Leveraging Machine Learning Approaches to Predict Organic Carbon Abundance in Mars-Analog Hypersaline Lake Sediments

Floyd Nichols[1], Alexandra Pontefract[2], Andrew L. Masterson[1], Mia L. Thompson[1], Christopher E. Carr[3], Mia T. Tuccillo[1], Magdalena R. Osburn[1]

[1]Department of Earth & Planetary Sciences, Northwestern University, Evanston IL, USA
[2]Space Exploration Sector, Johns Hopkins University Applied Physics Laboratory, Laurel MD, USA
[3]School of Earth & Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA, USA

+ corresponding authors [floydnichols2025@u.northwestern.edu; maggie@northwestern.edu]

Key Points

- Predictive machine learning models have the potential to aid in life detection efforts beyond Earth using organic geochemical datasets.
- We predicted organic carbon abundance using only XRF-derived elemental abundances with greater than 80% accuracy.
- *Post-hoc* interpretation of our models highlights the importance of elements associated with clays in determining organic carbon concentrations.
- We've developed a user-friendly interface to improve the accessibility of our classification model to predict organic carbon from XRF-derived elemental abundances.

Abstract

Modern advancements in laboratory and instrumental techniques in astrobiology have improved our life detection capabilities on both Earth and beyond. These advancements have also increased the complexity of data often resulting in datasets that are characterized by complex and non-linear relationships. Machine learning methods are underutilized in astrobiology; however, these methods are extremely effective at revealing structure and patterns in complex datasets when paired with the right algorithms. Here, we employ a series of classification and regression algorithms to predict the abundance of organic carbon (OC) from X-ray fluorescence (XRF) data

in dynamic Mars-analog hypersaline lake sediments. More specifically, we constructed models using the random forest (RF), k-nearest neighbors (KNN), support vector machine (SVM), and logistic regression (LR) algorithms. Overall, our trained models showed good performance with predicting the abundance of OC, with accuracies from 80% to 94%. Our results show how applying predictive models to astrobiology datasets can help life detection efforts. Machine learning approaches such as classification and regression algorithms offer insight into complex data while providing agnostic insights, ultimately creating a more efficient search for OC. We applied our trained model on XRF data from Martian soil using PIXL and Odyssey datasets to produce probability predictions of OC abundance. Our predictions show a high probability that OC abundance is low which is comparable to OC data from recently landed missions. These results highlight the potential for machine learning models to be trained on data from analog environments on Earth and then transferred (transfer learning) to extraterrestrial targets.

Plain Language Summary

Modern datasets have become large as a by-product of the desire to discover unknown or characterize complex non-linear relationships. Machine learning approaches are extremely valuable for tackling such problems; however, those methods are underutilized in astrobiology and therefore have not been refined for these types of data. Here, we employ machine learning approaches to predict organic carbon abundance from a series of Mars-analog hypersaline lake sediment core and a freshwater lake from X-ray fluorescence-derived elemental abundances. Overall, our models successfully predicted organic carbon concentration, with average accuracies between 80% and 94% and root mean square errors within 1.0 wt% organic carbon. Furthermore, we applied our model to Martian instruments including PIXL and Odyssey. We compute probability predictions that corroborate organic carbon that has been measured on the Martian surface. Our study demonstrates the potential for machine learning methods to be employed to aid in life detection efforts.

Introduction

Exploring the potential for ancient or extant life beyond Earth poses many challenges including but not limited to sample selection, sample priority, and the search for an ideal site (Warren-Rhodes *et al*., 2023; Theiling *et al*., 2022). These challenges and limitations are often due

to a lack of consensus for what establishes favorable features or conditions for life detection beyond Earth (Theiling *et al*., 2022). Although underutilized in astrobiological research, machine learning methods can be used to mitigate these limitations. Machine learning excels at detecting patterns and structures within large and/or complex datasets (Warren-Rhodes *et al*., 2023; Theiling *et al*., 2022; Peaple *et al*., 2021). As such, machine learning methods can be extremely valuable for using data that is less resource intensive (such non-destructive spectroscopy including X-ray Fluorescence (XRF)) to predict information that is resource intensive (such as organic carbon (OC) or biosignature analysis; Warren-Rhodes *et al*., 2023; Jacq *et al*., 2019). For example, the Curiosity Rover can analyze an effectively unlimited number of samples using its remote laser-induced breakdown spectroscopy (LIBS) instrument, ChemCam (Maurice et al. 2012), but the Sample Analysis at Mars (SAM) instrument has a limited number of sample cups for analyzing powdered samples via gas chromatography mass spectroscopy (Mahaffy et al. 2012). Similarly, the Planetary Instrument for X-ray Lithochemistry (PIXL) and the Scanning Habitable Environments with Raman & Luminescence for Organics & Chemicals (SHERLOC) instruments on the Perseverance Rover can detect elemental abundances using XRF and scan for organics using Raman spectroscopy; however, it is limited in its ability to detect OC beyond the surface of minerals or within inclusions of transparent evaporite minerals (Bhartia *et al*., 2021).

For Earth-based studies, major element analysis via XRF is often used on sediment cores to reconstruct paleohydroclimate (Shea *et al*., 2022; Puleo *et al*., 2020; Zhang *et al*., 2020). In contrast, these analyses applied beyond Earth such as Martian research are mostly limited to exploring the surface to provide insight into the processes that have guided surficial processes and the evolution of the Martian crust-mantle system (Allwood *et al*., 2015; Hahn *et al*., 2007). As such, there are rapid and streamlined procedures for the determination of the major element composition of sediments on Earth. XRF can yield abundance information for more than 30 elements, ultimately producing relatively large and complex datasets. Due to this complexity, most researchers select a few of the most important elements to probe based on geologic relevance (Puleo *et al*., 2020; Evans *et al*., 2019; Zhang *et al*., 2019; Rothwell and Croudace, 2015).

Our greater understanding of Earth processes compared to Mars allows for selection of a few elements to explore; however, it is imperative that human biases are eliminated for studies beyond Earth. Thus, similarly to the maximum entropy principle, all pieces of available information should be utilized for exploration to provide agnostic insight into interpretation of the

data (Uffink, 1995). Ultimately, viewing data holistically reintroduces the challenges inherent to large and complex datasets. Here, we employ machine learning approaches to reveal patterns and make rapid and agnostic predictions of OC abundance in Holocene-aged Mars-analog hypersaline lake sediment cores from major XRF-derived element abundances. We utilize both unsupervised and supervised learning approaches to understand the structure of the data and make predictions of OC abundance. More specifically, we use unsupervised learning for exploratory data analysis. We then build classification models for broad scale characterization of sediments and then construct regression models for granular predictions of OC abundance.

Study Area

We targeted a series of hypersaline lakes located within the Cariboo Plateau of South-Central Interior British Columbia, Canada including Salt Lake, Last Chance Lake, and the Basque Lakes. The chemistry and geography of these lakes are described in greater detail in Nichols *et al*., 2023. In short, these lakes are closed basins and situated within a rain-shadow (~300 mm precipitation per year) contributing to their hypersaline nature. Despite the aridity, heavily vegetated catchment areas surround many of these lakes. Additionally, these lakes feature unusual chemistries dominated by high concentrations of magnesium sulfate (Salt Lake and the Basque Lakes) and of sodium carbonate (Last Chance Lake) ions making them ideal analogs to Mars. To improve the generalization of our model performance we also include a freshwater lake in Greenland near Narsarsuaq, informally named Mel3. Unlike the lakes from the Cariboo Plateau, Mel3 is open-basin, oligotrophic and surrounded by a sparsely vegetated catchment, with the closest meteorological station recording ~650 mm precipitation per year.

Figure 1. Location of Study Sites. A.) Last Chance Lake: 51º19'40.8" N, 121º38'9.6" W; **B.)** Salt Lake: 51º04'25.44" N, 121º35'11.244" W; **C.)** Basque Lake #1: 50º36'1.8" N, 121º21'32.4" W; **D.)** Basque Lake #2: 50º35'36.6" N, 121º20'58.2" W; **E.)** Basque Lake #4: 50º35'20.304" N, 121º20'34.397" W.; **F.)** Mel3: 61º07'46.81" N, 45º20'10.68" W. Photo Credit: Mitchell Barklage, PhD (Canadian Lakes) and Pete J.K. Puleo (Greenland Lake).

Methods

Geochemical Sediment Analysis

In British Columbia, Canada, we sampled nine sediment cores in the summer of 2018 and the summer of 2019 using a Unicoring device (2018) or an SDI Vibecore mini adapted to 3 inch diameter polycarbonate pipe (2019). Sediment water interfaces were stabilized with Gelzan and packing material. Sediment cores were capped and kept cold within 6 hours of collection, transported to the lab, and stored at 4ºC until further processing. The Mel3 sediment core was collected in August 2022 using an Aquatic Research Instruments “Universal” check-valve and

percussion coring system. The core was sealed, transported, and stored at 4ºC until analysis. Each sediment core was split using a GeoTek core splitter for bulk analysis and XRF analysis, respectively. The sediment elemental abundance (Ag, Al, As, Bi, Ca, Cd, Co, Cr, Cu, Fe, Hg, K, Light Elements (LE), Mg, Mn, Mo, Nb, Ni, P, Pb, Rb, S, Sb, Se, Si, Sn, Sr, Th, U, V, W, Y, Zn, Zr, Ti), magnetic susceptibility (MS), and core imaging was determined using a Geotek Multi-Sensor Core Logger (MSCL-S) paired with an Olympus Delta X-ray fluorescence (XRF) analyzer, a Bartington MS2E magnetic susceptibility meter, and a 50mm Canon camera of the core split. LE have weaker X-ray energies, thus, are harder to resolve individually. Accordingly, the LE are grouped into a single category. After scanning, bulk sediment material was taken throughout the Canadian cores in approximately 10 cm intervals for ^{14}C-dating and calibration. The sediment core was then sub-sectioned and homogenized into 3 cm increments for further organic analysis. Bulk sediment samples for ^{14}C dating were sent to the National Ocean Sciences Accelerator Mass Spectrometry facility at Woods Hole Oceanographic Institution.

Total organic carbon (TOC) and total organic nitrogen (TON) abundances were measured in the Northwestern Stable Isotope Biogeochemistry Lab with an elemental analyzer isotope ratio mass spectrometer (EA-IRMS; Costech 4010 EA coupled to a Thermo Delta V+ IRMS through a Conflo IV interface). Freeze-dried samples were weighed then were treated with 1M HCl to remove inorganic carbon and acid soluble salts, rinsed with MilliQ water, then freeze-dried and weighed again. Fourier Transform Infrared Spectrometric (FTIR) analysis on Mel3 sediments confirmed no presence of carbonates (So et al., 2020), and thus these sediments were not acidified before analysis. The homogenized samples were loaded into tin capsules for analysis. Standards were run every 10 samples including IU-acetanilide (precision: ± 1.0%) and urea (precision: ± 0.1%; Schimmelman *et al*., 2009). Additionally, TOC and TON values were corrected for the loss of acid soluble material determined gravimetrically.

