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

Nicole M. North[†], Jessica B. Clark[†], Abigail A. A. Enders[†], Alex J. Grooms[†], Salmika G. Wairegi[†], Kezia A. Duah[†], Efthimia I. Palassis-Naziri[†], Abraham Badu-Tawiah[†], 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

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

Keywords: Carbohydrate, saccharide, sugar, lipid, fatty acid, ocean, supervised learning, Raman

Introduction

Marine organic composition drives many of the methods in which the ocean interacts with the other chemical systems of Earth. The organic compounds have the ability to influence atmospheric chemistry when partitioning to the surface of the ocean and contributing to sea spray aerosols.[1–9] They can also act as feedstocks or markers of metabolism within biological systems like algal blooms.[10–12] Measuring marine organic compounds also improves the ability to detect and remediate potential marine disasters such as oil spills.[13,14] The largest challenges in attaining a large scale understanding of ocean chemistry arise from the incredibly diverse array of compounds present in marine samples.[7,8,15–17]

Vibrational spectroscopy is used extensively to describe marine chemistry and aqueous environments.[6,18–20] Raman spectroscopy is used, in particular, in deep ocean probes due to the durable instrumentation and ability to analyze aqueous environments without major disruption from the vibrational signature of water itself.[21,22] Raman spectroscopy is also used as a method to identify chemical markers to understand physical properties (e.g. chemical kinetics, thermochemistry, and chemical building blocks) and biologic activity in marine systems.[18,23] In our prior work, we used Raman spectroscopy to identify ion pairing in aqueous solutions of NaCl and KCl.[24] Ion pairing is detected by observing how the vibration of water is affected by being in different solvation shells of ions. Detecting ion pairing requires a high signal-to-noise ratio as these interactions may only make small perturbations in the OH symmetrical and asymmetrical stretching regions of the spectra.

Utilizing machine learning in tandem with vibrational spectroscopy has been of high interest in recent years,[25,26] particularly in the areas of real-time reaction monitoring[27] and medical diagnostics.[28–31] The vibrational fingerprints of different analytes of interest are proving to be powerful features for machine learning models. De Medeiros Back and colleagues published a paper in 2022 describing methodology utilizing vibrational spectroscopy to identify microplastics in the Mediterranean Sea. They found that support vector classification showed the best performance out of the machine learning methods that they evaluated. Our group has also successfully utilized machine learning and vibrational spectroscopy to identify organic concentrations in complex chemical matrices.[32] In the prior work, attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR) data was used to evaluate the ability of six different machine learning methods to identify concentrations of glucose in a complex matrix of differing concentrations of egg serum albumin. Ultimately, we found that support vector regression had the highest accuracy in identifying glucose. To further analyze the extent of the expandability of the training, more chemically complex samples (containing sucrose, glucose, egg serum albumin, bovine serum albumin and 1-butanol) were created. It was found that the model could not only identify the concentration of glucose alone, but a sum concentration of saccharide (glucose and sucrose).

Here, we expand upon our results in our previous work by evaluating three different analytes' concentrations simultaneously, rather than just one. Our three analytes of interest have been curated due to their relevance and impact on marine chemistry and the marine ecosystem. We evaluate a total of eight machine learning models to this end. Each of these models was trained using two different datasets. The first dataset, the spiked lab (**SL**) sample dataset, is created with ultrapure water and spikes of various concentrations of the three analytes. This dataset focuses on giving the models access to highly resolved calibration spectra with little matrix effect. The second dataset, the spiked marine (**SM**) sample dataset, utilizes the same spikes as in the previous dataset but instead of using ultrapure water, unspiked marine (**UM**) samples are used. This dataset provides real-world samples and works to highlight the effect of the matrix (salts, other organics, potential fluorophores) as well as secondary chemical effects of the complex chemical environment. This work presents, to the best of our knowledge, the first multi-output machine learning models to describe the organic components of ocean chemistry quantitatively.

Methods

Selection of Representative Analytes

Describing the vast chemical complexity of ocean samples in just a few analytes of interest is incredibly challenging. This current work aims to focus on a saccharide, a fatty acid, and a proxy for proteins. Due to time and technique constraints, only one to two representatives could be chosen for each class of molecule. The compounds selected need to be ones that would be expected to be found in marine samples. As for concentration range, the total concentration sum was < 300 mM,[46] arising from estimated total organic carbon (TOC) for marine samples. This average varies globally depending on marine system, time of year, and local ocean productivity.[47–50] This adds the constraint that the analytes of interest should be soluble in room temperature water at a concentration of close to 300 mM.

Marine proteins vary greatly with type and size. These variables make defining the concept of a total concentration challenging. To standardize and simplify this analysis, this study looks at amino acids rather than a specific protein. Glycine and histidine were chosen as analytes of interest. These amino acids have been defined as potential markers for gluconeogenesis (non-sugar metabolism) and antifungal properties among others.[51–54] Glycine has also been noted as partitioning into sea spray aerosols and being transported into cloud water.[55] Amino acids have been reported to make up 11% by mass of the dissolved organic carbon within submicron sea spray aerosol particles.[56] Note although histidine was chosen as a representative analyte it was not found to be in UM samples above the LOQ of the utilized mass spectral calibration and thus could not be analyzed through our Raman and ML combined approach.

For fatty acids, the analyte needed to be marine relevant and not have a strong partitioning to the aqueous surface. This second criterion limited the options to fatty acids with a carbon chain length of three or less. Butyric acid has a carbon chain length of three and has been noted as one of the most abundant short chain fatty acid in algal bloom metabolic processes.[10,11,57,58] Butyric acid can also be an indicator of ocean oxygenation.[59] As algal bloom populations collapse, the dissolved oxygen is depleted, causing negative impacts to ocean health.[60,61] This lack of oxygen also increase ocean acidification.[62]

The chosen analyte representative for saccharides is glucose. This saccharide is one of the most abundant of the saccharides in marine systems.[63] It is also a common feedstock for small scale marine life like algae and has been used in the past as a biomarker of algal bloom presence and stage.[64,65] Glucose, along with other saccharides have also been known to partition into aerosols [9,63] where they can act as potential ice nucleators.[66]

Solution Preparation

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

Raman Spectroscopy

A total of 210 Raman spectra (100 SL, 100 SM, 10 UM) were collected using a custom-built Raman spectrometer. This instrument contains a diode-pumped 532 nm CW laser containing built-in laser line (±0.5 nm) and polarization filters (>100:1) (CrystaLaser). The excitation source is directly coupled to a custom-built fiber-optic polarized Raman probe system (InPhotonics) allowing 235 mW power at the sample with a spectral range of 90–4200 cm^{-1}. The output, both polarized and depolarized scattered light, is collected by two independent fiber-optic terminated ports. The two polarization output ports are fiber coupled directly to a spectrograph through a 50 μm slit with a 1200 g mm^{-1} grating with a 750 nm blaze, which is calibrated to Ar/Ne emission lines (IsoPlane 320, Princeton Instruments), and is detected with a liquid-nitrogen cooled CCD detector (Pylon, 1340 × 400 pixels, Princeton Instruments). Each 200 μm core fiber is directly coupled to the spectrograph and allows for the simultaneous collection of the perpendicular (HV, depolarized) and parallel (VV, polarized) spectra. Measurements of all of the concentrations were performed at a room temperature of 21 ± 2 °C. Spectra were collected by signal averaging 50 frames each with a 0.4 s integration time. Only the parallel (VV, polarized) spectra were used for analysis.

