Saccharide concentration prediction from proxy ocean samples analyzed via infrared spectroscopy and quantitative machine learning

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

Corresponding author

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

Abstract

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

Keywords
Machine learning, regression modeling, ocean chemistry, spectroscopy, support vector regression, decision trees, gradient boosted regression, multilayer perceptrons, k-nearest-neighbors

Introduction

The sea surface microlayer (SSML) is a multifaceted, deeply complex region of the ocean.[1–7] As the interface between the Earth's atmosphere and ocean, the SSML performs vital functions that affect climate[5,8–10] and ice formation.[4,11–13] Because of unique interfacial anisotropy,[14–17] the physical and chemical properties of the SSML are of interest for their divergence from bulk water behavior. Generally, the SSML is enriched with lipids, proteins, and saccharides (also referred to as sugars or carbohydrates) which contribute to the total dissolved organic carbon (DOC).[18–22] Understanding the chemical composition of the SSML provides insight into the biological activity and productivity within the SSML and enables predictions of cloud condensation[23] or ice nucleation,[4] ultimately aiding climatological models.[24–27] Recent analyses of saccharide concentrations in SSML have shown concentrations of about 500 nM from eight unique compounds.[20] The dynamic nature and chemical complexity of the SSML make monitoring the region difficult, and yet increasingly necessary.

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. Traditional methods to analyze SSML samples are typically 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 infrared (IR) spectroscopy methods, specifically, attenuated total reflectance Fourier transform infrared (ATR-FTIR) to estimate the saccharide concentration via machine learning (ML) implementations. IR methods provide information on chemical composition and concentration by probing the vibrations of chemical bonds, rather than relying on mass fragmentation. Identification and quantification of specific chemical classes from IR spectra is carried out by analyzing peaks characteristic to specific chemical bonds.[30] We note that the limit of detection for ATR-FTIR spectroscopy is higher than for mass spectroscopy, however the speed of analysis for this method is superior.

ML provides a unique avenue to explore relationships among data that cannot be otherwise deduced. The applications to improve or expand chemical systems via ML are broad and present throughout all chemistry fields. Materials design,[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. Advances in molecular dynamics in combination with machine learning have also paved the way for bridging the connection between molecular structure and physical characteristics.[39,40] 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.[41,42] Takamura and colleagues explored methods to identify donor biological sex from urine samples.[43] 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.[44] 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.[45] 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. 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.[46] Absorbance FTIR spectra generally follow a linear relationship of intensity with respect to concentration, which is advantageous for determining new sample composition. Recent work has utilized multiple LR to identify heavy metals, including investigating the effect of surface chemistry on vanadium[47] and lead[48] toxicity. However, the simplicity of the method ultimately restricts the model's usefulness in more complex, dynamic systems. The largest difference between Beer-Lambert Law linear regression and ML linear regression is that all features (in our work, wavenumbers) are used simultaneously to make the multi-variate model's assignments.[49]

Of the techniques considered here, SVR is the most mathematically advanced ML model.[50] SVR fits training data to the best function by minimizing the distance of each value from the fitting equation to be able to predict continuous values. Not all data is appropriate for SVR, but in cases where concentration is being predicted and is linearly correlated with absorbance, it can be a well-suited model. A 2020 report by Mohammadi and colleagues presented an application of SVR to predict different functional group fractions in crude oil.[51] As another example, ATR-FTIR and SVR were employed by Chen et al. 2022 to predict bio-oil characteristics quickly.[52]

The work described herein provides a discussion on an improved approach to monitoring the SSML. We explore ML approaches to achieve precise and accurate quantitative analysis of simplified proxies of glucose and egg serum albumin (ESA). Glucose is used as our saccharide proxy for training data as it is commonly observed in field measurements and saccharides are frequently reported as a concentration of glucose.[28,53,54] 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.[55–57] While an unlikely protein to find in field samples, ESA provides a complex matrix of amino acids that are abundant in the ocean's water column.[5,7,58–60] The use of ML in conjunction with vibrational spectroscopy enables greater exploration of chemical space and identifying connections between data. Our results present, to our knowledge, a first account of predicting saccharide concentration from FTIR spectra of ocean proxy samples using ML.