Model Selection and Descriptions

All models in this study were constructed using the Python package Scikit-learn (Pedregosa, 2011). The explanatory variables were XRF-derived elemental abundances from sediments whereas the response variables were OC abundance divided amongst three classes (high, moderate, and low) as defined and justified below in the *Geochemical and Sediment Analysis* section. There are a variety of machine learning models available, but to limit the number of

models used in this study, we choose those that are most common and interpretable/explainable. Interpretable and explainable are defined as the degree to which a human can understand the cause of a decision or consistently predict the model's result and perform well with small datasets (Belle & Papantonis, 2021; Molnar, 2019; Probst *et al*., 2018; Kim *et al*., 2016; Pal & Mather, 2005). We employed t-distributed stochastic neighbor embedding (t-SNE; van der Maaten & Hinton, 2008), principal component analysis (PCA; Jolliffe & Cadima, 2016), logistic regression (Maalouf, 2011), k-nearest neighbor (Taunk *et al*., 2019), random forest (Biau & Scornet, 2016), and support vector machine (Cortes & Vapnik, 1995). Even though deep learning models such as neural networks have become increasingly common due to their state-of-the-art performance, they require large datasets (thousands of training examples) to perform well (Cronin, 2021), and are therefore not appropriate for our study. Below we give a brief description of the models and hyperparameter tuning we applied; however, a more detailed description of the mathematics involved can be found in the references above.

Unsupervised Learning: Dimensionality Reduction

We used unsupervised learning approaches to visualize information such as patterns exclusively from unlabeled data. There are a variety of such algorithms, but we chose two distinct yet common approaches to do so. This included t-distributed stochastic neighbor embedding (t-SNE) and principal component analysis (PCA). The t-SNE method is a non-linear dimensionality reduction data visualization technique that preserves the local structure of data by minimizing the Kullback-Leibler divergence between the two distributions with respect to the locations of the points in the map (van der Maaten & Hinton, 2008). t-SNE excels at revealing structure at many different scales which is very important for high-dimensional data (van der Maaten & Hinton, 2008). Conversely, PCA is a linear dimensionality reduction data visualization method that preserves the global structure of the data. To do so, PCA implements an orthogonal transformation, resulting in a number of components equal to or less to the number of original variables (Platzer, 2013). As such, using these two methods in tandem allows for the visualization of the data separability as a first-order analysis of the structure of the data.

Supervised Learning

We employed four classification algorithms to make broad scale classifications of OC abundance and three regression algorithms to make granular predictions of OC abundance. More specifically, we use logistic regression (LR), k-nearest neighbor (KNN), support vector machine (SVM), and random forest (RF) due to their commonality and explainability. Explainability can be defined as simulatability (ability to be simulated by a human), decomposability (ability to break down a model into parts), and algorithmic transparency (ability to understand the procedure the model goes through to generate its output; Belle & Papantonis, 2021; Lipton, 2016). We applied a Synthetic Minority Oversampling Technique (SMOTE) to address issues of imbalance between classes prior to model construction. The parameters for each model constructed were determined through optimization via an iterative run of parameters. Although by definition SVM and RF are not considered explainable algorithms, *post-hoc* analysis such as feature importance extraction was used to improve their explainability including mean decrease in impurity, permutation importance, and Shapley additive explanations.

We calculate a variety of metrics including accuracy, precision, recall, and F1 score to evaluate the errors of each ML model. These metrics were chosen due to their popularity and interpretability. In short, each metric can be defined as follows: *accuracy* is the measure of the fraction of predictions the model got correct:

$$\text{Eq. 1: } \left(\frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions}\right)$$

precision is the measure of the proportion of positive identifications correctly identified:

$$\text{Eq. 2: } \left(\frac{true\ positive}{true\ positive + false\ positive}\right)$$

recall is the measure of the proportion of actual positives identified correctly:

$$\text{Eq. 3: } \left(\frac{true\ positive}{true\ positive + false\ negatives}\right)$$

and *F1 score* is a combined measurement of recall and precision that computes how many times a model made a correct prediction across the entire dataset:

$$\text{Eq. 4: } \left(2\times\frac{Precision \times Recall}{Precision + Recall}\right)$$

In addition to evaluation of errors, we performed a two-step process to prevent and evaluate overfitting. First, we employed a cross-validation technique to check each model's ability to generalize the data. Specifically, we used a repeated stratified k-fold method for cross-validation. The model performance associated with each split was then averaged, allowing for a more accurate estimate of the model's performance. Additionally, since training machine learning models on

small datasets can cause the model to memorize all training examples, in turn leading to overfitting, our second step to prevent and evaluate overfitting was introducing varying levels of gaussian noise to the input data. Gaussian noise can lead to an improvement in the generalization of the model performance as it adds structured noise to the input data that is consistent with natural perturbations. Furthermore, when different levels are added, the robustness of each model can be evaluated by comparing the accuracy of the models with respect to each noise level.

Results

Age-Depth Modelling

The age of the sediment cores was calculated using Bayesian statistics age-modelling with the R Bacon package (Supporting Information; Blaauw & Christen, 2011). Due to a paucity of plant debris in the sediment cores, the age model was calculated using radiocarbon ages derived from bulk OM from the sediments. All sediments were Holocene-aged with ages ranging from present-day to ~6.5 ka with the oldest sediments found at the base of the Salt Lake core. A coherent age model was found for Salt Lake, Basque Lake #1, and Basque Lake #2 whereas significant reversals were noted in the Last Chance Lake cores. Disrupted ages in the upper sediments were likely caused by mixing during coring or through mixing via salt growth and ice growth throughout the year. The radiocarbon age model of the Mel3 core is not reported here; however, preliminary chronologies derived from terrestrial plant macrofossils throughout the sediments indicate the record spans roughly 4.2 ka to present.

Geochemical Sediment Analysis

The TOC for our samples showed a frequency distribution that had a high density around 2.5 wt% TOC that decreased with increasing wt% TOC (Figure 2). Generally, sediments are classified as low in OC if the TOC is less than 1.0 wt% (Fox *et al*., 2017). Considering our distribution and sparsity of samples with less than 1.0 wt% TOC we considered samples less than 2.5 wt% as low. Conversely, sediments are generally classified as organic rich when the TOC is greater than 10.0 wt% (Fuller *et al*., 2021; Fox *et al*., 2017). We used the distribution from our data and previously defined boundaries to determine our boundary conditions for the three-class model where low concentration is less than 2.5 wt% TOC, moderate concentrations are between 2.5 and 10.0 wt%, and high concentrations are greater than 10.0 wt% TOC.

The TOC of our samples ranged from 0.3 wt% to 20.4 wt% (Figure 3). Overall, we observed the highest average TOC from Basque Lake #2 of 8.2 wt% with minimum and maximum values of 0.6 wt% and 20.4 wt%, respectively. Conversely, Last Chance Lake had the lowest average TOC of 1.5 wt% with minimum and maximum values of 0.3 wt% and 2.9 wt%, respectively. The remaining lakes including Basque Lake #1, Salt Lake, and Mel3 had similar average TOC values of 4.1 wt%, 5.0 wt%, and 6.0 wt%, respectively. We observed minimum TOC values of 0.8 wt%, 0.7 wt%, and 2.0 wt% for Basque Lake #1, Salt Lake, and Mel3, respectively. Conversely, we observe maximum TOC values of 9.8 wt%, 10.5 wt%, and 9.2 wt%.

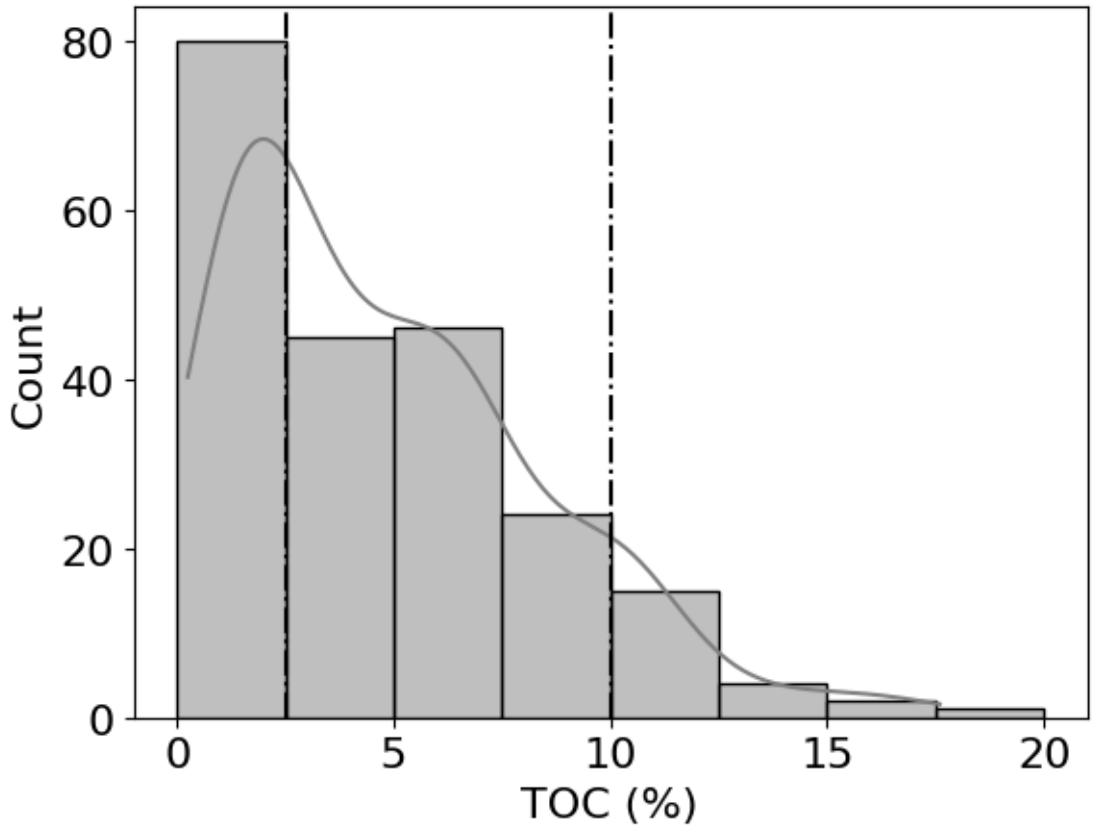


Figure 2. Histogram and Kernel Density Distribution of Organic Carbon Abundance. For this study, we divided OC into three classes: low, moderate, and high (dashed lines). This classification was based on a combination of the distribution of OC in our sediments and general classification of sediment OC concentration (Fuller *et al*., 2021; Fox *et al*., 2017) where low < 2.5 wt%, 2.5 wt% < moderate < 10.0 wt%, and 10.0 wt% < High. We employed a Synthetic Minority Oversampling Technique (SMOTE) to address the issue of class imbalance prior to model construction.