Paper Spray Ionization Mass Spectrometry (PSI-MS):

Mass spectral data was used to benchmark the model results. The mass spectral data was not used to train the models, but to develop a validation set of concentrations for the field samples so that an appropriate error analysis of the ML models could be performed. The mass spectrometry (**MS**) method used herein for all calibrations consisted of a paper spray ionization (**PSI**) platform which has been utilized as a valuable ambient ionization MS method for direct, targeted, and rapid analysis of analytes within a native sample.[33,34] In PSI-MS sample is deposited directly onto an untreated Whatman #1 filter paper triangle substrate produced in-house. All samples (i.e. SL and SM samples) were deposited on the paper substrate (10 μL sample size) and allowed to dry completely before the application of methanol extraction solvent. Ionization was facilitated by the application of a high DC voltage (6 kV) to the ionization apparatus, thus inducing an electrospray ionization mechanism from the paper substrate. Methanol extraction solvent was applied directly onto the paper substrate with the paper triangle secured from the rear via a copper clip. Paper substrates were held at a 5 mm distance from the inlet of the mass spectrometer which was held at 250 °C inlet capillary temperature. Spectra were recorded over a total acquisition time of two minutes with 0.25 minutes analyte and internal standard averaging for all calibrant and UM solutions. The MS was operated in positive-ion mode for butyric acid, glycine, and histidine analytes with analysis of protonated pseudomolecular ion and negative-ion mode for glucose for analysis of the chloride adduct pseudomolecular ion. Protonation of butyric acid was facilitated via the high DC voltage ionization mechanism.[35] Protonation of glycine and histidine was assisted via addition of 0.1 % formic acid. Glucose chloride adduct formation was assisted via addition of 10 mM ammonium chloride.[36]

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

Mass Spectral Quantification - Internal Standard Calibration Curve

Using the PSI-MS platform, we sought to quantify each analyte in the UM samples and constructed internal standard calibration curves (Figure S2). This was done using standard solutions of each analyte made in neat water (13-100 mM) with appropriate internal standards (800 mM). We placed 50 μL of the prepared internal standard solution into 2 mL of the standard solution to prepare a 16 mM solution for analysis. We then took 10 μL aliquot of the 16 mM solution and place this on the paper triangle, allowing for 1 minute of dry time before extraction solvent application onto the paper and applying a 6 kV high DC voltage for subsequent analysis in the positive ion mode (butyric, glycine, and histidine) and negative ion mode (glucose). Tandem MS (MS/MS) mode was implemented for analysis, using the appropriate transitions for each compound and its corresponding internal standard (Figure S2a (butyric), S2b (glucose), S2c (glycine), and S2d (histidine)). We monitored the ratio of the intensity of the analyte-to-internal standard (A/IS) as a function of the analyte concentration – consistent with MS based calibration. Figure S2a-d shows the linearity achieved with R^2 values that fall within the 0.99 range. With these results, we moved forward with the quantitative analysis of the selected compounds using the PSI-MS set-up with UM samples. Under analogous conditions to calibration, the UM samples were analyzed, and their spectrum confirmed the presence of butyric acid, glucose, glycine, and histidine in the ocean water samples via MS/MS.

Field Collection for SM and UM Samples

Water was collected from two locations in Cocoa Beach, Florida in January 2023. Sampling site one was the Atlantic Ocean and site two was the Banana River within the Indian River Lagoon System. The Banana River is a brackish waterway connected via ocean inlet with mangrove shorelines; the conditions provide a unique aqueous environment on the west side of the Florida barrier islands. Samples are categorized as surface microlayer (**SML**) and bulk sea/river water (**BW**). We operationally define the SML as the top 1 mm of the sampled water and BW as the top 1 m of the sampled water. All samples were stored at room temperature and shipped; once received, samples were stored at 2°C until analyzed.

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

BW was collected by first copiously rinsing a jar, replacing the lid, submerging the covered jar, and finally removing the lid underwater. Jars were filled to avoid head space. SML water was collected according to methods detailed by Harvey and Burzell.[37] Briefly, a clean hydrophilic glass plate (Millipore Sigma, unframed, H × W × D 200 mm × 260 mm × 4 mm) was submerged perpendicular to the surface to about the top inch, the plate was then withdrawn from the water at a rate of approximately 20 cm/s. Adsorbed water and organics were collected via silicone squeegee into a copiously rinsed glass jar.

An additional sample was collected from the tropical saltwater aquarium, ~ 200-gallon capacity, within the Center for Life Sciences Education at the Ohio State University.

Data Preprocessing

There was a large degree of observed Raman fluorescence in the SM and UM sample datasets. This presented itself as broad band elevated baselines **(Appendix A – Figure S1)**. Fluorescence was expected from the large number and variety of naturally occurring organic compounds in solution. Multiple methods of preprocessing were evaluated to see if this baseline variation could be corrected and if that correction led to higher model accuracies. To ensure that all data was treated the same way, all preprocessing was completed on the SM, UM, and the SL datasets even though the fluorescence was not observed in the SL data **(Appendix A – Figure S1)**. The highest accuracies came from taking the average of the Raman spectra from 1283 to 2640 cm^{-1}. This average was then subtracted from all intensities from 346 to 3117 cm^{-1}. This baseline corrects some of the observed Raman fluorescence in the SM and UM dataset. Next, the entire spectrum is normalized with respect to 3343 cm^{-1} which is correlated with the isosbestic point between the symmetric and asymmetric O-H stretching bands. This further corrects for the fluorescence.

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

Python Scripts

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

Regression Methods

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

Decision Trees (DT)[38]

Decision trees (DT) utilize iterative binary splits of the data to identify concentrations of new data. A fitting criterion of absolute error was used with a best splitter to separate the data into leaves that had a minimum of 5 samples. Fewer than 5 samples per leaf did not lead to increased model accuracy.

Random Forest (RF)[39]

Random forests (RF) utilize many decision trees to improve model accuracy. In this context, 100 trees were trained independently of each other (non-bootstrapped) by minimizing squared error. All of the trees were then used simultaneously to make model assignments. As few as 10 and as many as 100 trees were evaluated in steps of 10 and the most successful model is presented here.

Gradient Boosted Regression (GBR)[40] and Histogram Gradient Boosted Regression (HGBR)[41]

Gradient boosted regression (GBR) and histogram gradient boosted (HGBR) models are made similarly to random forest models in the fact that the base architecture is a decision tree. However, the difference is that as new trees are trained in GBR, models learn from the previous trees. For this context, 100 trees are used reducing a loss of squared error. A learning rate of 0.5 was used with a max depth of 1. HGBR utilizes a histogram estimator to improve the speed of computation. The learning rate and max depth are both the model defaults and adjustments to these values didn't lead to increased model accuracy.

K Nearest Neighbors (KNN)[42]

K nearest neighbors (KNN) models utilize the distance from previous datapoints to estimate quantifications for new samples. The presented models utilize the 5 nearest neighbors to make their assignments. Neighbor numbers between 1 and 10 were evaluated and 5 neighbors performed the best.

Support Vector Regression (SVR)[43]

Support vector regression (SVR) models work to optimize high dimensionality hyperplanes to fit datasets with many features. The kernel being utilized in the presented models is a radial bias function. All of the available kernels were tested, and the radial bias function performed the best.

