Environmental Science & Technology

pubs.acs.org/est

Article

Exploring the Complexities of Dissolved Organic Matter Photochemistry from the Molecular Level by Using Machine Learning Approaches

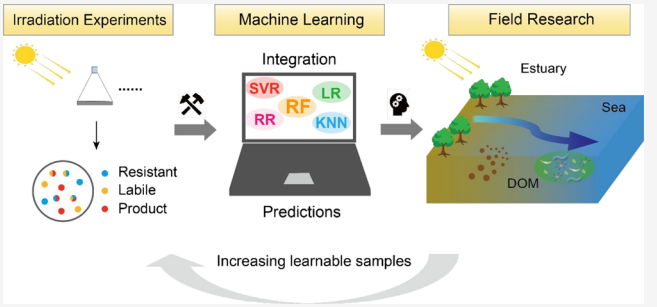
Chen Zhao,[$] Xinyue Xu,[$] Hongmei Chen, Fengwen Wang, Penghui Li, Chen He, Quan Shi, Yuanbi Yi, Xiaomeng Li, Siliang Li, and Ding He*

Cite This: https://doi.org/10.1021/acs.est.3c00199

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Dissolved organic matter (DOM) sustains a substantial part of the organic matter transported seaward, where photochemical reactions significantly affect its transformation and fate. The irradiation experiments can provide valuable information on the photochemical reactivity (photolabile, photoresistant, and photoproduct) of molecules. However, the inconsistency of the fate of irradiated molecules among different experiments curtailed our understanding of the roles the photochemical reactions have played, which cannot be properly addressed by traditional approaches. Here, we conducted irradiation experiments for samples from two large estuaries in China. Molecules that occurred in irradiation experiments were characterized by the Fourier transform ion cyclotron resonance mass spectrometry and assigned probabilistic labels to define their photochemical reactivity. These molecules with probabilistic labels were used to construct a learning database for establishing a suitable machine learning (ML) model. We further applied our well-trained ML model to "unmatched" (i.e., not detected in our irradiation experiments) molecules from five estuaries worldwide, to predict their photochemical reactivity. Results showed that numerous molecules with strong photolability can be captured solely by the ML model. Moreover, comparing DOM photochemical reactivity in five estuaries revealed that the riverine DOM chemistry largely determines their subsequent photochemical transformation. We offer an expandable and renewable approach based on ML to compatibly integrate existing irradiation experiments and shed insight into DOM transformation and degradation processes.



KEYWORDS: *dissolved organic matter, machine learning, molecular composition, photochemistry, estuarine carbon cycling*

Downloaded via HONG KONG UNIV SCIENCE TECHLGY on August 22, 2023 at 16:16:22 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

1. INTRODUCTION

Dissolved organic matter (DOM) is one of the largest reactive carbon pools on the earth, and its chemical composition and reactivity are closely associated with aquatic carbon and nutrient cycling,[1,2] trace-element transport,[3] microbial metabolism,[4] and reactions with environmental contaminants.[5] Photochemical reactions (photoproduction and photodegradation) are essential components altering DOM chemistry, by directly mineralizing it to CO_2 or indirectly inducing its biogeochemical function changes for subsequent microbial mediation.[6−9] The state-of-the-art ultrahigh resolution mass spectrometry, Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS), can unprecedentedly provide thousands of molecular formulas (MFs) within one DOM sample and provide an opportunity to directly link the molecular chemical composition and photochemical reactivity.[10,11] Previous studies demonstrated significant DOM compositional conversions after photoirradiation,[12,13] and the MFs linked to photochemical processing are operationally classified as photoresistant, photolabile, and photoproduct types according to their occurrence before and after irradiation experiments.[14,15]

To study the potential photochemical reactivity of samples without employing irradiation experiments, a favorable practice is to compare the MFs occurring in samples from various aquatic ecosystems with the classified MFs acquired from previously performed irradiation experiments.[16−19] Despite the success of this "molecular matching" approach in evaluating the photochemical quality of the samples, many concerns remain. First, considering the complexity of the DOM transformation, the relationship between the classes of photochemistry related MFs and their chemical composition is complicated and probably nonlinear, which cannot be simply

Special Issue: Data Science for Advancing Environmental Science, Engineering, and Technology