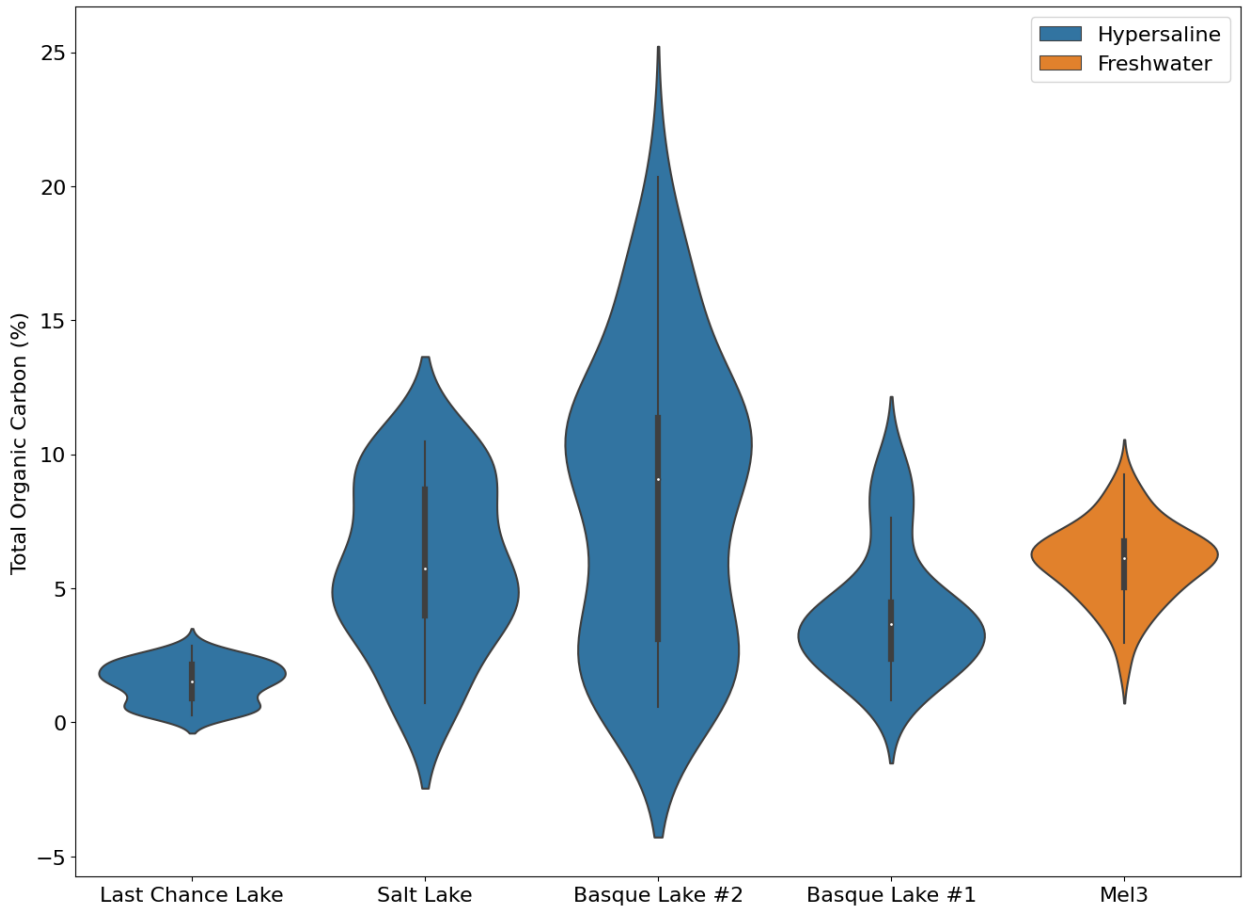


Figure 3. Violin plot of the distribution of TOC by lake. Blue represents hypersaline Mars-analog systems and orange represents the freshwater system.

Our XRF analysis shows that the most abundant elements on average were the light elements (LE, atomic mass < Mg); however, of the individual elements that can be resolved, the five most abundant were Mg, Fe, Si, S, and Ca (Figure 4). The average concentrations of the most abundant elements were 2.7 wt%, 1.3 wt%, 3.4 wt%, 3.8 wt%, and 2.4 wt%, respectively. The majority of the other elements had concentrations less than 0.01 wt%.

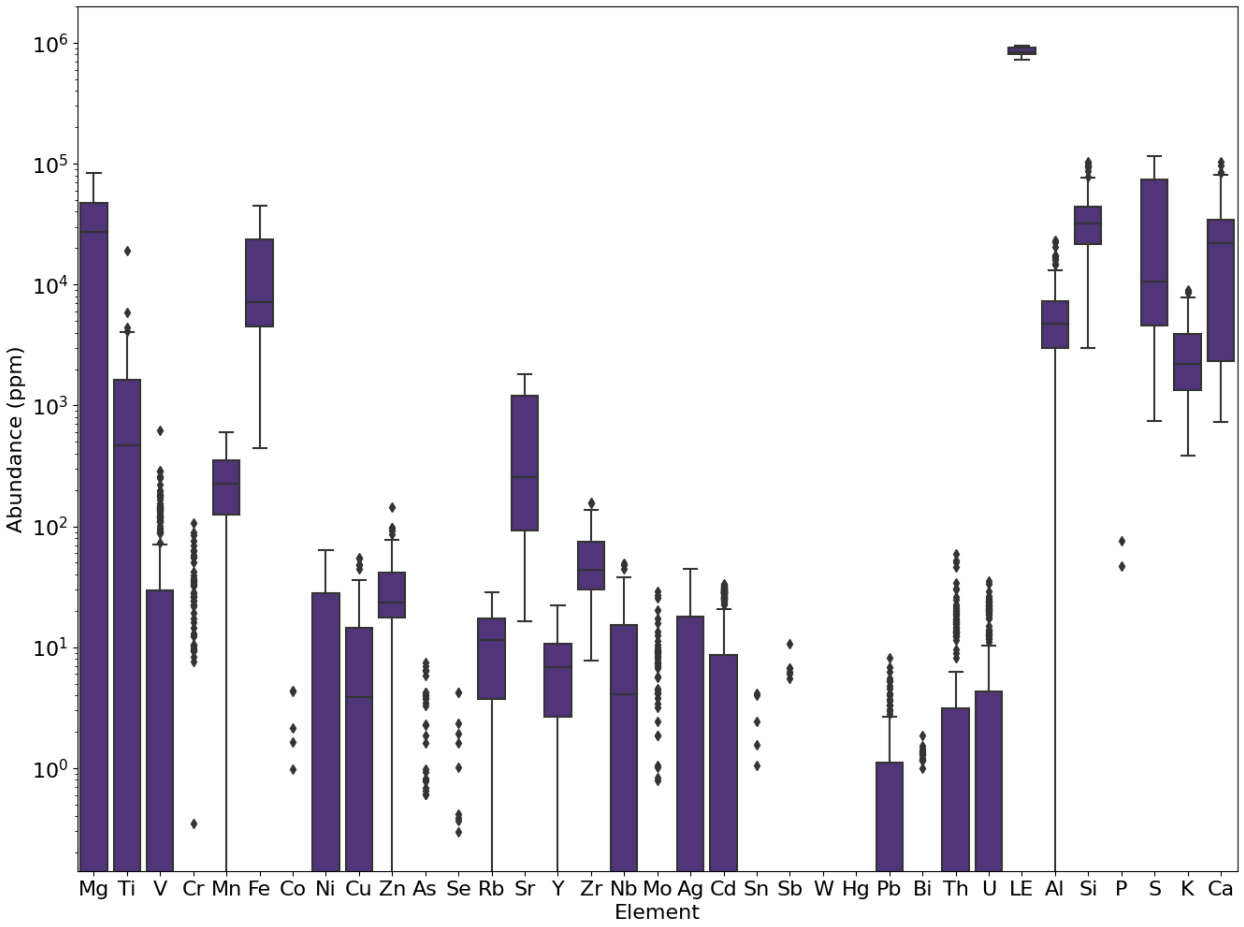


Figure 4. Box and whisker plot of the distribution of detected elements from XRF in our samples with respect to their concentrations in parts per million (ppm). Points represent outliers (points farther than 1.5 × interquartile range). The y-axis is plotted on a $\log_{10}$ scale. Additionally, there is an absence of a boxplot in a few elements due to extremely low abundances and little variance.

Machine Learning

Unsupervised Learning

We first employed unsupervised learning approaches to visualize the structure of the data. The two main approaches used were t-SNE and PCA. In our analysis, we observed that t-SNE showed 4 distinct clusters while PCA showed 3 distinct clusters (Figure 5). More specifically, in the t-SNE approach, each OC classification generally formed its own cluster. Conversely, with PCA, there are clear clusters of low and moderate OC classes, but there is more overlap with

respect to the high OC class. Overall, both methods were able to identify distinct clusters within the data, but t-SNE showed better structure in the clusters.

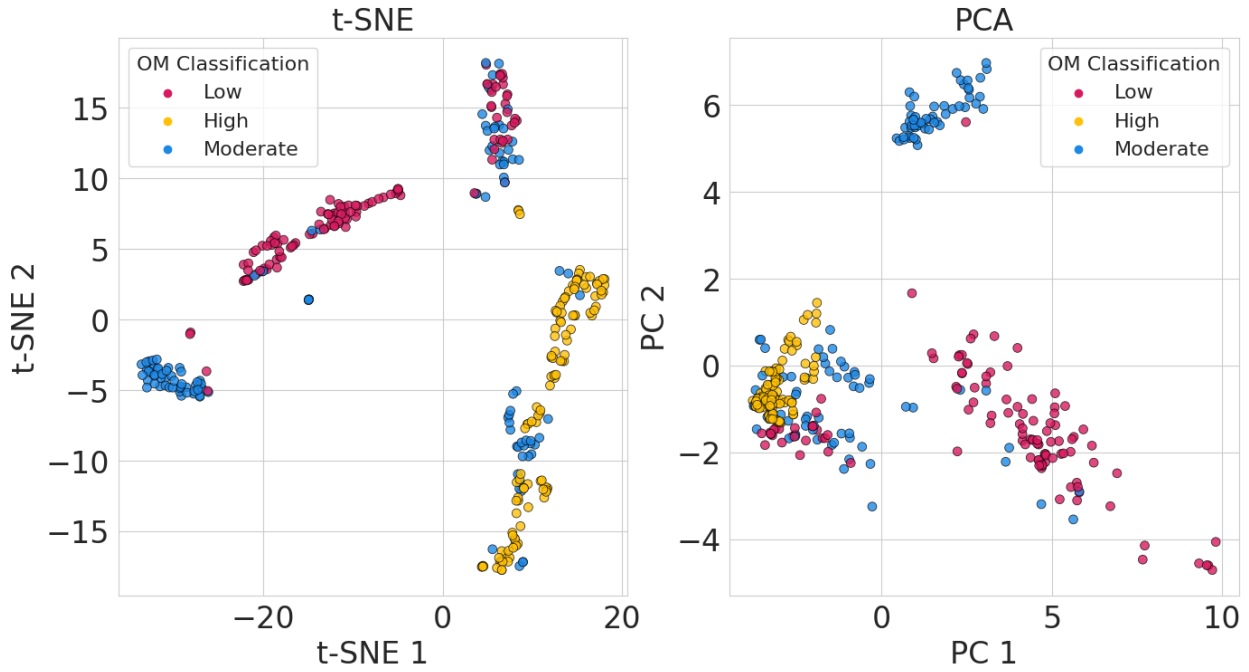


Figure 5. Unsupervised Learning: Dimensionality Reduction. We applied two unsupervised learning methods including t-distributed stochastic neighbor embedding (t-SNE; *left*) and principal component analysis (PCA; *right*). Both methods clustered the data to some capacity; however, t-SNE revealed more structure in the data as shown by the four distinct clusters representing the different OC classifications.