Multi-Layer Perceptron (MLP)[44]

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

Multi-Variate Linear Regression (MLR)[45]

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

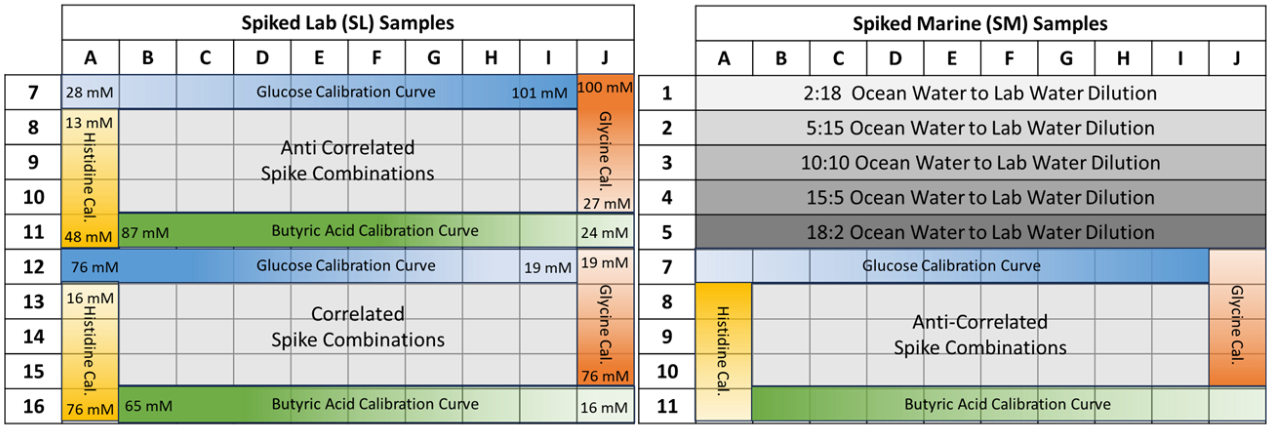
Results and Discussion

The two datasets, Spiked Lab (SL) and Spiked Marine (SM), consisting of only Raman spectroscopic data, were each made with a different perspective of the chemical system in mind. The SL dataset works on making a clean calibration curve of each chemical analyte as well as showing the direct interaction between the analytes of interest. The SM dataset includes the contribution of the salts and other organics present in the marine samples which can greatly affect the vibrational signatures observed. All data was labeled with a "true" concentration using the mass spectral data. Models trained on each of these datasets were used to predict the concentrations of the Unspiked Marine (UM) samples to evaluate their ability to be applied to new marine samples.

Sample Organization

After selection of the analytes of interest, a methodology was developed to make unique combinations of organic concentrations to generate the datasets for training. Four distinct calibration curves, two with 10 datapoints and two with 5 datapoints, were utilized in the method. The calibrations and the sample combinations that are developed make up a single array and each dataset contains two sample arrays. Each sample array contains 50 samples. This is done in different ways for the SL samples and the SM samples (**Figure 1**).

For the SL samples, the first sample array has anti-correlated calibration curves (**Figure 1** SL samples rows 7-11). This means that the analyte concentration gradients on opposite sides of the sample array are changing inversely to one another. This ensures that the models are penalized for trying to correlate any of the concentrations during the training. For the second sample array, the opposing analyte concentrations change proportionally to one another (**Figure 1** SL samples rows 12 – 16). This second array was to ensure that there wasn't in inverse correlation that could be picked up by the model either.



	Spiked Lab (SL) Samples									
	A	B	C	D	E	F	G	H	I	J
7	28 mM			Glucose Calibration Curve					101 mM	100 mM
8	13 mM Histidine Cal.			Anti Correlated Spike Combinations						Glycine Cal.
9										
10										27 mM
11	48 mM	87 mM			Butyric Acid Calibration Curve					24 mM
12	76 mM			Glucose Calibration Curve					19 mM	19 mM
13	16 mM Histidine Cal.			Correlated Spike Combinations						Glycine Cal.
14										
15										76 mM
16	76 mM	65 mM			Butyric Acid Calibration Curve					16 mM

	Spiked Marine (SM) Samples									
	A	B	C	D	E	F	G	H	I	J
1	2:18 Ocean Water to Lab Water Dilution									
2	5:15 Ocean Water to Lab Water Dilution									
3	10:10 Ocean Water to Lab Water Dilution									
4	15:5 Ocean Water to Lab Water Dilution									
5	18:2 Ocean Water to Lab Water Dilution									
7				Glucose Calibration Curve						
8	Histidine Cal.			Anti-Correlated Spike Combinations						Glycine Cal.
9										
10										
11					Butyric Acid Calibration Curve					

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

For the SM samples, the setup involved associating each column of the sample arrays with a marine sample (**Table 1**). This allowed for a dilution series to be made for the first sample array (**Figure 1** SM samples rows 1-5). Due to the UM samples already having unique concentrations this dilution series took the place of the anti-corelated sample array in the SL dataset. The second sample array contained the same organic spikes that the correlated calibration data of the second SL sample array (**Figure 1** SM samples rows 7-11). Together, these ensured that the concentrations of the UM samples would be within the calibration. A full spreadsheet describing the concentrations of each analyte in each sample correlated with the same alphanumerical matrix described in Figure 1 is available on the GitHub associated with the project (https://github.com/Ohio-State-Allen-Lab/multi_compound_marine_regression).

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

Table 1. *Marine samples associated with the UM and SM datasets. Concentrations of glycine, butyric acid, and glucose were calculated through mass spectrometry and will be used as the "true" values of concentration for these samples. Histidine concentrations were all beneath the LOQ for the mass spectral method.*

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

The concentration combinations within the spike-containing sample arrays are created by taking the row or column associated with each of the calibration curves and spiking those concentrations into either lab or marine water depending on the dataset.

After training all the models, initial assessments on internal accuracy were made. **Figure 2** depicts all the error associated with each chemical species (glycine, butyric acid, and glucose) for each of the machine learning methods trained on SL data (left) and SM data (right). The errors associated with the SM models are, on average, higher than the models trained on the SL data. Within each set, the ensemble methods (GBR, RF, and HGBR) tend to perform better than the single models. There doesn't tend to be an immediately visible trend between error and chemical species, suggesting that different models are able to optimize different chemical species more effectively.

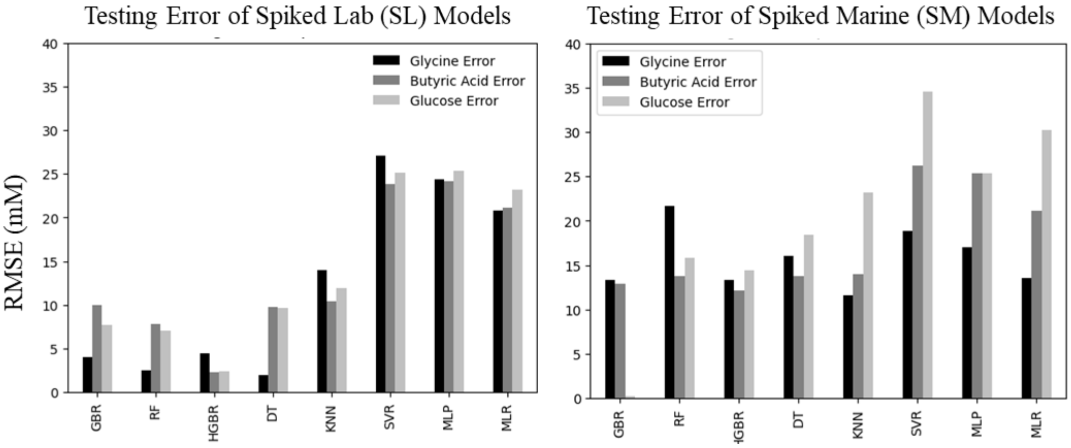


Figure 2. *Test stage root mean squared error (RMSE) values for each combination of ML approach and chemical species.*

All the models can then be used to predict the concentrations of the UM samples. **Figure 3** has the model assignments for each of the different compounds. The models trained with the LS data are on the left (circles) and the models trained with the SM data are on the right (triangles). The dotted lines show a boundary of +/- 20% of the highest concentration of that analyte in a single marine sample. The models trained on the SM models show much more clustering of assignments within this +/-20% region. It is also possible to identify models that are performing more poorly across the board these include MLP, MLR, and RF.

