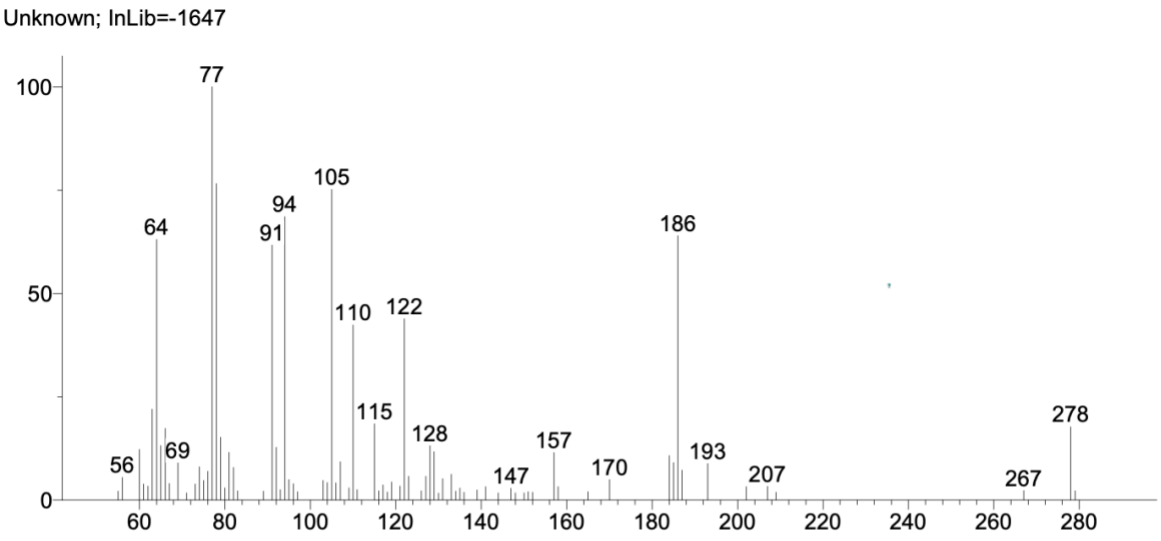
186 999	77 801	105 735	94 614	64 553	91 532	122 408	110 402	78 378	115 228

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 S5. MS of GC retention from January 11, 2023, river surface sample from the Banana River in Cocoa Beach, Florida.



(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 S6. MS of bulk surface water sample from Banana River in Cocoa Beach, Florida on January 10, 2023.