Supervised Learning

The accuracies of our supervised models ranged from 80% to 94% (Table 1). More specifically, KNN performed with an accuracy of 90%, precision of 89%, recall of 90% and F1 score of 89%. Random Forest (RF) performed with an accuracy of 94%, precision of 94%, recall of 95% and F1 score of 94%. SVM performed with an accuracy of 80%, precision of 81%, recall of 95%, and F1 score of 80%. Logistic regression performed with an accuracy of 91%, precision of 91%, recall of 91% and F1 score of 91%. We used cross-validation to evaluate the generalization capabilities of our models. The cross-validation accuracies of our RF, KNN, SVM, and LR were 91%, 89%, 84%, and 88%, respectively.

Table 1. Model Performance Metrics.

Model	Training Accuracy	Precision	Recall	F1	CV
KNN	90	89	90	89	88
RF	94	94	95	94	91
SVM	80	81	81	80	84
LR	91	91	91	91	88

To further interrogate the performance of our models, we computed a confusion matrix for each model (Figure 6). In short, a confusion matrix allows for the visualization of correctly labeled and mislabeled classifications. Generally, all models correctly labeled high and low OC abundance with high accuracy (~80%). Conversely, all models correctly labeled moderate OC abundance with a lower accuracy. The SVM and KNN models were more likely to misclassify moderate OC abundance (54% and 70%, respectively) than the RF and LR (79%). We computed feature importance calculations using mean decrease in impurity, permutation importance, and Shapley additive explanations to extract additional information from our models, (Figure 7). These calculations highlight that the top elemental abundances contributing to the classification of OC were Al, Fe, Si, Ti, U, and Zn.

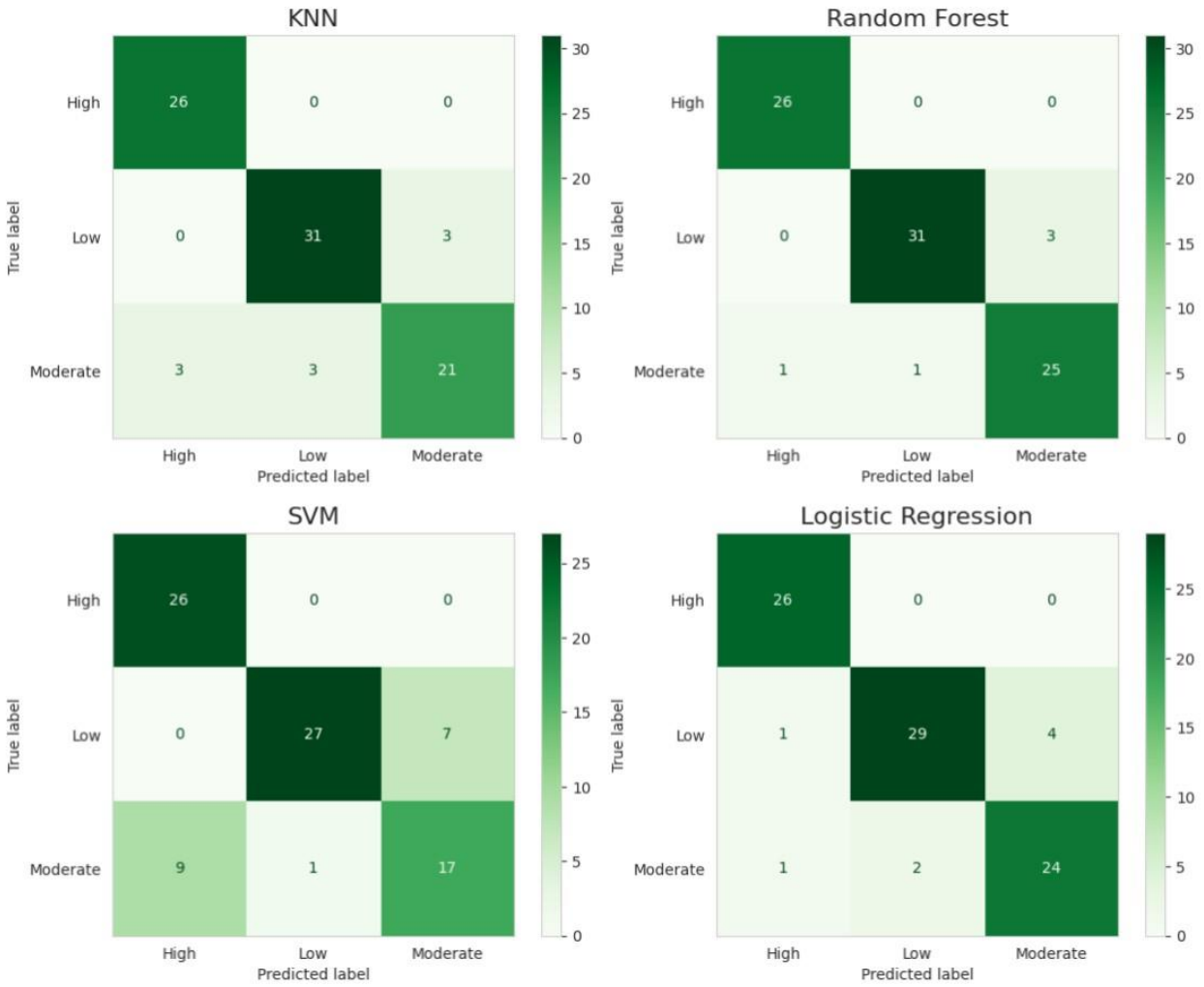
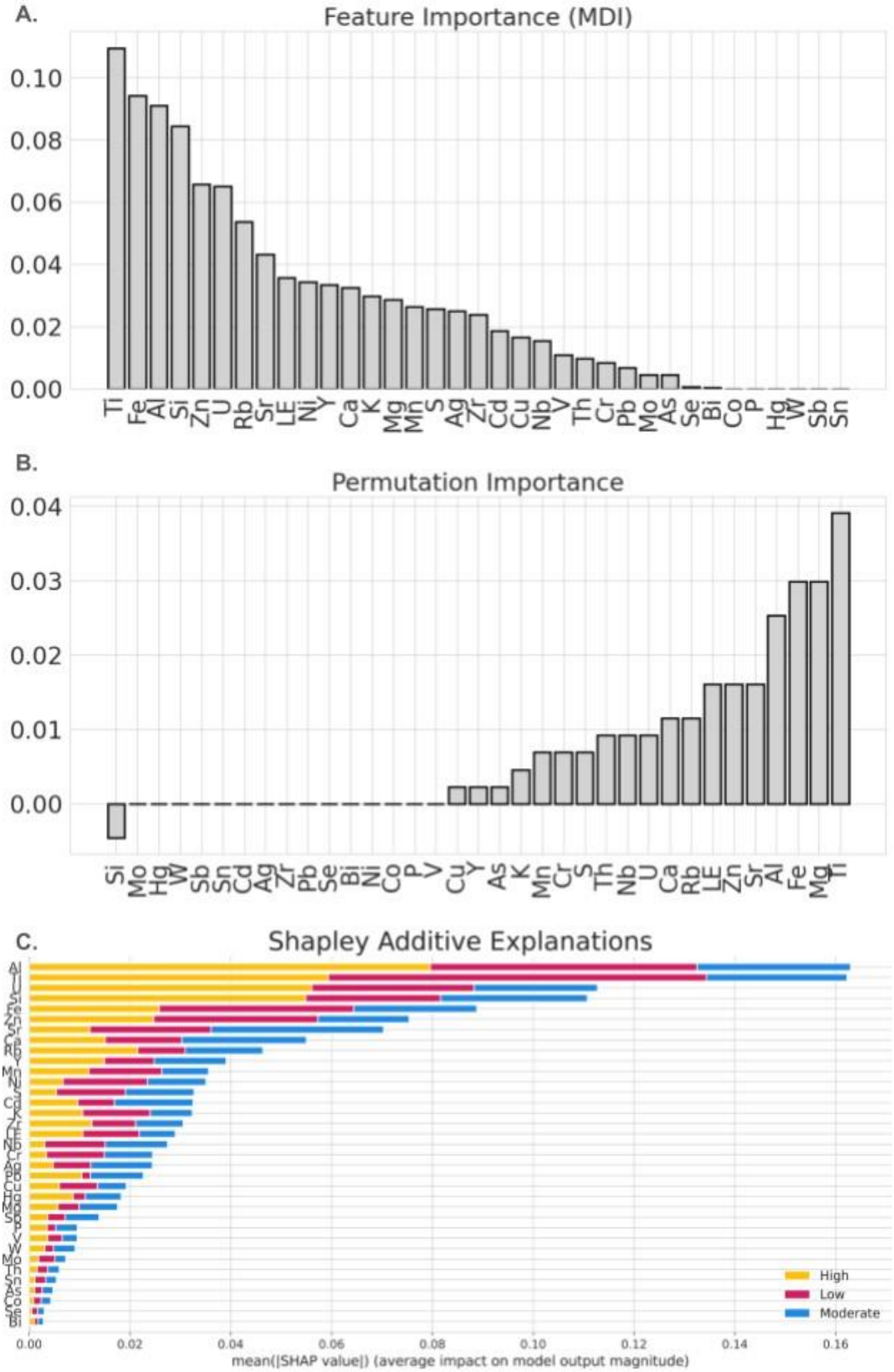


Figure 6. Confusion Matrix. Visual representation of correctly labeled classes against mislabeled classes.


A.
Feature Importance (MDI)
0.10
0.08
0.06
0.04
0.02
0.00
Ti Fe Al Si Zn U Rb Sr LE Ni Y Ca K Mg Mn S Ag Zr Cd Cu Nb V Th Cr Pb Mo As Se Bi Co P Hg W Sb Sn
B.
Permutation Importance
0.04
0.03
0.02
0.01
0.00
Si Mo Hg W Sb Sn Cd Ag Zr Pb Se Bi Ni Co P V Cu Y As K Mn Cr S Th Nb U Ca Rb LE Zn Sr Al Fe Mg Ti
C.
Shapley Additive Explanations
High
Low
Moderate
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16
mean(|SHAP value|) (average impact on model output magnitude)

334

Figure 7. Interpretation of Model Results. A.) Feature Importance Mean Decrease in Impurity (MDI), **B.)** Permutation Importance, **C.)** Shapley Additive Explanations.