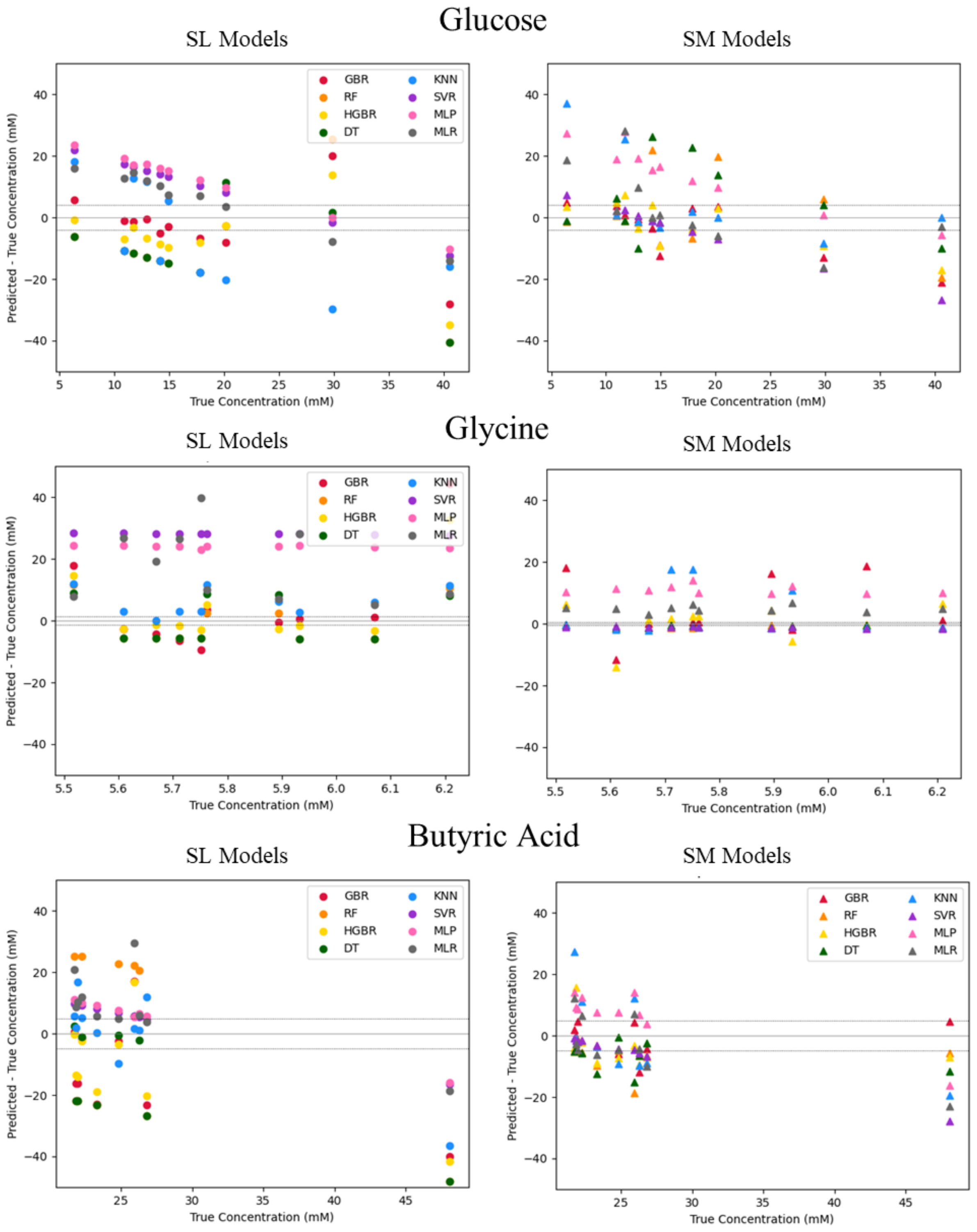


Figure 3*. UM sample estimates from each ML approach on SM models (left - circles) and on SL models (right - triangles). Solid grey line denotes a difference between actual and predicted concentrations of 0. The dotted lines represent +/- 20% of the most concentrated marine sample for the given chemical species (glucose, glycine, and butyric acid). The SM models show more clustering within these boundaries than the SL models suggesting that the SM models were more accurate at identifying the concentrations within the UM samples.*

To improve the visualization of the models that are making assignments in the +/-20% range, the number of assignments in this region were counted for each model and for each compound (**Figure 4**). This confirms that the SM models perform better than the LS models at identifying the UM samples. This increase in accuracy likely comes as a function of the increased similarities between the training data and the final validation data. These similarities include non-analyte organics which are leading to the observed Raman fluorescence. These organic compounds likely change from sample to sample, but they work to make the training data more chemically similar to the final validation data.

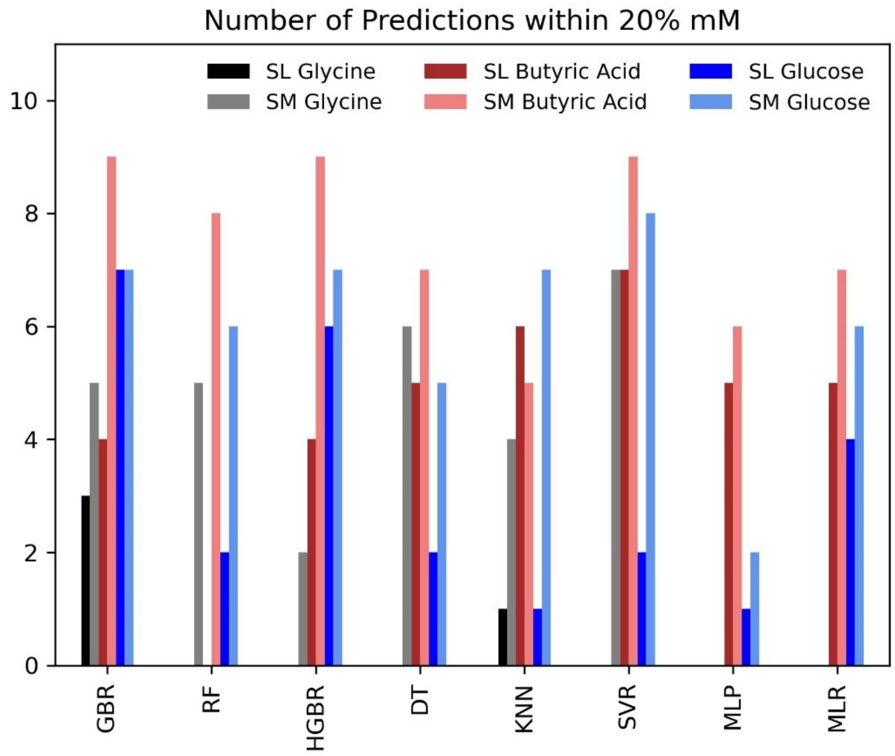


***Figure 4**. Counted values out of 10 for the correctly quantified UM samples within 20% of the max true values in a single UM sample. These counts are separated by ML approach and chemical species. Importantly, the SM models perform higher than the SL models in nearly every case. SVR achieved the highest accuracies for all three analyte concentrations.*

SVR performed the best at identifying the concentrations of glycine, butyric acid and glucose assigning 7/10, 9/10, and 8/10 within 20% of the true value, respectively. Butyric acid was also well described through the GBR and HGBR methods (**Table 2**).

Table 2. *Highest performing models for each analyte compound.*

Highest Accuracy Model for Each Analyte Compound		
Glycine	Butyric Acid	Glucose
Support Vector Regression (SVR) **7/10 Marine Samples**	Support Vector Regression (SVR) Or Gradient Boosted Regression (GBR) Or Histogram Gradient Boosted Regression (HGBR) **9/10 Marine Samples**	Support Vector Regression (SVR) **8/10 Marine Samples**

As mentioned in **Figure 1**, the SM sample models (highest performing) are trained on dilutions of the marine samples. To further analyze the accuracy of these models, it is important to measure the marine sample accuracy if the model hadn't been trained on dilutions of that exact sample. To accomplish this, the highest performing models (**Table 2** –SVR (for glycine, butyric acid, and glucose), HGBR (for butyric acid), and GBR (for butyric acid)) were trained another 10 times each. For each model, training one column of marine samples was dropped, (e.g., SM samples: column A) then the model was evaluated using the UM sample associated with that marine sample (for column A: sample A6). This allows for the analysis of the model if it was shown a truly new marine sample. This was then repeated with each of the remaining columns independently. **Figure 5** shows these results.