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

Both the protein and saccharide have IR absorbances from 1800 to 900 cm^{-1}. The peaks were well resolved, with minimal convolution. Inorganic salts were excluded in our matrix, but we provide spectra of the O-H stretching region in the SI to emphasize the limited effect that they have on the IR spectra. Concentrations were selected based on literature precedent from field study results.[26,27,29] Solutions were measured in triplicate via ATR-FTIR spectroscopy (PerkinElmer Spectrum 3) with a single beam KRS-5/diamond ATR assembly. Spectra were acquired in the "SingleBeam" mode without the use of a continuous reference and were detected using a liquid nitrogen cooled HgCdTe (MCT) detector over 32 scans (approximately one minute) from 4000 to 450 cm^{-1} with a resolution of 1 cm^{-1}. Spectra were converted to absorbance with a water-only background spectrum (R_o) using the established relationship of $-\log(R/R_o)$. Baseline correction was done using a linear fit model to correct for inconsistent baseline between measurements. Water-only backgrounds were obtained every 5 sample measurements. Triplicate measurements were used as individual spectra, rather than an average of the three, to provide more machine learning training and testing data (Figure 1).

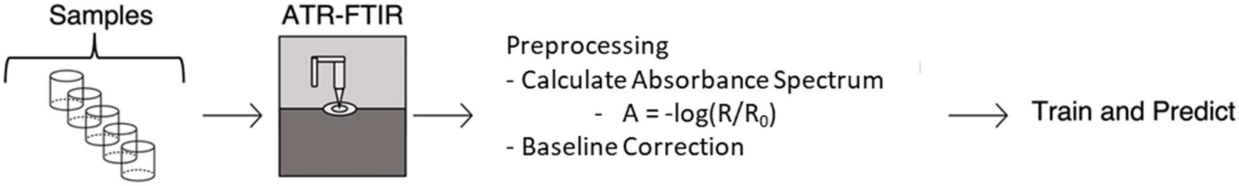

Figure 1. Schematic flow chart of data collection process to the ML pipeline.

Lab Generated Simplified and Ocean Proxy Sample Preparation and Sampling

To test the models' accuracies with increasing chemical complexity, ocean proxy samples were made in the lab with a greater diversity of chemical constituents than the simplified proxies. For these test data, stock ocean proxy-solution was prepared to have 0.1 M sucrose (Sigma Aldrich, ≥99.5% (GC)), 0.1 M glucose, 0.5 mg/mL ESA, 3.323 mg/mL bovine serum albumin (BSA) (Sigma Aldrich, ≥ 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.

	Ocean Proxy A	*Ocean Proxy B*	*Ocean Proxy C*
Concentration of Sucrose (M)	0.1	0.075	0.05
Concentration of Glucose (M)	0.1	0.075	0.05
Concentration of Saccharide (M)	0.2	0.150	0.1
Concentration of ESA (mg/mL)	0.5	0.375	0.250
Concentration of BSA (mg/mL	3.323	2.492	1.662
Concentration of 1-Butanol (M)	0.1	0.075	0.05

Table 1. Concentrations of all species in the lab-made ocean proxy samples for evaluation of model accuracy on more chemically diverse conditions.

Machine Learning Methods

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

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

Preprocessing

All data, the entire training set of simplified proxy and the ocean proxy samples were standardized using the scikit-learn standard scalar function. The function subtracts the mean of each feature (wavenumber) and divides each feature by the standard deviation of that feature. After standardization, the ocean proxy data was separated from the data that would then be used for training. The data was then split 70::30 into training and test/validation sets. The latter of which was then split 50::50 into testing and validation datasets. A random state was set to to split the data the same way every time into the training, testing, and validating datasets each time to ensure consistency in the datasets. The training and testing sets were used to train each of the models (210 spectra for training 45 for testing). The withheld validation data (45 spectra) were then used to further explore the models' accuracy on previously unseen data that was similar to the data the models were trained on. After training of the models, the results were then reverse transformed back to molar quantities by mathematically undoing the scaling that occurred prior to training allowing for the direct interpretation of the results.

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

Multivariate Linear Regression (MLR)

MLR is the least computationally complex method that we utilized. All features are fit with a hyperplane in which the dimensionality is determined by the number of features, each feature has an associated weight. This hyperplane is then used to identify concentrations of new samples in the same way that a line would be used for regression with only one feature. Multiple linear models including Lasso, ElasticNet, and Orthogonal Matching Pursuit were tested, but the highest performing estimator was the Ridge regressor.[61] This method tends to perform well when you have a large number of features compared to your number of spectral samples.