We also introduced three levels of gaussian noise (1%, 5%, and 10%) to improve generalization and evaluate overfitting for each model (Table 2). We observe that in general, the model performance decreases slightly with added gaussian noise. More specifically, there is a gradual decrease in overall model performance including accuracy, precision, recall, and F1 scores with each gaussian noise percentage step. Despite the performance of the models decreasing with respect to the original models with no added noise, we still obtain accuracies above 80% even at 10% added gaussian noise.

Table 2. Model Performances with Varying Levels of Gaussian Noise.

Gaussian Noise	Model	Accuracy	Precision	Recall	F1
1%	KNN	87	88	87	87
	RF	84	84	84	83
	SVM	85	86	85	85
	LR	90	90	89	90
5%	KNN	86	87	86	86
	RF	84	84	84	83
	SVM	82	83	81	81
	LR	89	89	88	88
10%	KNN	84	86	85	83
	RF	85	87	86	84
	SVM	83	85	84	82
	LR	84	84	85	84

In addition to constructing a classification model, we built three regression models including random forest regression (RF), support vector regression (SVR), and k-nearest neighbor regression (KNN) to examine OC prediction at a more granular level. Overall, our models show good performance across all metrics including coefficient of determination (r^2), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percent error (MAPE; Table 3). Of the models, RF had the best metrics, followed by KNN, and then SVR (Table 3). We compute a moderate average r^2 value of 0.63 across all models. Additionally, we compute an average RMSE value of 1.05 across all models.

Table 3. Statistic metrics for our regression models.

Model	r^2	RMSE	MAE	MAPE
RF	0.80	0.80	0.61	0.17
SVR	0.51	1.24	0.95	0.37
KNN	0.60	1.12	0.85	0.29

Discussion

Sediment Geochemistry

On Earth, several competing mechanisms control the abundance of OC in sediments including mineral protection from clays and salts, selective preservation of refractory biomolecules, chemical speciation of trace metals and their affinity towards organic compounds, and redox conditions (Hemingway et al, 2019; Burdige, 2007; Aubrey *et* al., 2006; Bilali *et al.*, 2002; Hedges et al, 2001). Our feature extraction analysis converged on common elements that had the greatest effect on the model predictions for OC abundance. These elements included Al, Fe, Si, Ti, U, and Zn. The abundance of OC in sedimentary systems can be indirectly linked to elemental abundances (Evans *et al*., 2019; Bilali *et al*., 2002). For instance, the concentration of OC in lake systems often depends on aridity, salinity, and other hydrologic variations which can be captured by elements linked to phyllosilicates such as Al, Fe, and Si or detrital elements such as Ti and Zn (Evans *et al*., 2019; Zhang *et al*., 2019; Rothwell and Croudace, 2015; Bilali *et al*., 2002). Conversely, OC may be directly linked to elements such as trace metals including U in which OC influences its mobilization within sediments (Bone *et al*., 2019; Bone *et al*., 2017; Cumberland *et al*., 2016; Bilali *et al*., 2002). These relationships can be complex and challenging to unravel in lake sedimentary systems as these elements and OC can show varying degrees of correlations due to their potential to become bound to the sediment or released into the water and removed from the system (Frings *et al*., 2014; Makinen *et al*., 2005; Bilali *et al.*, 2002). Improving our understanding of these complex mechanisms will be key to understanding areas of interest for OC on the Martian landscape.

Appropriately, we capture a wide range of OC concentrations from lean samples (<1 wt% TOC) to organic rich samples (>10 wt% TOC) in addition to varying levels of XRF-derived elemental abundances. The varying levels of TOC and elemental abundances, especially within an individual lake, highlight the dynamic nature of these lakes. In our analysis of the relationship between TOC and the top elements from our feature extraction analysis, we show that Al, Fe, Si,

Ti, and Zn have very strong positive correlations between each other (Figure 8). This strongly suggests that Al, Fe, Si, Ti, and Zn inputs are linked to weathering as they are generally associated with phyllosilicates and detrital elements (Evans *et al*., 2019; Zhang *et al*., 2019; Rothwell and Croudace, 2015). Conversely, TOC shows moderately positive correlations with U. This relationship between OC and U is potentially due to organic ligands forming stable complexes with U (Bone *et al*., 2019; Bone *et al*., 2017; Bilali *et al*., 2002). Bone *et al*., 2017 show that as OC increases, U increases proportionally as the U has a higher potential to adsorb to the surface of organic matter. Another potential mechanism controlling this relationship between OC and U is due to a redox effect. It is known that oxidized U (VI) is soluble whereas reduced U (IV) is insoluble. Organic matter acts as the reductant that immobilizes U through its reduction (Cumberland *et al*., 2016). Ultimately, both of these potential mechanisms influence the fate of U in subsurface sediments with the same behavior that we observe in the sediments in this study.

On the other hand, we observe moderately negative correlations with Al, Fe, Si, Ti, and Zn (Figure 8). We suggest the following mechanisms for the observed correlation. As previously mentioned, the concentrations of Al, Fe, Si, Ti and Zn are controlled by hydrologic processes including weathering and aridity due to their association with phyllosilicates and detrital material. As weathering increases and aridity decreases, the loading of these elements to the sedimentary system increases as described by the chemical index of alteration (Wang *et al*., 2020). Considering the negative correlation that we observe with OC abundance with respect to Al, Fe, Si, Ti, and Zn we suggest that periods of increased lake desiccation and aridity are the primary driving forces for increased OC. It has been shown that the increasing desiccation of lakes drives an increase of nutrients and ions to the system, ultimately increasing OC productivity (Sarkar *et al*., 2023; Duarte *et al*., 2008; Jones and Decampo, 2003). Lake desiccation can also drive evaporite mineral dilution effects such that as the sediments fluctuate between low and high weight percentage of soluble evaporite minerals, the signal of the elements associated with phyllosilicates and detrital material are diluted out. This effect would also influence the negative correlation that we observe with OC. Additionally, an increase in salinity is known to slow remineralization rates of OC, resulting in a positive correlation of OC abundance with drier periods (Jellison *et al*., 1996).

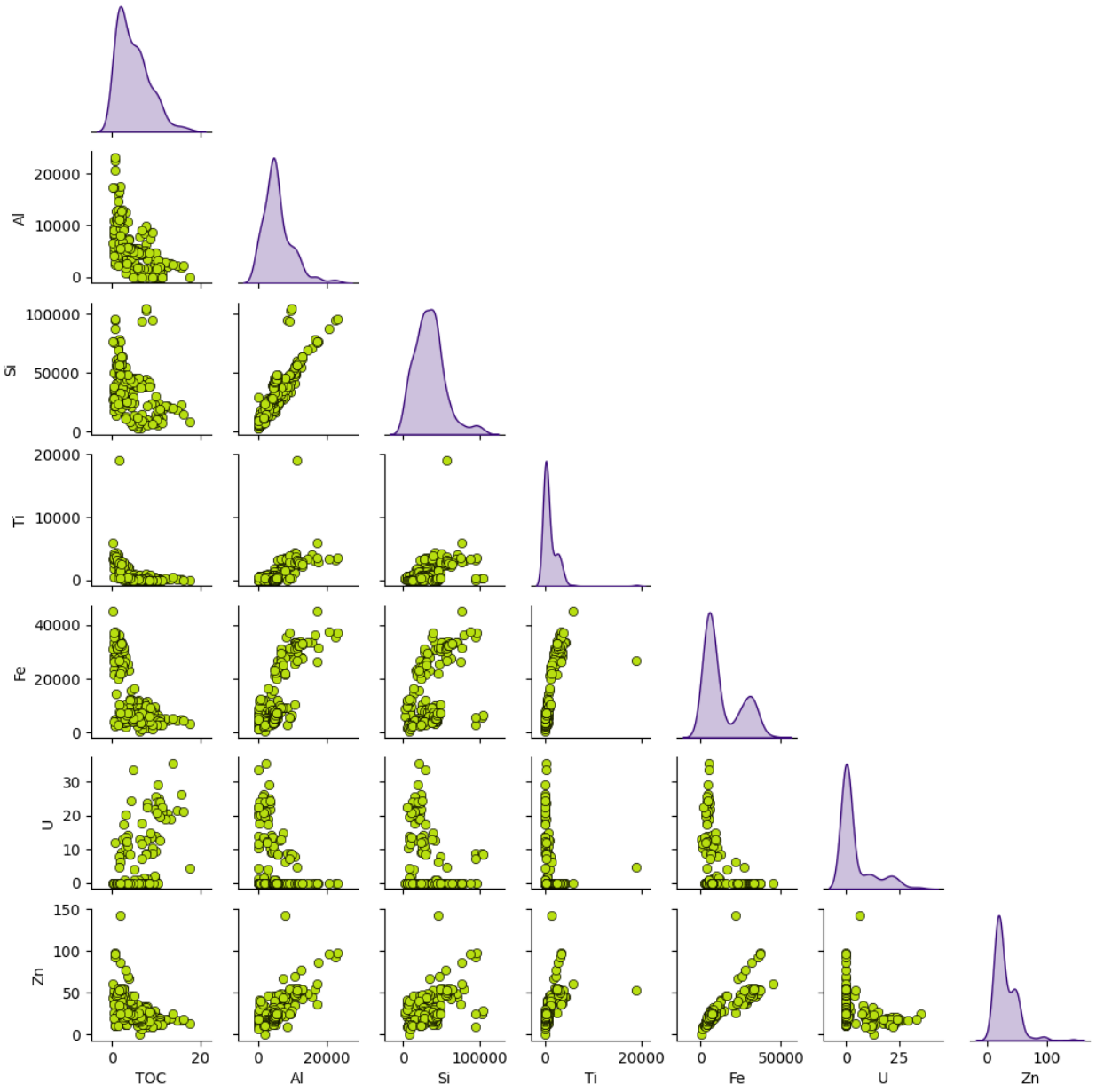


Figure 8. Pairplot of relationship between TOC and the top elemental abundance features. Using the top five features from our model, we compute the pairwise relationship to evaluate their correlation with OM. Abundance is in ppm.

Model Performance and Validation

PCA is widely used for visualization of high dimensional data and data pre-processing; however, while robust in some scenarios such as handling linearly separable data, it does not always perform well where the data has a non-linear structure. Additionally, the clustering of

points in a PCA is highly affected by outliers. In contrast, t-SNE has the ability to reveal more structure, underlying patterns, assess data separability, and better handles outliers. Despite these advantages, t-SNE is also limited in its use as a critical drawback of its approach compared to PCA is its inability to preserve the global structure of the data, thus, making it useful only for data visualization and not pre-preprocessing. As such, the tandem use of these methods allows for holistic exploratory data analysis.