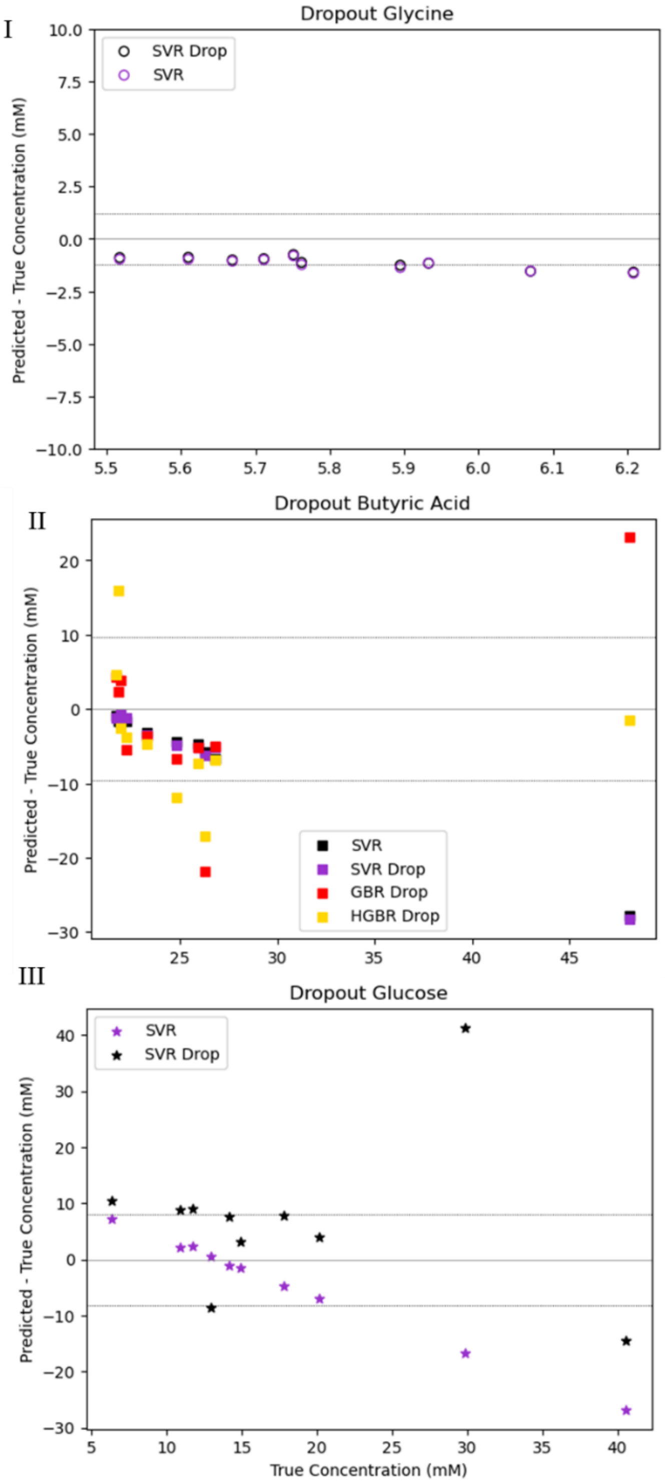


***Figure 5**. Marine sample analysis using dropout sample method. For each model training one column of samples was dropped (e.g., SM samples column A) then the model was evaluated using the UM sample associated with that marine sample (for column A: sample A6). The dropped sample results are in black or grey and the original analysis is left in the color associated with that ML approach in **Figure 4**. The accuracy of models is well maintained for glycine (I) and butyric acid (II). The largest loss in accuracy was in the measurement of glucose. This variance, due to it mostly being overestimates, may be associated with the presence of other saccharides in these samples that cannot be determined using the stated mass spectral method.*

The accuracy of analysis with and without sample dropout is maintained well in analyzing glycine and butyric acid. SVR performed the best out of the three possible butyric acid models in the drop out test. The model was able to achieve accuracy for 9/10 samples even with the sample dropout. GBR (8/10 correct) and HGBR (7/10 correct) both experienced reductions in accuracy in identifying concentrations of butyric acid while using sample dropout. The largest variance was found in the analysis of glucose where there is a trend in over estimation from the SVR model (**Table 3**). This perpetual overestimation may suggest that there are other saccharides in these field samples.[32] Our "true" value for glucose is limited to only glucose based on the limitations of our mass spectral method, which can only evaluate one stated analyte at a time. Machine learning models can accurately identify a generalized saccharide concentration through a sum of glucose and sucrose;[32] this is consistent with the vast majority of errors being positive, as observed here **(Figure 5**).

Table 3. *Effects of dropout sample test on highest performing models for each analyte compound.*

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

Future work should add samples to the SM dataset to help improve its stability, to lessen the reliance on any given marine sample and to further increase dataset size as the limited size of 100 spectra may be contributing to over fitting. Other experimental methods to benchmark and confirm the concentrations of the marine samples with additional representative analytes should be developed to improve the scope of the "true" concentrations. Other complementary analyte models should also be added to improve the overall organic compositional analysis. With sufficient analyte models it may also be possible to look for correlations between analyte models which would suggest which compounds may lead to systematic errors when coexisting in solution.

Future work should also include evaluating these models in their ability to identify changes in concentrations in entirely field based systems. A limitation of this current work is that all concentrations have been artificially spiked on top of true field samples. This limits the current ability to make conclusions on the field systems themselves. With a larger sample set from a field campaign through collecting spatially and/or temporally spaced samples it is possible that models, like those described in this work, will allow for the analysis of marine systems in many capacities including potential origin of life or ecological studies among many others.

Conclusion

Eight machine learning models were tested for their ability to identify four different analyte concentrations in a complex marine matrix. Two different Raman spectral datasets of organic spiked arrays were made on ultrapure water and on marine samples to approach the complex system in different ways. The results show that support vector regression had the highest accuracy in identifying all three analytes. Butyric acid was also well described through gradient boosted regression and histogram gradient boosted regression however these approaches performed more poorly than the support vector regression during the sample drop out test. In nearly every case the spiked marine (SM) dataset, in which the spikes were added to marine samples with their internal chemical complexity, outperformed the spiked lab (SL) dataset. Upon testing sample dropout to remove potential internal correlation in concentrations from the dilution series making up half of the SM dataset, it was found that butyric acid and glycine were largely unaffected. When this dropout approach was used with glucose, it led to an increase in overestimating glucose concentrations which suggests that there are saccharides in solution that are contributing to the same vibrational modes. This work reveals that it is possible to achieve accurate estimates of selected organics in an increasingly complex chemical matrix using Raman spectroscopy with machine learning. This combination of Raman and ML stands to improve our rapid response and characterization of marine samples both in the lab and in the field due to the durability and transportability of Raman instrumentation and the ease of use and rapid computations of a pretrained machine learning model.