K-Nearest Neighbors (KNN)

KNN is a method of supervised learning that uses the proximity of previously explored data to make predictions by looking at the Euclidian distance (calculating distance using the Pythagorean theorem) between the neighbors and the test datapoint and using that to adjust the predictions.[62] Numbers of utilized neighbors between 2 and 10 neighbors were tested, but the model performed the highest when 5 were used.

Decision Trees (DT)

DT work to separate the large dataset into smaller pieces repeatedly based on optimized features to be used as split points.[63] These smallest components, or leaves, then are used to identify predictions for new data. The model utilized in this work terminated splitting once 2 features were unable to be split further. The model then worked to minimize squared error between training predictions and true values. The original splits were randomized.

Gradient Boosted Regression (GBR)

GBR is an example of an ensemble algorithm that allows for the utilization of many smaller models, in this context decision trees.[64] This method is more computationally complex than a single DT and can identify more complex patterns. The model presented here utilizes a Huber loss function, and 2,000 estimators with a learning rate of 0.5 and a max depth of 1.

Multilayer Perceptron (MLP)

MLP is an example of an artificial neural network, a framework of interconnected nodes referred to as neurons.[65] Each neuron has an associated weight, which is adjusted with each training step through a mathematical process of backpropagation. The model presented in this work uses a tanh activation function, an Adam solver, and 500 training steps.

Support Vector Regression (SVR)

SVR utilizes the power of high dimensionality data to identify patterns.[66] By pulling the data into a higher dimensionality space, it allows for the fitting of the model with different mathematical approaches. The kernel describes the function used to fit the data that has been pulled into the high dimensionality hyperplane. This model utilizes a radius bias function as the kernel for fitting the dataset.

Model Analysis

To evaluate the models after training, the R^2 values are calculated to quantify the closeness of fit of the model to the data. Error was also calculated at three different places within the training set. First, the error within the training set is evaluated by comparing the predicted values to the true values with each model. This also describes how well the model was able to fit the training data. Next, the error of the testing data was calculated in the same way. This highlights the final accuracy of the training as the training and datasets are both used to adjust model parameters. Finally, the error within the validation data focuses on the ability of the model to evaluate data that it has not previously been exposed to. We also evaluate the prediction of the models on the ocean proxy samples to determine how well they perform on data that is chemically different than the data that the models were trained on.

Results and Discussion

Evaluating Feasibility of Using IR Spectra to Quantify Saccharide Concentration

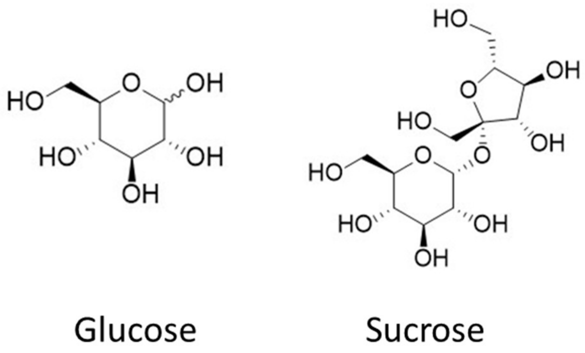

Figure 2. Molecular structures of both glucose (left) and sucrose (right).

The chemical complexity of simplified proxy and ocean proxy samples is explored with ATR-FTIR spectroscopy and quantitative machine learning approaches to develop a simple and accurate method of analysis. The FTIR spectra provide chemical information about the sample components and their concentrations, which have a linear correlation with absorbance. The correlation diverges from a linear relationship at high absorbance values, which is not of concern in the presently studied concentration ranges. A single figure containing all the acquired spectra is presented in the SI (Figure S1). Glucose has many vibrational modes that can be used for analysis (Figure S2).

Heat maps can be used to visualize the spectral dataset in its entirety. The data was sorted with respect to the concentration of glucose and then plotted against the wavenumber and the intensity at that wavenumber for a given spectrum. This allows for the visualization of the entire dataset in the context of changing glucose concentration and is presented as a heat map in Figure 3. A band of increasing intensity can be seen between 1200 and 1000 cm^{-1} correlating to the increasing concentration of glucose in solution, specifically with the C-C and C-O vibrational modes. The presence of this band supports the ability of the machine learning models to have representative features that will allow for the concentration analysis of glucose.