In this study, we show that t-SNE clustered the data more distinctly with respect to PCA. These results suggest that our data is complex and has a non-linear structure, as highlighted by our OC and elemental abundance data, rather than a linear structure. The supervised learning algorithms that we employed validates our results from the unsupervised learning as the highest performing algorithms for both classification and regression analysis were RF and KNN. Similarly, to t-SNE, RF and KNN are generally better with handling data with non-linearities compared to SVM or LR (Acito, 2023; Chauhan *et al*., 2019; Auret & Aldrich, 2012; Maalouf, 2011). It is important to note that SVM can overcome non-linear challenges when different kernels are employed such as the radial basis function (RBF); however, it often does not perform as well algorithms specifically designed to handle non-linearities (Dong *et al*., 2014). Our results from our unsupervised and supervised learning approaches are promising as it is very likely when considering life detection beyond Earth, specifically Mars, the environments for which life may have thrived were likely in dynamic and non-linear states as the planet evolved. Subsequently, choosing the correct algorithms for predictions will require knowledge of the environment.

Additionally, we used our trained models to predict the OC abundance of a new sediment core to validate the models and produce a probability map of the OC abundance within that core (Figure 9). This sediment core (Last Chance Lake 5) from British Columbia was relatively low in OC with the exception of the topmost point which was greater than 2.5% OC. All classification models showed good performance (~80% accuracy) with predicting the actual TOC classification of the sediment core. We then examine the predictive capability of our regression models for the same sediment core (Figure 10). Similarly, to our classification models, all three regression models were generally able to predict the TOC to within 1.0 wt% of the observed data. More specifically, ~80% of the predicted samples were within 1.0 wt% TOC of the observed data. In general, RF and KNN outperformed SVR as they predicted TOC more closely to the real data. It is worth mentioning that all three of our models generally predicted slightly higher than the observed

values. This behavior is likely due to the tested core (Last Chance Lake 5) being exceptionally lean in TOC (mean = 1.8 wt%) with respect to the training data, thus overpredicting values.

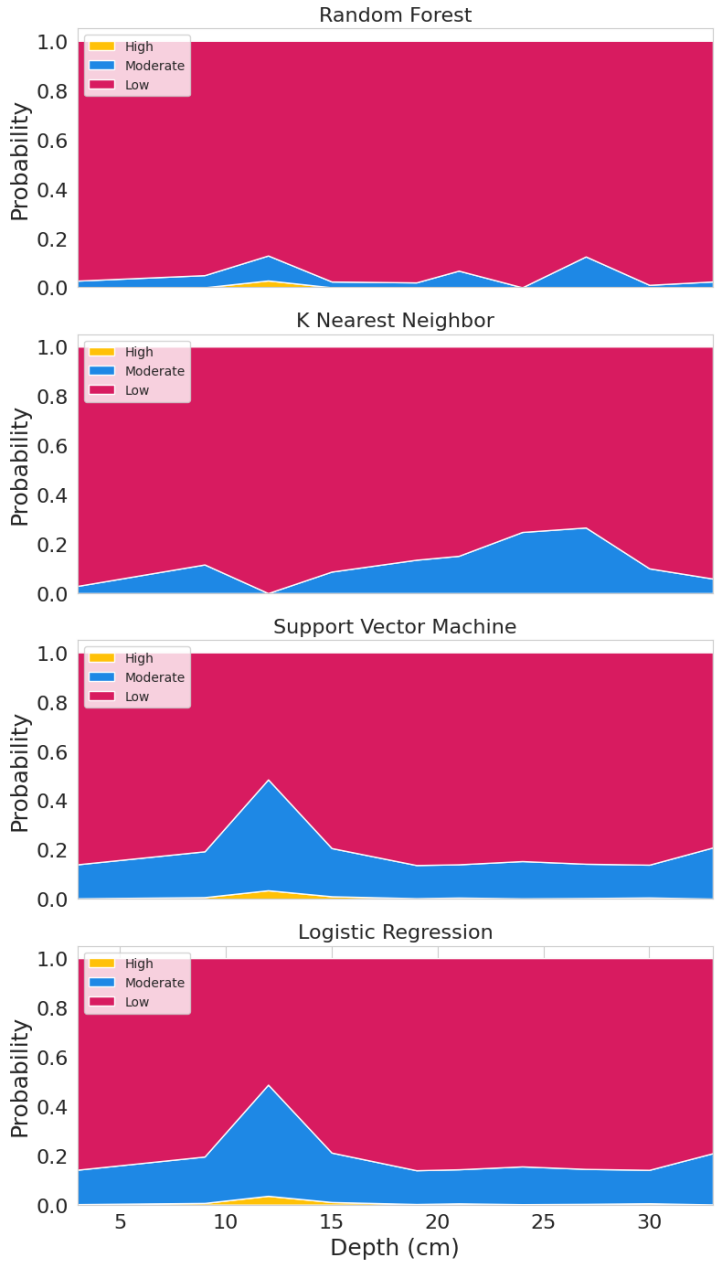


Figure 9. Model Validation. Area Plot of Classification Probabilities. To validate the models, we apply new data that the models have never seen. Similar to the training sets, all models classified the sediment organic carbon abundance with ~80% accuracy.

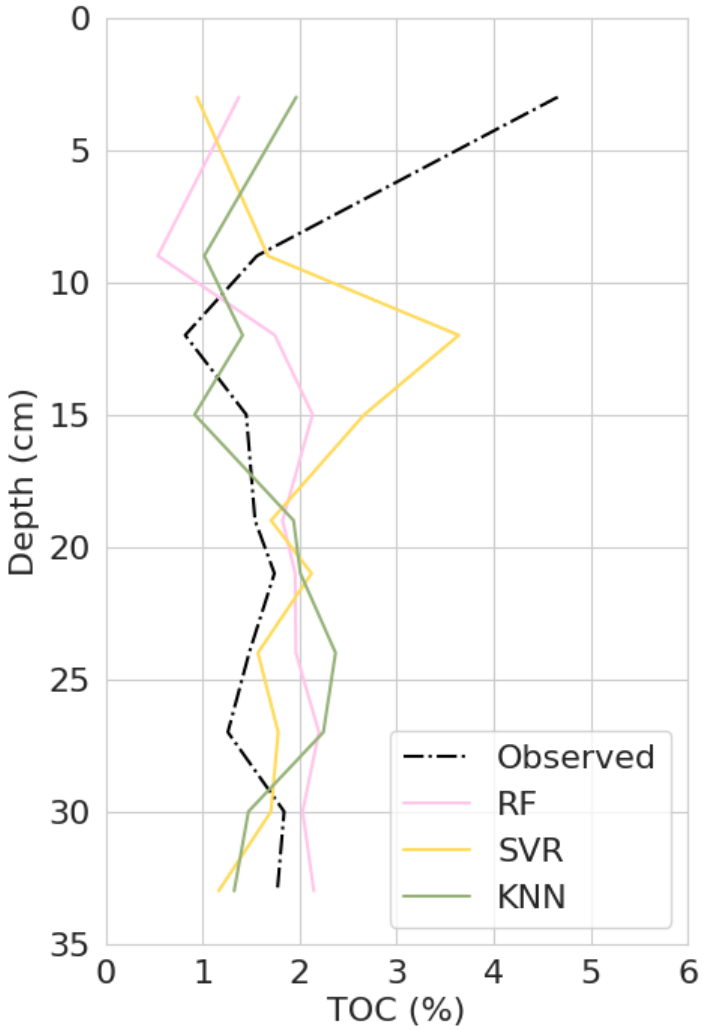


Figure 10. Regression Model Comparison to Real Data. We plot our modelled predictions (solid-colored lines) with respect to the observed data of Last Chance Lake 5 (black dashed line) to compare the accuracy of their predictions. Generally, all models predicted within 1.0% of the actual data. Additionally, RF and KNN outperformed SVR.

Graphical User Interface: Organic Matter Abundance Predictor (OMAP)

To improve accessibility of the model and provide rapid prediction of OC abundance from XRF-derived elemental abundances, we have developed an open-source graphical user interface: Organic Matter Abundance Predictor (OMAP; Nichols, 2023). This application is an interactive data visualization tool and predictor for the constructed model (Figure 11). Due to the random forest algorithm having the best performance for our testing and validation, we use it as the primary algorithm for the application. The application consists of three main components including model

data and performance visualization, organic carbon probability predictor, and geographic distribution of samples that the model is based on. Additionally, this application serves as an open-source database for others to add OC and XRF-derived elemental abundances from other sedimentary systems to improve and expand upon the model.

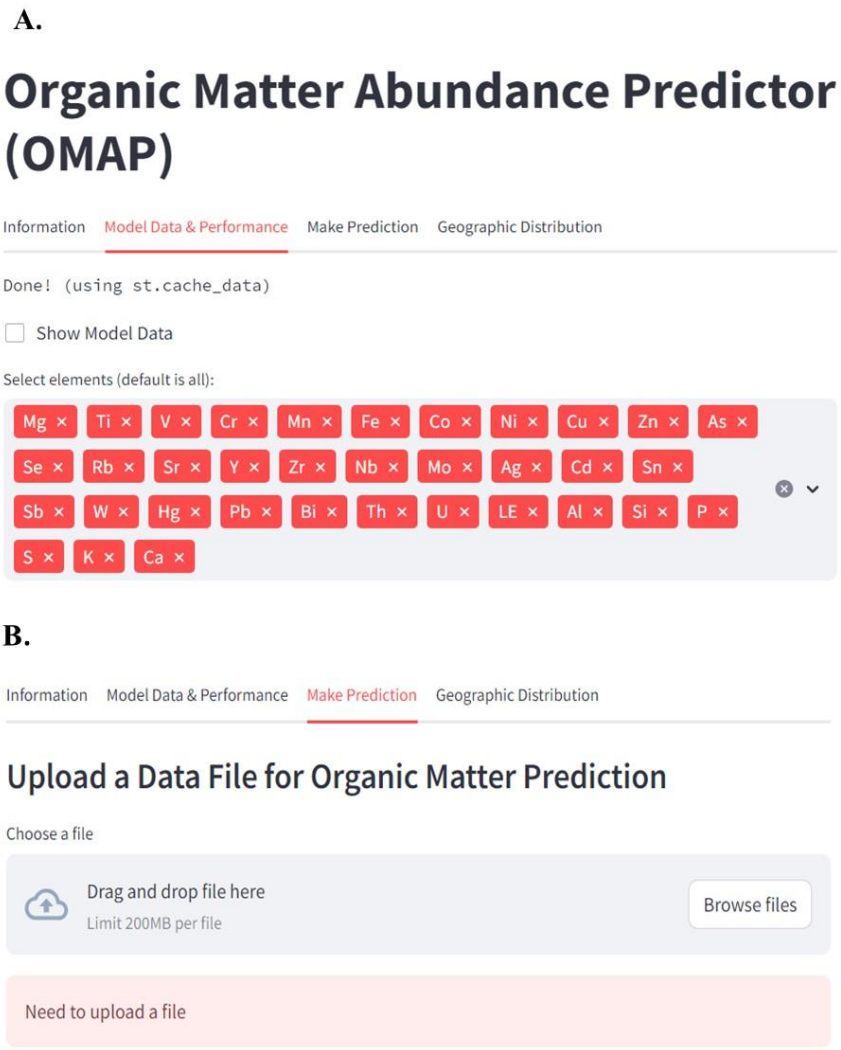


Figure 11. Layout of the Graphical User Interface: Organic Matter Abundance Predictor (OMAP). A.) Model data and performance tab which allows the user to selectively choose the parameters to make OC predictions including specific elements and boundary conditions for OC classification. Additionally, once the elements and boundary conditions are chosen, the data dimensionality reduction and model performance metrics will be printed and **B.)** The make a prediction tab allows for the user to make a probability prediction of OC abundance based on the

elements and OC boundary conditions chosen in the Model data and performance data via uploading a CSV that includes a sample column and elements of interest.