Supplemental Information

Appendix A. Raman Spectra Before and After Preprocessing
Appendix B. Calibration Curves for Mass Spectral Analysis
Appendix C. Spiked Lab (SL) Models Test Accuracy
Appendix D. Spiked Lab (SL) Models - Marine sample accuracy per model
Appendix E. Spiked Marine (SM) Models Test Accuracy
Appendix F. Spiked Marine (SM) Models - Marine sample accuracy per model

Author Contributions

N.M.N designed the study and the sample organization scheme. H.C.A conceived of and supervised the project. N.M.N, J.B.C, and A.A.A.E collected the marine samples in Florida. E.I.P.N collected the marine sample from the Ohio State University's Center for Life Sciences Education salt water tank. E.I.P.N and K.A.D collected the Raman data. A.B.T, A.J.G, and S.G.W analyzed the unspiked marine samples with mass spectrometry. N.M.N organized and wrote the jupyter notebooks and completed the python analysis. N.M.N wrote the paper and all authors contributed to the edits.

Acknowledgements

N.M.N, J.B.C, A.A.A.E, and H.C.A acknowledges support from the National Science Foundation through the Center for Aerosol Impacts on Chemistry of the Environment (CAICE) under Grant No. CHE-1801971 for funding the sampling trip to Cocoa Beach, FL. N.M.N thanks S.M. and B.S. in Ohio State University's Center for Life Sciences Education for contributing marine sample F6 from their saltwater aquarium. 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.A.E. acknowledges funding from CAICE under Grant No. CHE-1801971. J.B.C. and H.C.A. acknowledge funding support from the National Science Foundation through Grant No. CHE-2102313. N.M.N thanks M.J.N, B.X.N and Z.A.N for their support throughout the project.

Works Cited

(1) Frossard, A. A.; Gérard, V.; Duplessis, P.; Kinsey, J. D.; Lu, X.; Zhu, Y.; Bisgrove, J.; Maben, J. R.; Long, M. S.; Chang, R. Y.-W.; Beaupré, S. R.; Kieber, D. J.; Keene, W. C.; Nozière, B.; Cohen, R. C. Properties of Seawater Surfactants Associated with Primary Marine Aerosol Particles Produced by Bursting Bubbles at a Model Air–Sea Interface. *Environ. Sci. Technol.* **2019**, *53* (16), 9407–9417. https://doi.org/10.1021/acs.est.9b02637.

(2) 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.

(3) Sauer, J. S.; Mayer, K. J.; Lee, C.; Alves, M. R.; Amiri, S.; Bahaveolos, C. J.; Franklin, E. B.; Crocker, D. R.; Dang, D.; Dinasquet, J.; Garofalo, L. A.; Kaluarachchi, C. P.; Kilgour, D. B.; Mael, L. E.; Mitts, B. A.; Moon, D. R.; Moore, A. N.; Morris, C. K.; Mullenmeister, C. A.; Ni, C.-M.; Pendergraft, M. A.; Petras, D.; Simpson, R. M. C.; Smith, S.; Tumminello, P. R.; Walker, J. L.; DeMott, P. J.; Farmer, D. K.; Goldstein, A. H.; Grassian, V. H.; Jaffe, J. S.; Malfatti, F.; Martz, T. R.; Slade, J. H.; Tivanski, A. V.; Bertram, T. H.; Cappa, C. D.; Prather, K. A. The Sea Spray Chemistry and Particle Evolution Study (SeaSCAPE): Overview and Experimental Methods. *Environ. Sci. Process. Impacts* **2022**, *24* (2), 290–315. https://doi.org/10.1039/D1EM00260K.

(4) Schiffer, J. M.; Mael, L. E.; Prather, K. A.; Amaro, R. E.; Grassian, V. H. Sea Spray Aerosol: Where Marine Biology Meets Atmospheric Chemistry. *ACS Cent. Sci.* **2018**, *4* (12), 1617–1623. https://doi.org/10.1021/acscentsci.8b00674.

(5) Frossard, A. A.; Long, M. S.; Keene, W. C.; Duplessis, P.; Kinsey, J. D.; Maben, J. R.; Kieber, D. J.; Chang, R. Y.-W.; Beaupré, S. R.; Cohen, R. C.; Lu, X.; Bisgrove, J.; Zhu, Y. Marine Aerosol Production via Detrainment of Bubble Plumes Generated in Natural Seawater With a Forced-Air Venturi. *J. Geophys. Res. Atmospheres* **2019**, *124* (20), 10931–10950. https://doi.org/10.1029/2019JD030299.

(6) Russell, L. M.; Hawkins, L. N.; Frossard, A. A.; Quinn, P. K.; Bates, T. S. Carbohydrate-like Composition of Submicron Atmospheric Particles and Their Production from Ocean Bubble Bursting. *Proc. Natl. Acad. Sci.* **2010**, *107* (15), 6652–6657. https://doi.org/10.1073/pnas.0908905107.

(7) 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.

(8) Dommer, A. C.; Wauer, N. A.; Angle, K. J.; Davasam, A.; Rubio, P.; Luo, M.; Morris, C. K.; Prather, K. A.; Grassian, V. H.; Amaro, R. E. Revealing the Impacts of Chemical Complexity on Submicrometer Sea Spray Aerosol Morphology. *ACS Cent. Sci.* **2023**, *9* (6), 1088–1103. https://doi.org/10.1021/acscentsci.3c00184.

(9) 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.

(10) Wu, H.; Liang, C.; Zhang, C.; Chang, H.; Zhang, X.; Zhang, Y.; Zhong, N.; Xu, Y.; Zhong, D.; He, X.; Zhang, L.; Ho, S.-H. Mechanisms and Enhancements on Harmful Algal Blooms Conversion to

Bioenergy Mediated with Dual-Functional Chitosan. *Appl. Energy* **2022**, *327*, 120142. https://doi.org/10.1016/j.apenergy.2022.120142.

(11) Chang, H.; Wu, H.; Zhang, L.; Wu, W.; Zhang, C.; Zhong, N.; Zhong, D.; Xu, Y.; He, X.; Yang, J.; Zhang, Y.; Zhang, T.; Liao, Q.; Ho, S.-H. Gradient Electro-Processing Strategy for Efficient Conversion of Harmful Algal Blooms to Biohythane with Mechanisms Insight. *Water Res.* **2022**, *222*, 118929. https://doi.org/10.1016/j.watres.2022.118929.

(12) Barile, P. J. Widespread Sewage Pollution of the Indian River Lagoon System, Florida (USA) Resolved by Spatial Analyses of Macroalgal Biogeochemistry. *Mar. Pollut. Bull.* **2018**, *128*, 557–574. https://doi.org/10.1016/j.marpolbul.2018.01.046.

(13) Passow, U.; Lee, K. Future Oil Spill Response Plans Require Integrated Analysis of Factors That Influence the Fate of Oil in the Ocean. *Curr. Opin. Chem. Eng.* **2022**, *36*, 100769. https://doi.org/10.1016/j.coche.2021.100769.

(14) Zapelini de Melo, A. P.; Hoff, R. B.; Molognoni, L.; de Oliveira, T.; Daguer, H.; Manique Barreto, P. L. Disasters with Oil Spills in the Oceans: Impacts on Food Safety and Analytical Control Methods. *Food Res. Int.* **2022**, *157*, 111366. https://doi.org/10.1016/j.foodres.2022.111366.

(15) Jiang, M.; Chen, S.; Li, J.; Liu, L. The Biological and Chemical Diversity of Tetramic Acid Compounds from Marine-Derived Microorganisms. *Mar. Drugs* **2020**, *18* (2), 114. https://doi.org/10.3390/md18020114.

(16) Xu, K.; Li, X.-Q.; Zhao, D.-L.; Zhang, P. Antifungal Secondary Metabolites Produced by the Fungal Endophytes: Chemical Diversity and Potential Use in the Development of Biopesticides. *Front. Microbiol.* **2021**, *12*.

(17) Yan, X.; Liu, J.; Leng, X.; Ouyang, H. Chemical Diversity and Biological Activity of Secondary Metabolites from Soft Coral Genus Sinularia since 2013. *Mar. Drugs* **2021**, *19* (6), 335. https://doi.org/10.3390/md19060335.