Appendix C. Optimization of support vector regression model

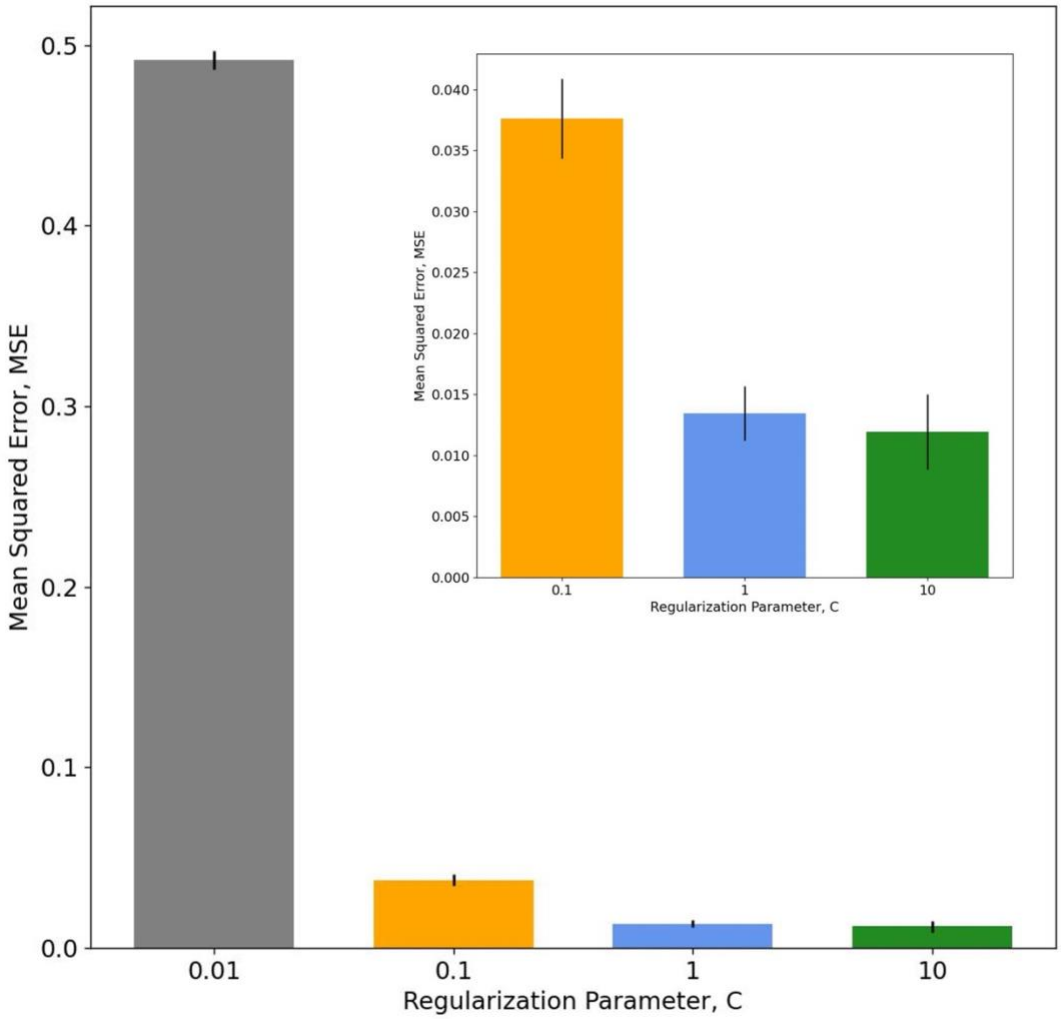


Figure S3. Optimization of regularization parameter C for the support vector regression (SVR). Variability shown is that of changing ε, the tolerance limit, which varies little compared to the optimization of C.

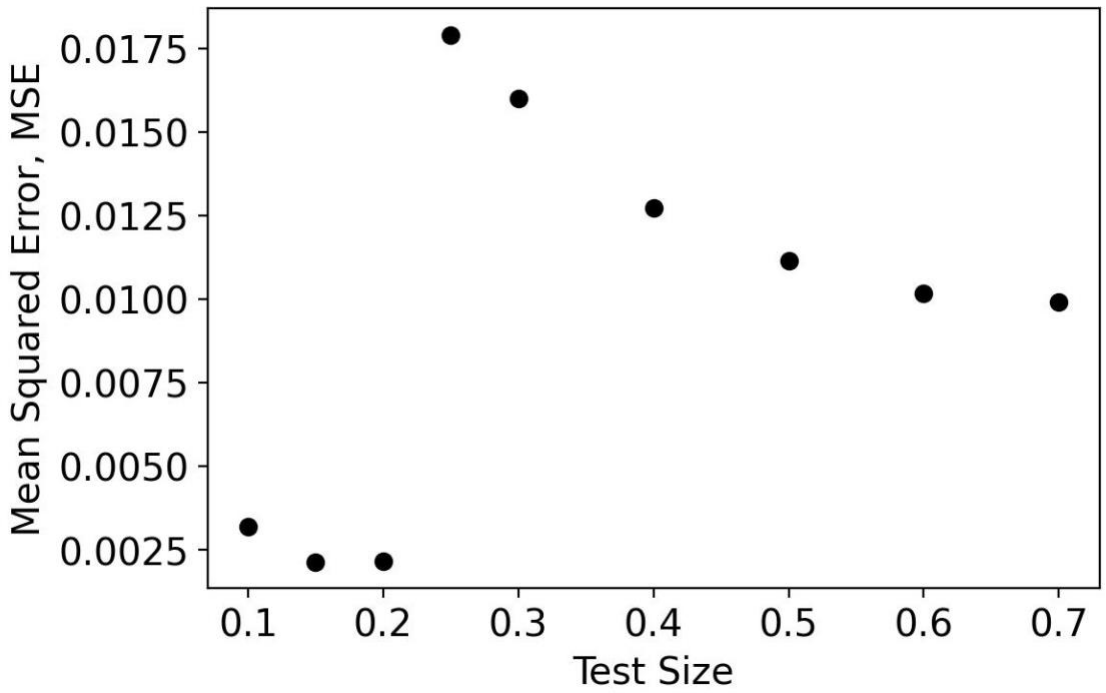


Figure S4. Optimization of train-test size split for the SVR. Minimization of MSE is prioritized for model performance. An 80/20 split minimizes MSE and has literature precedence.