Transfer Learning: Mars Regolith

As a proof-of-concept, we applied our model to the average elemental abundance composition from Mars regolith samples as determined by Perseverance's Planetary Instrument for X-ray Lithochemistry (PIXL; Christian *et al*., 2023) (Table 4). Our model computes a 73.5% probability that the OC abundance is low ([OC] < 2.5 wt% TOC), a 17.8% probability that it is moderately abundant in OC (2.5 wt% < [OC] < 10.0 wt%), and an 8.2 % probability that OC is high in abundance (10.0 wt% < [OC]). These predictions corroborate OC analyses that have been done on Martian soils using resource intensive combustion methods from Curiosity in which they determined the sediment to contain ~0.1% OC (Stern *et al*., 2022). Importantly, this calculation is only for the average composition of Mars regolith. In addition, we also applied our model to elemental abundances from the Mars Odyssey Orbiter (which uses gamma-ray spectroscopy rather than XRF to determine elemental abundances) to make predictions for OC abundance. Unlike PIXL, Mars Odyssey determines elemental abundances at a much lower resolution. This lower resolution allows it to cover broader areas of the Martian surface. Hahn *et al*., 2007 determined the elemental abundance average for three different aged soils/rocks including Noachian, Hesperian, and Amazonian. We use this data to determine the probability of OC abundance as a comparison to PIXL. Interestingly, we compute very similar probabilities to those determined from PIXL elemental abundances (Table 4). On average, we calculate an 80.3% probability that OC abundance is low, a 19.7% probability that OC abundance is moderate, and a 0% probability that OC abundance is high.

Table 4. Classification probability of Martian soils.

Instrument/Sample	High	Moderate	Low
PIXL	8.2	18.2	73.5
Odyssey (Noachian)	0	17.7	82.2
Odyssey (Hesperian)	0	17.8	82.3
Odyssey (Amazonian)	0	23.7	76.3

Despite the overall good performance of our model, it is important to note that we cannot assume that OC behaves exactly the same on Mars as it does on Earth. As such, our model serves as a basis for transfer learning which in this instance is an algorithm that is trained using Earth-based data sets after which that knowledge is "transferred" and applied to Mars (Theiling *et al*., 2022). An advantage to the transfer learning concept is that the vast amounts of datasets on Earth can serve as a starting point to eventually adapt algorithms for Mars, which is currently necessary considering the paucity of data on Mars compared to Earth (Theiling *et al*., 2022).

Potential Earth-Based Application

Although our model was constructed with the goal of life detection beyond Earth, there are potential applications for Earth-based analysis. For instance, carbon flux through time in lake and marine sediments is an active area of research (Sarkar *et al*., 2023; Lee *et al*., 2019; Zhang *et al*., 2018; Leach *et al*., 2008). More specifically, high quality, continuous, and high-resolution records are sought after (Leach *et al*., 2008). One of the limitations to continuous and high-resolution studies is that due to the time intensive nature of OC analysis, it is more challenging to achieve the same level of resolution as elemental abundances from XRF, ultimately resulting in a sparsity of OC data (Lee *et al*., 2019; Zhang *et al*., 2018). Our model potentially provides a useful screening tool to evaluate where to target biomarker work and/or assess necessary analytical amounts for TOC and biomarker work.

Our model has the potential to be especially useful for hypersaline lakes as they are widely distributed across the globe; however, they are often neglected in climate models as they are assumed to be smaller in number and sparsely distributed (Sarkar *et al.,* 2023; Marce *et al*., 2019). Contrary to this belief, hypersaline lakes contribute significantly to the global lake volume (85 x 10^3 km^3 to the total 190 x 10^3 km^3; Sarkar *et al*., 2023; Williams, 2002). This assumption has created a blind spot in the global carbon cycle, which is estimated by the amount of primary production and long-term sediment storage of OC (Marce *et al*., 2019). Considering this, hypersaline environments have a great potential to influence the global carbon budget and ultimately global climate. Given their potential to strongly influence global climate, it is crucial that we identify a rapid way to identify the amount of potential carbon flux into the atmosphere.

Conclusion

This work provides a proof-of-concept application for leveraging machine learning models to aid in life detection efforts. More specifically, we use data that is relatively easy to collect (XRF-derived elemental abundances) to make predictions about data that is more resource intensive (OC analysis). All models constructed in this study showed good performance (>80% accuracy) for predicting OC abundance in lake sediments from XRF data; however, the RF algorithm outperformed the others for both classification and regression predictions. Warren-Rhodes *et al*., 2023 showed that a random biosignature search yielded only a 9.2% probability of detecting biosignatures. As such, our models improve the probability of detecting OC with varying levels by >70% compared to random searches. Ultimately, our model has the potential to be used to make predictions about OC abundance prior to sample analysis on high-resolution but narrow field of view Martian rovers to save time and resources for life detection. Additionally, it can be used to create a prediction map at the global scale using orbiters such as Mars Odyssey which are low-resolution but have a broad field of view.

Despite the overall good performance of our model, we recognize that a model is only as strong as its training set, and this model is most adapted to hypersaline, lacustrine sediments. Further efforts for broader classification tools will require more data from diverse environments and older sediments or rocks. Additionally, our model lacks an abundance of samples that are very lean in OC (<1 wt% [OC]), thus, it can also benefit from an addition of more samples lean in organic material. While the models constructed in this work have showed good performance, future work will focus on adding more data from diverse environments as described above in addition to extending the explanatory variables to mineralogy. Our work is expected to be critical for better understanding which samples and/or sites on Mars are most likely to harbor an abundance of OC and will further propel future life detection missions.

Open Research

All data and Python scripts written for data analysis, visualization, and model construction are hosted at https://zenodo.org/records/10433398 and can be cited as "Nichols, Floyd. (2024). FloydNichols97/OMAP: OMAP. Release (1.0) [Dataset]. Zenodo. https://zenodo.org/records/10433398"

Acknowledgements

This work was supported by a grant from NASA Exobiology to AP, CEC, and MRO (NNH17ZDA001N-EXO). This work would not have been possible without the aid in sample collection in British Columbia from Hannah Dion-Kirschner. Additionally, the Mel3 sediment core collection was made possible by the people of Greenland for access to the field site, Bailey Nash, Yarrow Axford, Peter Puleo, Grace Schellinger, Aidan Burdick, Aaron Hartz, Polarfield Services, US Air National Guard, Sermeq Helicopters, and Blue Ice Adventures and the NSF grant that funded MEL3 core collection and analysis (NSF Polar Program award #2002515). Additionally, Caleb Scharf, Ph.D. and Mary Beth Wilhelm, Ph.D. provided valuable insights into improving machine learning approaches as well as interpretation of sediment geochemistry behavior.

References

Acito, F. (2023). k Nearest Neighbors. In: Predictive Analytics with KNIME. Springer, Cham.

Allwood, A., Clark, B., Flannery, D., Hurowitz, J., Wade, L., Elam, T., ... & Knowles, E. (2015, March). Texture-specific elemental analysis of rocks and soils with PIXL: The Planetary Instrument for X-ray Lithochemistry on Mars 2020. In *2015 IEEE Aerospace Conference* (pp. 1-13). IEEE.

Aubrey, A., Cleaves, H. J., Chalmers, J. H., Skelley, A. M., Mathies, R. A., Grunthaner, F. J., ... & Bada, J. L. (2006). Sulfate minerals and organic compounds on Mars. *Geology*, *34*(5), 357-360.

Auret, L., & Aldrich, C. (2012). Interpretation of nonlinear relationships between process variables by use of random forests. *Minerals Engineering*, *35*, 27-42.

Belle, V., and Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. *Front. Big Data* 4:688969

Bhartia, R., Beegle, L. W., DeFlores, L., Abbey, W., Razzell Hollis, J., Uckert, K., ... & Zan, J. (2021). Perseverance's scanning habitable environments with Raman and luminescence for organics and chemicals (SHERLOC) investigation. Space Science Reviews, 217(4), 58.

Biau, G., and Scornet, E. (2016). A Random Forest Guided Tour. *TEST* 25, 197-227.

El Bilali, L., Rasmussen, P. E., Hall, G. E. M., & Fortin, D. (2002). Role of sediment composition in trace metal distribution in lake sediments. *Applied Geochemistry*, *17*(9), 1171-1181.

Bone, S. E., Dynes, J. J., Cliff, J., & Bargar, J. R. (2017). Uranium (IV) adsorption by natural organic matter in anoxic sediments. *Proceedings of the National Academy of Sciences*, *114*(4), 711-716.

Bone, S. E., Cliff, J., Weaver, K., Takacs, C. J., Roycroft, S., Fendorf, S., & Bargar, J. R. (2019). Complexation by organic matter controls uranium mobility in anoxic sediments. *Environmental Science & Technology*, *54*(3), 1493-1502.

Brownlee, J. (2018). Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions. *Machine Learning Mastery*.

Burdige, D. J. (2007). Preservation of organic matter in marine sediments: controls, mechanisms, and an imbalance in sediment organic carbon budgets? *Chemical reviews*, *107*(2), 467-485.

Chauhan, V. K., Dahiya, K., & Sharma, A. (2019). Problem formulations and solvers in linear SVM: a review. *Artificial Intelligence Review*, *52*(2), 803-855.

Christian, J.R., VanBommel, S.J., Elam, W.T., Ganly, B., Hurowitz, J.A., Heirwegh, C.M., Allwood, A.C., Clark, B.C., Kizovski, T.V. and Knight, A.L. (2023). Statistical characterization of PIXL trace element detection limits. *Acta Astronautica 212*, 534-540.

Cifuentes, G. R., Jimenez-Millan, J., Quevedo, C. P., Galvez, A., Castellanos-Rozo, J., & Jimenez-Espinosa, R. (2021). Trace element fixation in sediments rich in organic matter from a saline lake in tropical latitude with hydrothermal inputs (Sochagota Lake, Colombia): The role of bacterial communities. *Science of The Total Environment*, *762*, 143113.

Cortes, C., and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning* 20, 273-297.

Cronin, N.J. (2021). Using Deep Neural Networks for Kinematic Analysis: Challenges and Opportunities.

Cumberland, S. A., Douglas, G., Grice, K., & Moreau, J. W. (2016). Uranium mobility in organic matter-rich sediments: A review of geological and geochemical processes. *Earth-Science Reviews*, *159*, 160-185.

Dong, L., Li, X., & Xie, G. (2014). Nonlinear methodologies for identifying seismic event and nuclear explosion using random forest, support vector machine, and naive Bayes classification. In *Abstract and applied analysis* (Vol. 2014, pp. 1-8). Hindawi Limited.