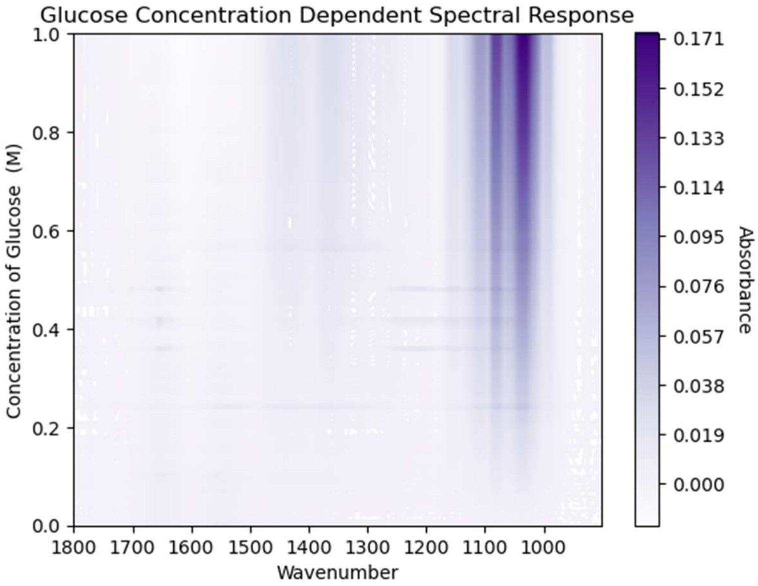

Figure 3. Heat map of the ATR-FTIR dataset as sorted by the concentration of glucose (0 – 1 M). The band of intensity growing in between 1100 and 1000 cm^{-1} corresponds to the increasing C-O stretching within the IR fingerprint region from the increased concentration of glucose. We do not see a strong spectral signature for the ESA relative to that of glucose also in solution (0 – 5 mg/mL) where we would expect the amide bands to exist between 1700 and 1500 cm^{-1}.

It is also possible to describe the ability of the ATR-FTIR dataset to be fit through pre-training feature selection analysis. Typically, this approach is utilized to reduce data dimensionality and avoid overfitting large datasets or to reduce computational costs. For our uses, we can analyze the feature selection scores (Figure 4) to determine how much of the variance with respect to concentration is described in each wavenumber. Ideally, these scores would correlate a higher score with the areas of the spectrum where we see vibrational contributions from glucose. To do this analysis, we used the SelectKBest function within scikit learn. This function explores each feature (wavenumber) independently and calculates a score quantifying the degree to which that feature describes the target variable (concentration). The function is provided to all of the testing dataset so it is exposed not only to peaks from the analyte of interest but also any species in the matrix. For solutions, it is ideal to see peaks associated with the analyte of interest having high scores and any peaks associated with the matrix to be minimized. In Figure 4 the C-O and C-H bands from 1150 to 500 cm^{-1} appear to describe the majority of the variation within the data. This is likely due to the fact that they have a much higher relative intensity compared to the bending modes around 1400 cm^{-1}. The feature selection analysis also contains additional features between 1500 and 1400 cm^{-1}, which arise as artifacts within the glucose data from mathematically removing water from the samples via -log(R/R_0).

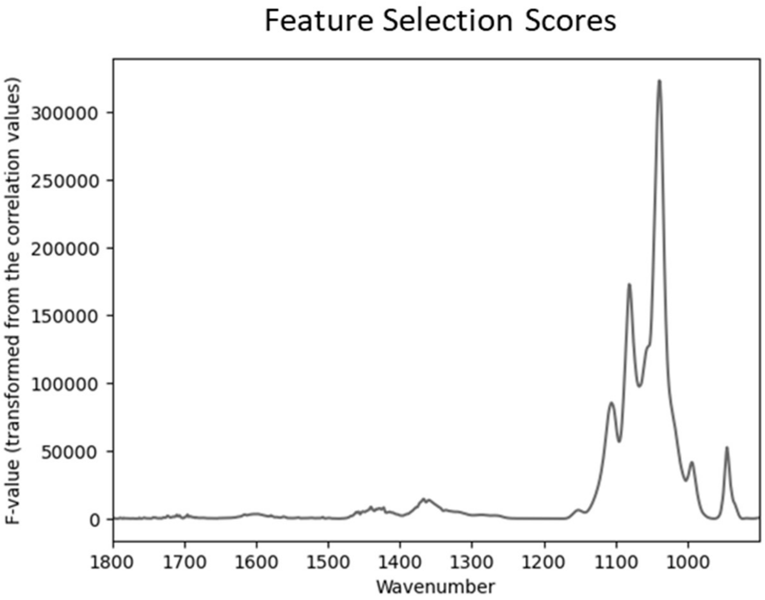

Figure 4. Feature selection scores as calculated by the SelectKBest function from Scikit Learn.