(18) Puzzarini, C.; Barone, V. Diving for Accurate Structures in the Ocean of Molecular Systems with the Help of Spectroscopy and Quantum Chemistry. *Acc. Chem. Res.* **2018**, *51* (2), 548–556. https://doi.org/10.1021/acs.accounts.7b00603.

(19) Back, H. de M.; Vargas Junior, E. C.; Alarcon, O. E.; Pottmaier, D. Training and Evaluating Machine Learning Algorithms for Ocean Microplastics Classification through Vibrational Spectroscopy. *Chemosphere* **2022**, *287*, 131903. https://doi.org/10.1016/j.chemosphere.2021.131903.

(20) Tripathy, B.; Dash, A.; Das, A. P. Detection of Environmental Microfiber Pollutants through Vibrational Spectroscopic Techniques: Recent Advances of Environmental Monitoring and Future Prospects. *Crit. Rev. Anal. Chem.* **2022**, *0* (0), 1–11. https://doi.org/10.1080/10408347.2022.2144994.

(21) ZHang, X.; Kirkwood, W. J.; Walz, P. M.; Peltzer, E. T.; Brewer, P. G. A Review of Advances in Deep-Ocean Raman Spectroscopy. *Appl. Spectrosc.* **2012**, *66* (3), 237–249. https://doi.org/10.1366/11-06539.

(22) *Raman Spectroscopy in the Deep Ocean: Successes and Challenges*. https://doi.org/10.1366/0003702041389319.

(23) Pirutin, S. K.; Jia, S.; Yusipovich, A. I.; Shank, M. A.; Parshina, E. Y.; Rubin, A. B. Vibrational Spectroscopy as a Tool for Bioanalytical and Biomonitoring Studies. *Int. J. Mol. Sci.* **2023**, *24* (8), 6947. https://doi.org/10.3390/ijms24086947.

(24) Wang, L.; Morita, A.; North, N. M.; Baumler, S. M.; Springfield, E. W.; Allen, H. C. Identification of Ion Pairs in Aqueous NaCl and KCl Solutions in Combination with Raman Spectroscopy, Molecular Dynamics, and Quantum Chemical Calculations. *J. Phys. Chem. B* **2023**, *127* (7), 1618–1627. https://doi.org/10.1021/acs.jpcb.2c07923.

(25) Qi, Y.; Hu, D.; Jiang, Y.; Wu, Z.; Zheng, M.; Chen, E. X.; Liang, Y.; Sadi, M. A.; Zhang, K.; Chen, Y. P. Recent Progresses in Machine Learning Assisted Raman Spectroscopy. *Adv. Opt. Mater.* **2023**, *11* (14), 2203104. https://doi.org/10.1002/adom.202203104.

(26) Guo, S.; Popp, J.; Bocklitz, T. Chemometric Analysis in Raman Spectroscopy from Experimental Design to Machine Learning–Based Modeling. *Nat. Protoc.* **2021**, *16* (12), 5426–5459. https://doi.org/10.1038/s41596-021-00620-3.

(27) Ke, J.; Gao, C.; Folgueiras-Amador, A. A.; Jolley, K. E.; de Frutos, O.; Mateos, C.; Rincón, J. A.; Brown, R. C. D.; Poliakoff, M.; George, M. W. Self-Optimization of Continuous Flow Electrochemical Synthesis Using Fourier Transform Infrared Spectroscopy and Gas Chromatography. *Appl. Spectrosc.* **2022**, *76* (1), 38–50. https://doi.org/10.1177/00037028211059848.

(28) Ralbovsky, N. M.; Lednev, I. K. Towards Development of a Novel Universal Medical Diagnostic Method: Raman Spectroscopy and Machine Learning. *Chem. Soc. Rev.* **2020**, *49* (20), 7428–7453. https://doi.org/10.1039/D0CS01019G.

(29) Ryzhikova, E.; Ralbovsky, N. M.; Sikirzhytski, V.; Kazakov, O.; Halamkova, L.; Quinn, J.; Zimmerman, E. A.; Lednev, I. K. Raman Spectroscopy and Machine Learning for Biomedical Applications: Alzheimer's Disease Diagnosis Based on the Analysis of Cerebrospinal Fluid. *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* **2021**, *248*, 119188. https://doi.org/10.1016/j.saa.2020.119188.

(30) Zhang, L.; Li, C.; Peng, D.; Yi, X.; He, S.; Liu, F.; Zheng, X.; Huang, W. E.; Zhao, L.; Huang, X. Raman Spectroscopy and Machine Learning for the Classification of Breast Cancers. *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* **2022**, *264*, 120300. https://doi.org/10.1016/j.saa.2021.120300.

(31) Coe, J. V.; Chen, Z.; Li, R.; Nystrom, S. V.; Butke, R.; Miller, B.; Hitchcock, C. L.; Allen, H. C.; Povoski, S. P.; Jr, E. W. M. Molecular Constituents of Colorectal Cancer Metastatic to the Liver by Imaging Infrared Spectroscopy. In *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XIII*; SPIE, 2015; Vol. 9328, pp 98–104. https://doi.org/10.1117/12.2079884.

(32) North, N.; Enders, A.; Clark, J.; Allen, H.; Duah, K. Saccharide Concentration Prediction from Proxy Sea Surface Microlayer Samples Analyzed via Infrared Spectroscopy and Quantitative Machine Learning. ChemRxiv January 4, 2024. https://doi.org/10.26434/chemrxiv-2023-d2ztk-v3.

(33) Grooms, A. J.; Burris, B. J.; Badu-Tawiah, A. K. Mass Spectrometry for Metabolomics Analysis: Applications in Neonatal and Cancer Screening. *Mass Spectrom. Rev. n/a* (n/a), e21826. https://doi.org/10.1002/mas.21826.

(34) Wang, H.; Liu, J.; Cooks, R. G.; Ouyang, Z. Paper Spray for Direct Analysis of Complex Mixtures Using Mass Spectrometry. *Angew. Chem. Int. Ed.* **2010**, *49* (5), 877–880. https://doi.org/10.1002/anie.200906314.

(35) S. Kulyk, D.; V. Baryshnikov, G.; S. Damale, P.; Maher, S.; K. Badu-Tawiah, A. Charge Inversion under Plasma-Nanodroplet Reaction Conditions Excludes Fischer Esterification for Unsaturated Fatty Acids: A Chemical Approach for Type II Isobaric Overlap. *Chem. Sci.* **2024**, *15* (3), 914–922. https://doi.org/10.1039/D3SC05369E.

(36) Amoah, E.; Kulyk, D. S.; Callam, C. S.; Hadad, C. M.; Badu-Tawiah, A. K. Mass Spectrometry Approach for Differentiation of Positional Isomers of Saccharides: Toward Direct Analysis of Rare Sugars. *Anal. Chem.* **2023**, *95* (13), 5635–5642. https://doi.org/10.1021/acs.analchem.2c05375.

(37) Harvey, G. W.; Burzell, L. A. A Simple Microlayer Method for Small Samples. *Limnol. Oceanogr.* **1972**, *17* (1), 156–157. https://doi.org/10.4319/lo.1972.17.1.0156.

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

(39) *sklearn.ensemble.RandomForestClassifier*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed 2024-01-13).

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

(41) *sklearn.ensemble.HistGradientBoostingRegressor*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html (accessed 2024-01-13).

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

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

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

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

(46) 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.

(47) Volpe, A. M.; Esser, B. K. Real-Time Ocean Chemistry for Improved Biogeochemical Observation in Dynamic Coastal Environments. *J. Mar. Syst.* **2002**, *36* (1), 51–74. https://doi.org/10.1016/S0924-7963(02)00125-2.