Duarte, C. M., Prairie, Y. T., Montes, C., Cole, J. J., Striegl, R., Melack, J., & Downing, J. A. (2008). CO_2 emissions from saline lakes: A global estimate of a surprisingly large flux. *Journal of Geophysical Research: Biogeosciences*, *113*(G4).

Evans, G., Augustinus, P., Gadd, P., Zawadzki, A., & Ditchfield, A. (2019). A multi-proxy μ-XRF inferred lake sediment record of environmental change spanning the last ca. 2230 years from Lake Kanono, Northland, New Zealand. *Quaternary Science Reviews*, *225*, 106000.

Fuller, K. M., Fox, A. L., Jacoby, C. A., & Trefry, J. H. (2021). Biological Abundance and Diversity in Organic-Rich Sediments From a Florida Barrier Island Lagoon. *Frontiers in Marine Science*, *8*, 768083.

Fox, P. M., Nico, P. S., Tfaily, M. M., Heckman, K., & Davis, J. A. (2017). Characterization of natural organic matter in low-carbon sediments: Extraction and analytical approaches. *Organic Geochemistry*, *114*, 12-22.

Frings, P. J., Clymans, W., Jeppesen, E., Lauridsen, T. L., Struyf, E., & Conley, D. J. (2014). Lack of steady-state in the global biogeochemical Si cycle: emerging evidence from lake Si sequestration. *Biogeochemistry*, *117*, 255-277.

Hahn, B. C., McLennan, S. M., Taylor, G. J., Boynton, W. V., Dohm, J. M., Finch, M. J., ... & Williams, R. M. (2007). Mars Odyssey Gamma Ray Spectrometer elemental abundances and apparent relative surface age: Implications for Martian crustal evolution. Journal of Geophysical Research: Planets, 112(E3).

Hedges, J.I., Baldock, J.A., Gelinas, Y., Lee, C., Peterson, M., and Wakeham, S. (2001). Evidence for Non-Selective Preservation of Organic Matter in Sinking Marine Particles. *Nature* 409, 801-804.

Hemingway, J.D., Rothman, D.H., Grant, K.E., Rosengard, S.Z., Eglinton, T.I., Derry, L.A., and Galy, V.V. (2019). *Nature* 570, 228-231.

Jacq, K., Perrette, Y., Fanget, B., Sabatier, P., Coquin, D., Martinez-Lamas, R., ... & Arnaud, F. (2019). High-resolution prediction of organic matter concentration with hyperspectral imaging on a sediment core. Science of the Total Environment, 663, 236-244.

Jellison, R., Anderson, R. F., Melack, J. M., & Heil, D. (1996). Organic matter accumulation in sediments of hypersaline Mono Lake during a period of changing salinity. *Limnology and Oceanography*, *41*(7), 1539-1544.

Jolliffe, I.T., and Cadima, J. (2016). Principal Component Analysis: A Review and Recent Developments. *Phil. Trans. R. Soc. A.* 374:20150202.

Jones, B. F., & Deocampo, D. M. (2003). Geochemistry of saline lakes. *Treatise on geochemistry*, *5*, 605.

Kim, B, Khanna, R. and Koyejo, O.O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. Advances in Neural Information Processing Systems.

Leach, C. J., Wagner, T., Jones, M., Juggins, S., & Stevenson, A. C. (2008). Rapid determination of total organic carbon concentration in marine sediments using Fourier transform near-infrared spectroscopy (FT-NIRS). Organic Geochemistry, 39(8), 910-914.

Lee, T. R., Wood, W. T., & Phrampus, B. J. (2019). A machine learning (kNN) approach to predicting global seafloor total organic carbon. Global Biogeochemical Cycles, 33, 37–46.

Lipton, Z. C. (2016). The Mythos of Model Interpretability. *arXiv preprint arXiv*:1606.03490.

Maalouf, M. (2011). Logistic Regression in Data Analysis: An Overview. *Int. J. Data Analysis Techniques and Strategies* 3(3), 281-299.

Makinen, J., & Pajunen, H. (2005). Correlation of carbon with acid-soluble elements in Finnish lake sediments: two opposite composition trends. *Geochemistry: Exploration, Environment, Analysis*, *5*(2), 169-181.

Mahaffy, P. R., Webster, C. R., Cabane, M., Conrad, P. G., Coll, P., Atreya, S. K., ... & Mumm, E. (2012). The sample analysis at Mars investigation and instrument suite. Space Science Reviews, 170, 401-478.

Mahesh, B. (2018). Machine Learning Algorithms - A Review. *IJSR* 9(1), 381-386.

Marcé, R., Obrador, B., Gómez-Gener, L., Catalán, N., Koschorreck, M., Arce, M. I., ... & von Schiller, D. (2019). Emissions from dry inland waters are a blind spot in the global carbon cycle. *Earth-science reviews*, *188*, 240-248.

Maurice, S., Wiens, R. C., Saccoccio, M., Barraclough, B., Gasnault, O., Forni, O., ... & Vaniman, D. (2012). The ChemCam instrument suite on the Mars Science Laboratory (MSL) rover: Science objectives and mast unit description. Space science reviews, 170, 95-166.

McKaig, J.M., MinGyu, K., and Carr, C.E. (2023). Translation as a Biosignature. *bioRxiv*.

Molnar, C. (2019). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Available Online. https://christophm.github.io/interpretable-ml-book/ (accessed on 1 September 2023).

Nichols, F., Pontefract, A., Dion-Kirschner, H., Masterson, A. L., & Osburn, M. R. (2023). Lipid Biosignatures From SO_4-Rich Hypersaline Lakes of the Cariboo Plateau. *Journal of Geophysical Research: Biogeosciences*, *128* (10), e2023JG007480.

Nichols, F. (2023). Organic Matter Abundance Predictor (OMAP). https://om-prediction-app.streamlit.app/

Nichols, Floyd. (2024). FloydNichols97/OMAP: OMAP. Release (1.0) [Dataset]. Zenodo. https://zenodo.org/records/10433398

Pal, M., and Mather, M. (2005). Support Vector Machines for Classification in Remote Sensing. *International Journal of Remote Sensing* 26(5), 1007-1011.

Peaple, M., *et al*. (2021). Identifying Plant Wax Inputs in Lake Sediments Using Machine Learning. *Organic Geochemistry* 156, 104222.

Pedregosa, F., *et al*. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12, 2825-2830.

Platzer, A. (2013). Visualization of SNPs with t-SNE. *PLOS ONE* 8(2).

Probst, P., Wright, M.N., and Boulesteix, A.L. (2018). Hyperparameters and Tuning Strategies for Random Forest. *WIREs Data Mining and Knowl. Discov.*

Puleo, P. J., Axford, Y., McFarlin, J. M., Curry, B. B., Barklage, M., & Osburn, M. R. (2020). Late glacial and Holocene paleoenvironments in the midcontinent United States, inferred from Geneva Lake leaf wax, ostracode valve, and bulk sediment chemistry. *Quaternary Science Reviews*, *241*, 106384.

Rothwell, R. G., & Croudace, I. W. (2015). Twenty years of XRF core scanning marine sediments: What do geochemical proxies tell us? *Micro-XRF studies of sediment cores: Applications of a non-destructive tool for the environmental sciences*, 25-102.

Sarkar, S., *et al*. (2023). Lake Desiccation Drives Carbon and Nitrogen Biogeochemistry of a Sub-Tropical Hypersaline Lake. *Hydrobiologia*.

Schimmelmann, A., Albertino, A., Sauer, P. E., Qi, H., Molinie, R., & Mesnard, F. (2009). Nicotine, acetanilide and urea multi-level 2H-, 13C-and 15N-abundance reference materials for continuous-flow isotope ratio mass spectrometry. Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry, 23(22), 3513-3521.

Shea, C.J., Steinman, B.A., Brown, E.T., and Schreiner, K.M. (2022). A Multi-Proxy Lake-Sediment Record of Middle Through Late Holocene Hydroclimate Change in Southern British Columbia, Canada. *J Paleolimnol* 67, 163-182.

So, R. T., Blair, N. E., Masterson, A. L. (2020). Carbonate mineral identification and quantification in sediment matrices using diffuse reflectance infrared Fourier transform spectroscopy. *Environmental Chemistry Letters*, 18, 1725-1730.

Stefanowicz, A. M., Kapusta, P., Zubek, S., Stanek, M., & Woch, M. W. (2020). Soil organic matter prevails over heavy metal pollution and vegetation as a factor shaping soil microbial communities at historical Zn–Pb mining sites. *Chemosphere*, *240*, 124922.

Stern, J.C., Malespin, C.A., Eigenbrode, J.L., Webster, C.R., Flesch, G., Franz, H.B., Graham, H.V., House, C.H., Sutter, B., Archer Jr, P.D. and Hofmann, A.E. (2022). Organic carbon concentrations in 3.5-billion-year-old lacustrine mudstones of Mars. *Proceedings of the National Academy of Sciences*, 119(27).

Taunk, K., De, S., Verma, S., and Swetapadma, A. (2019). A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. *International Conference on Intelligent Computing and Control Systems*.

Taylor, G. J. (2013). The bulk composition of Mars. Geochemistry, 73(4), 401-420.

Tien, P. L., & Waugh, T. C. (1970). Epsomite and hexahydrite from an underground storage area, Atchinson, Kansas. *Bulletin Kansas Geological Survey*, *199*, 3-8.

Uffink, J. (1995). Can the maximum entropy principle be explained as a consistency requirement?. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 26(3), 223-261.

van der Maaten, L., and Hinton, G. (2008). Visualizing Data Using t-SNE. *Journal of Machine Learning Research* 9, 2579-2605.

Wang, P., Du, Y., Yu, W., Algeo, T. J., Zhou, Q., Xu, Y., ... & Pan, W. (2020). The chemical index of alteration (CIA) as a proxy for climate change during glacial-interglacial transitions in Earth history. Earth-Science Reviews, 201, 103032.

Warren-Rhodes, K., Cabrol, N. A., Phillips, M., Tebes-Cayo, C., Kalaitzis, F., Ayma, D., ... & SETI Institute NAI Team. (2023). Orbit-to-ground framework to decode and predict biosignature patterns in terrestrial analogues. Nature Astronomy, 7(4), 406-422.

Westrich, J. T., & Berner, R. A. (1984). The role of sedimentary organic matter in bacterial sulfate reduction: The G model tested 1. *Limnology and oceanography*, *29*(2), 236-249.

Williams, W. D. (2002). Environmental threats to salt lakes and the likely status of inland saline ecosystems in 2025. *Environmental conservation*, *29* (2), 154-167.

Zhang, F., Xue, B., Yao, S., & Gui, Z. (2018). Organic carbon burial from multi-core records in Hulun Lake, the largest lake in northern China. *Quaternary International*, *475*, 80-90.

Zhang, X., Zhang, H., Chang, F., Ashraf, U., Peng, W., Wu, H., ... & Duan, L. (2020). Application of corrected methods for high-resolution XRF core scanning elements in lake sediments. *Applied Sciences*, *10*(22), 8012.