Evaluating Machine Learning Models' Fit of the Simplified Proxy Dataset

After training, the accuracy of the model's ability to identify the concentrations of the test and validation sets was evaluated to explore the influence of the true concentration of the simplified proxy dataset on the error. Ideally, there wouldn't be any effect and the error would be consistent regardless of concentration range. Figure 5 visualizes these results. MLR had the smallest errors and apart from the GBR none of the models had a strong error response with respect to concentration range. Also, there is little variance between the accuracies of the training and validation datasets suggesting that there is good fit amongst the models. R^2 values for each model have also been calculated and are presented in the SI (Table S2).

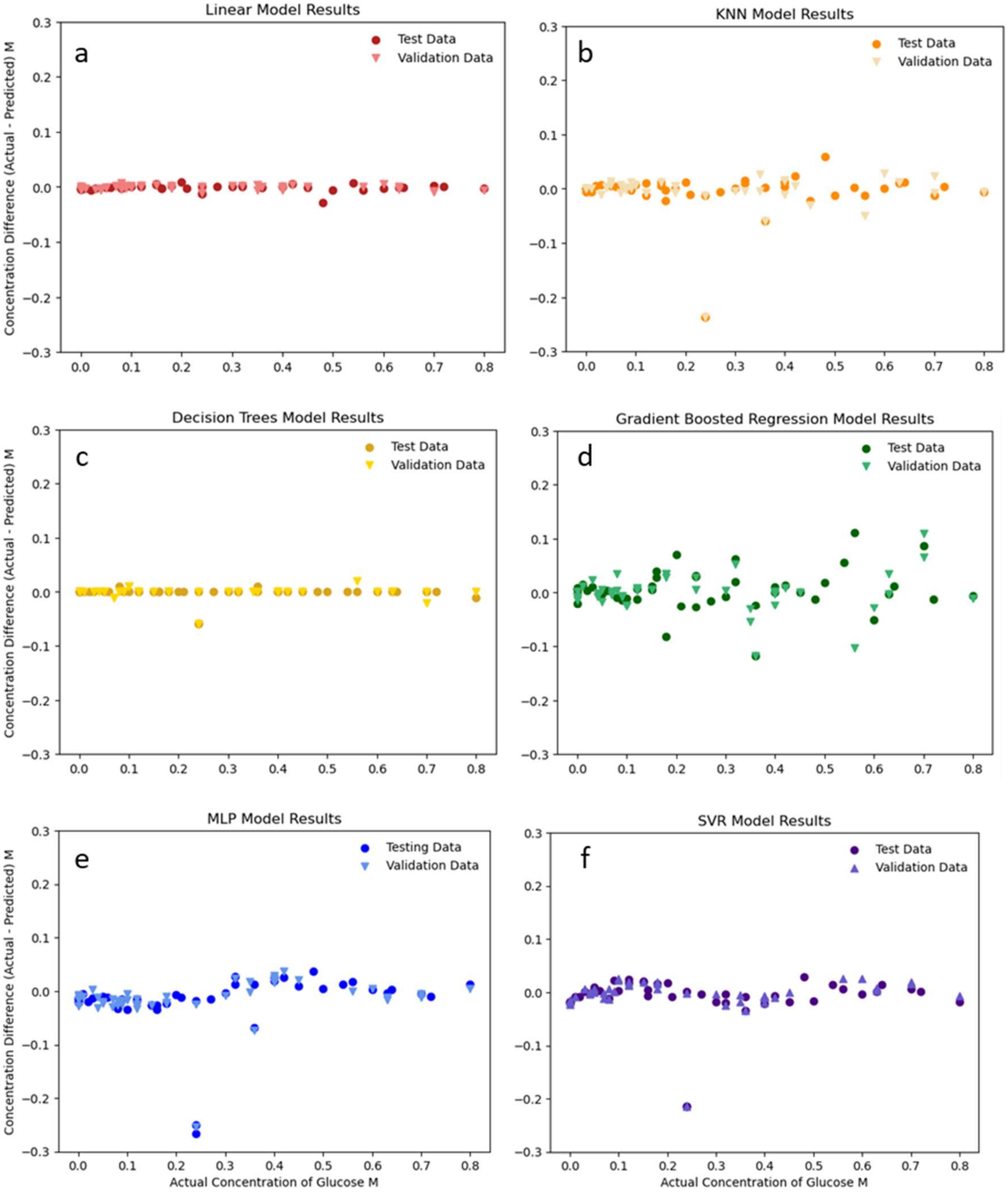

Figure 5 (a-f). Scatter plots depicting the accuracy of each of the utilized machine learning models on the **simplified proxy** dataset. The y-axis represents the difference between the model assigned and the actual concentrations of the testing dataset (darker circles) and the withheld validation dataset (lighter triangles). Apart from the gradient boosted regression model, the models are robust with respect to concentration meaning that the error changes minimally as the actual concentration increases.

To perform a more in-depth error analysis, the models' error was calculated between each step of the training by evaluating the training, testing, and validation sets' final accuracies. All of the models had smaller than 60 mM in error amongst the different steps. These results have been visualized in Figure 6.

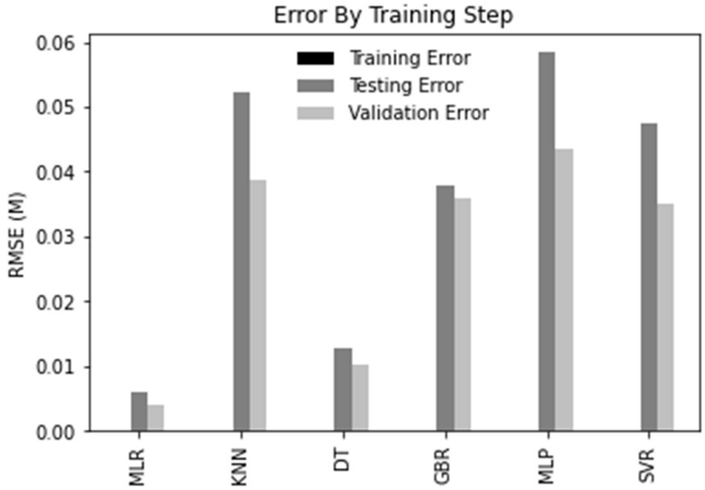