(48) Bates, N. R.; Astor, Y. M.; Church, M. J.; Currie, K.; Dore, J. E.; González-Dávila, M.; Lorenzoni, L.; Muller-Karger, F.; Olafsson, J.; Santana-Casiano, J. M. A Time-Series View of Changing Surface Ocean Chemistry Due to Ocean Uptake of Anthropogenic CO_2 and Ocean Acidification. *Oceanography* **2014**, *27* (1), 126–141.

(49) Takahashi, T.; Sutherland, S. C.; Chipman, D. W.; Goddard, J. G.; Ho, C.; Newberger, T.; Sweeney, C.; Munro, D. R. Climatological Distributions of pH, pCO2, Total CO2, Alkalinity, and CaCO3 Saturation in the Global Surface Ocean, and Temporal Changes at Selected Locations. *Mar. Chem.* **2014**, *164*, 95–125. https://doi.org/10.1016/j.marchem.2014.06.004.

(50) Hagens, M.; Middelburg, J. J. Attributing Seasonal pH Variability in Surface Ocean Waters to Governing Factors. *Geophys. Res. Lett.* **2016**, *43* (24), 12,528-12,537. https://doi.org/10.1002/2016GL071719.

(51) Matoo, O. B.; Lannig, G.; Bock, C.; Sokolova, I. M. Temperature but Not Ocean Acidification Affects Energy Metabolism and Enzyme Activities in the Blue Mussel, Mytilus Edulis. *Ecol. Evol.* **2021**, *11* (7), 3366–3379. https://doi.org/10.1002/ece3.7289.

(52) Xue, J.; Lee, C.; Wakeham, S. G.; Armstrong, R. A. Using Principal Components Analysis (PCA) with Cluster Analysis to Study the Organic Geochemistry of Sinking Particles in the Ocean. *Org. Geochem.* **2011**, *42* (4), 356–367. https://doi.org/10.1016/j.orggeochem.2011.01.012.

(53) Alonso-González, I. J.; Arístegui, J.; Lee, C.; Calafat, A. Regional and Temporal Variability of Sinking Organic Matter in the Subtropical Northeast Atlantic Ocean: A Biomarker Diagnosis. *Biogeosciences* **2010**, *7* (7), 2101–2115. https://doi.org/10.5194/bg-7-2101-2010.

(54) Wu, C.; Zhao, X.; Wu, X.; Wen, C.; Li, H.; Chen, X.; Peng, X. Exogenous Glycine and Serine Promote Growth and Antifungal Activity of Penicillium Citrinum W1 from the South-West Indian Ocean. *FEMS Microbiol. Lett.* **2015**, *362* (8), fnv040. https://doi.org/10.1093/femsle/fnv040.

(55) Triesch, N.; van Pinxteren, M.; Engel, A.; Herrmann, H. Concerted Measurements of Free Amino Acids at the Cabo Verde Islands: High Enrichments in Submicron Sea Spray Aerosol Particles and

Cloud Droplets. *Atmospheric Chem. Phys.* **2021**, *21* (1), 163–181. https://doi.org/10.5194/acp-21-163-2021.

(56) Triesch, N.; van Pinxteren, M.; Salter, M.; Stolle, C.; Pereira, R.; Zieger, P.; Herrmann, H. Sea Spray Aerosol Chamber Study on Selective Transfer and Enrichment of Free and Combined Amino Acids. *ACS Earth Space Chem.* **2021**, *5* (6), 1564–1574. https://doi.org/10.1021/acsearthspacechem.1c00080.

(57) Wu, H.; Liang, C.; Zhang, C.; Chang, H.; Zhang, X.; Zhang, Y.; Zhong, N.; Xu, Y.; Zhong, D.; He, X.; Zhang, L.; Ho, S.-H. Mechanisms and Enhancements on Harmful Algal Blooms Conversion to Bioenergy Mediated with Dual-Functional Chitosan. *Appl. Energy* **2022**, *327*, 120142. https://doi.org/10.1016/j.apenergy.2022.120142.

(58) Cai, J.; Chen, M.; Wang, G.; Pan, G.; Yu, P. Fermentative Hydrogen and Polyhydroxybutyrate Production from Pretreated Cyanobacterial Blooms. *Algal Res.* **2015**, *12*, 295–299. https://doi.org/10.1016/j.algal.2015.09.014.

(59) Srain, B. M.; Sobarzo, M.; Daneri, G.; González, H. E.; Testa, G.; Farías, L.; Schwarz, A.; Pérez, N.; Pantoja-Gutiérrez, S. Fermentation and Anaerobic Oxidation of Organic Carbon in the Oxygen Minimum Zone of the Upwelling Ecosystem Off Concepción, in Central Chile. *Front. Mar. Sci.* **2020**, *7*.

(60) US EPA, O. *The Effects: Dead Zones and Harmful Algal Blooms*. https://www.epa.gov/nutrientpollution/effects-dead-zones-and-harmful-algal-blooms (accessed 2024-01-16).

(61) *Eutrophication: Causes, Consequences, and Controls in Aquatic Ecosystems | Learn Science at Scitable*. https://www.nature.com/scitable/knowledge/library/eutrophication-causes-consequences-and-controls-in-aquatic-102364466/ (accessed 2024-01-16).

(62) Melzner, F.; Thomsen, J.; Koeve, W.; Oschlies, A.; Gutowska, M. A.; Bange, H. W.; Hansen, H. P.; Körtzinger, A. Future Ocean Acidification Will Be Amplified by Hypoxia in Coastal Habitats. *Mar. Biol.* **2013**, *160* (8), 1875–1888. https://doi.org/10.1007/s00227-012-1954-1.

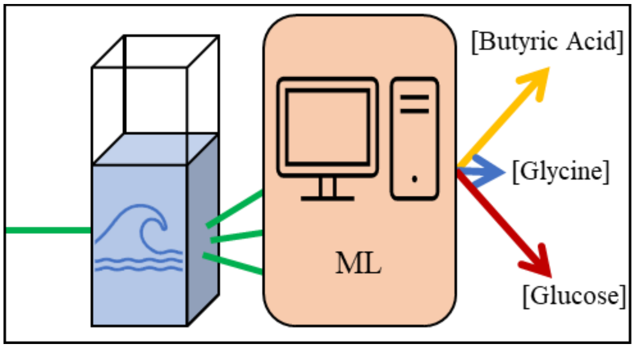
(63) Hasenecz, E. S.; Kaluarachchi, C. P.; Lee, H. D.; Tivanski, A. V.; Stone, E. A. Saccharide Transfer to Sea Spray Aerosol Enhanced by Surface Activity, Calcium, and Protein Interactions. *ACS Earth Space Chem.* **2019**, *3* (11), 2539–2548. https://doi.org/10.1021/acsearthspacechem.9b00197.

(64) Liu, N.; Yang, Y.; Li, F.; Ge, F.; Kuang, Y. Importance of Controlling pH-Depended Dissolved Inorganic Carbon to Prevent Algal Bloom Outbreaks. *Bioresour. Technol.* **2016**, *220*, 246–252. https://doi.org/10.1016/j.biortech.2016.08.059.

(65) Kisand, V.; Tammert, H. Bacterioplankton Strategies for Leucine and Glucose Uptake after a Cyanobacterial Bloom in an Eutrophic Shallow Lake. *Soil Biol. Biochem.* **2000**, *32* (13), 1965–1972. https://doi.org/10.1016/S0038-0717(00)00171-1.

(66) Zeppenfeld, S.; van Pinxteren, M.; Hartmann, M.; Bracher, A.; Stratmann, F.; Herrmann, H. Glucose as a Potential Chemical Marker for Ice Nucleating Activity in Arctic Seawater and Melt Pond Samples. *Environ. Sci. Technol.* **2019**, *53* (15), 8747–8756. https://doi.org/10.1021/acs.est.9b01469.

For TOC use only


[Butyric Acid]
ML
[Glycine]
[Glucose]