Figure 6. Bar graphs depicting the associated root mean squared error (RMSE) in each part of the training process for the **simplified proxy** dataset. Note, the training error for each model was 0.000 M, suggesting that each of the models fully converged on the training dataset. All models have a final validation error of less than 0.06 M, but the MLR performed the best in this evaluation.

Evaluating Machine Learning Models' Fit of the Ocean Proxy Dataset

The ocean proxy samples were then estimated using these same ML models. The "true" saccharide concentrations are defined as the sum of the concentrations of glucose and sucrose. This additive concentration, coupled with the increased complexity of the matrix extends these proxies beyond the chemical space that the models were originally trained on. For the purpose of identifying a generalized saccharide concentration, it is important to select for the models with the highest combined accuracy without valuing one sample over the others. A model performing poorly here doesn't suggest that the model is poorly trained, just that it doesn't have the capacity to generalize that far beyond the training. For example, MLR had the lowest error in test and validation datasets as seen in Figure 6 for the simplified proxies, however MLR only has an accuracy of 50-60% on the ocean proxies. This suggests that the MLR model is highly fit to glucose to the point that it does not generalize to sucrose which for other chemical contexts would be ideal.

The highest accuracy in identifying the combined saccharide concentrations came from the SVR model. It had a slight bias towards high concentrations but was fairly accurate, with over 90% accuracy in all 3 samples. The MLP also performed well, having an accuracy of 99.9% on the 0.1001 M saccharide but just below 90% accuracy on the other two proxies. These results are shown in Figure 7.

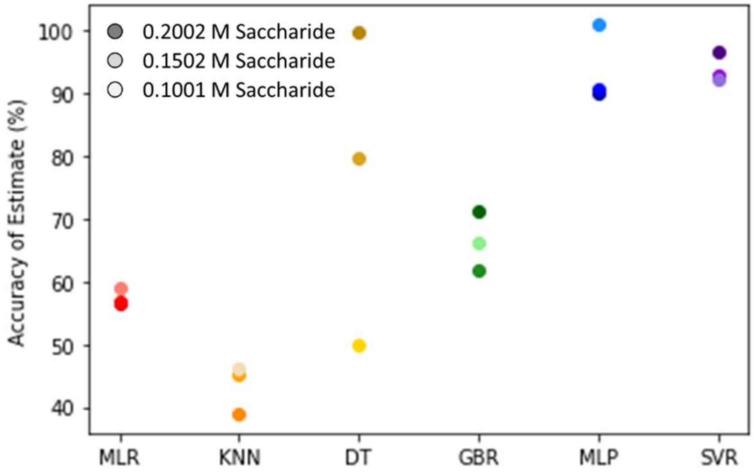

Figure 7. Accuracy of estimates from each of the models on interpreting the **ocean proxy** saccharide concentrations. The darkest markers in each column represent the highest concentration of saccharide in ocean proxy (0.2002 M) and the lightest represent the least concentrated (0.1001 M). The models have varied levels of success at identifying samples that are far removed from the original training set. The highest performers were MLP and SVR with SVR having a higher total accuracy for all three proxies.

Summary of Discussion

Our quantitative results indicate that a computationally inexpensive model, SVR, provides predictions of saccharide 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 more chemically complex test samples with more than just glucose and ESA. Even with increased sample complexity, including additional saccharide, protein, and lipid molecules, the SVR model accurately predicts the total saccharide concentration.

Conclusions

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

Supplemental Information

Appendix A. ATR-FTIR Spectra of all Training Samples.
Appendix B. Vibrational Analysis of Glucose and ESA
Appendix C. Highest Concentration of ESA and Glucose
Appendix D. Selected Single Peak Beer's Law Analysis
Appendix E. Analysis of MLR and SVR Weights
Appendix F. Tabulated Values for Accuracy and Fit for Each ML Model
Appendix G. Concentration Predictions for Ocean Proxy Solutions for Each ML Model

Acknowledgements

N.M.N. acknowledges funding support from NASA's Future Investigators of NASA Earth and Space Science Technology (FINESST) grant number 20-PLANET20-0067. 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. 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**, 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), 216. 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) Al Ibrahim, E.; Farooq, A. Transfer Learning Approach to Multitarget Temperature-Dependent Reaction Rate Prediction. *J. Phys. Chem. A* **2022**, *126* (28), 4617–4629. https://doi.org/10.1021/acs.jpca.2c00713.

(40) Freitas, R. S. M.; Lima, Á. P. F.; Chen, C.; Rochinha, F. A.; Mira, D.; Jiang, X. Towards Predicting Liquid Fuel Physicochemical Properties Using Molecular Dynamics Guided Machine Learning Models. *Fuel* **2022**, *329*, 125415. https://doi.org/10.1016/j.fuel.2022.125415.

(41) 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.

(42) 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.

(43) 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.

(44) 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.

(45) 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.

(46) 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.

(47) 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.

(48) 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.

(49) *sklearn.linear_model.LinearRegression*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.LinearRegression.html (accessed 2023-11-02).

(50) 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.

(51) 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.

(52) 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.

(53) 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.

(54) 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.

(55) 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.

(56) 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.

(57) Langmuir, I.; Waugh, D. F. The Adsorption of Proteins at Oil-Water Interfaces and Artificial Protein-Lipoid Membranesthe 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.

(58) 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.jpclett.2c00231.

(59) 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.

(60) 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.

(61) *sklearn.linear_model.Ridge*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.Ridge.html (accessed 2023-12-21).

(62) *sklearn.neighbors.KNeighborsRegressor*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html (accessed 2023-12-21).

(63) *sklearn.tree.DecisionTreeRegressor*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html (accessed 2023-12-21).

(64) *sklearn.ensemble.GradientBoostingRegressor*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (accessed 2023-12-21).

(65) *sklearn.neural_network.MLPRegressor*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.neural_network.MLPRegressor.html (accessed 2023-12-21).

(66) *sklearn.svm.SVR*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.svm.SVR.html (accessed 2023-12-21).

For TOC use only