Received: January 8, 2023
Revised: May 10, 2023
Accepted: May 12, 2023

© XXXX American Chemical Society

elucidated through the molecular matching approach. Different MFs may share comparable photochemical reactivity for their inherently similar chemical compositions. Significant amounts of MFs with potential photochemical reactivity would be neglected through a simple matching approach, i.e., MFs detected in the field but not in limited irradiation experiments. Second, the initial samples supporting the irradiation experiments were collected from varying aquatic ecosystems (e.g., rivers, estuaries, and open oceans). It will be risky to directly match MFs derived from samples with different background information when depending on the result from individual experiment. More importantly, there will be "label conflicts" when considering more than one irradiation experiment, which means inconsistency of the assignments of MFs into specific photochemical types. For instance, a MF assigned as a photoresistant molecule in one irradiation experiment can be alternatively assigned as a photolabile or photoproduct MF in another, leading to an ambiguous understanding of the photochemical reactivity of specific MFs. As such, molecular matching alone could likely lead to biased estimates of the photochemical reactivity of specific MFs among different samples (e.g., with a different number of "un-matched" formulas). It is urgent to reconcile the above predicament and contribute to the broader applications of existing photochemical irradiation experiment results.

In contrast to the "matching" approach to deal with complex molecular composition data, machine learning (ML) methodologies can learn from large, complex, and multidimensional data to develop predictive models and have proven to be promising tools for handling nonlinear correlations between the learnable features and target labels. Its goal is to continuously extract internal knowledge from the data during the learning process and make the model resemble the human thought process, thereby enhancing the predictive power.[20] Multiple instances have demonstrated that ML techniques can be applied to earth and environmental science.[21−23] As the amount of data collected has increased, it has become increasingly difficult to manually analyze and extract information from these massive data sets. Therefore, it is crucial to develop an intelligent system capable of acquiring and interpreting vast amounts of data more efficiently. This does not imply that traditional expert opinions will be abandoned but rather that the collected data will be better understood, and new perspectives will be offered. Intelligent systems can help us discover unseen relationships between features to better comprehend the outcomes of photochemical experiments conducted thus far.

To overcome the above limitations of the molecular matching approach, we explored the applicability of multiple ML techniques and established well-trained random forest regression models to investigate the underlying relationship between the chemical composition and photochemical reactivity of MFs. Four photoirradiation experiments covering freshwater and seawater samples from two large estuaries in China, the Yangtze River Estuary (YRE) and Pearl River Estuary (PRE), were conducted. To prevent the "label conflicts" problem, instead of assigning MFs a "hard label" of a specific type (photoresistant, photolabile, or photoproduct), we calculated the ratio of times where they occurred as one of three types to the total times where they were detected in the four irradiation experiments to obtain "soft labels" with probability. The photochemical reactivity of MFs which could not be matched in five estuaries worldwide (YRE, PRE, Delaware Estuary, Daliao River Estuary, and Jiulong River Estuary) were predicted by our well-trained models, and the photochemical specificity of DOM was further assessed within and across five estuaries.

By combining the benefits of molecular composition information and ML approaches, we aim to (i) offer an expandable approach to compatibly integrate existing irradiation experiments for their boarder applications; (ii) explore the MFs behaviors during the photochemical processing from a novel perspective, and (iii) semiquantitatively assess the spatial variations of DOM photochemical reactivity and its geochemical implications in estuarine environments.

2. SAMPLES AND ANALYTICAL METHODS

2.1. Molecular Formular Data Set Generated by FT-ICR MS. The samples for photoirradiation experiments were collected from two large estuaries (YRE and PRE) in July 2017. The water sample collecting details for the Daliao River Estuary (DLE, November 2016), the Delaware Estuary (DWE, August 2012), and the Jiulong River Estuary (JRE, July 2014 and May 2015) can be found from previous case research.[24−26] Two surface endmember samples obtained from each estuary (YRE-F, YRE-S, PRE-F, and PRE-S) were specially subjected to photoirradiation treatments (detailed in Supporting Information 1.1 and Figure S1). The irradiated samples were prefiltered with 0.2 μm precleaned polycarbonate membranes (Millipore). Although the microbial activities could not be completely prevented, previous biological incubation comparisons demonstrated only trivial changes in the bacterial abundance after filtering during the 14-day observation, implying that bacteria-dominated microbial activities did not serve as the primary factor controlling DOM transformation.[27,28] We suggest that constant monitoring of the bacterial abundance during photoirradiation would be better in future research.

The FT-ICR MS analysis followed our matured protocols (Supporting Information 1.2). Relative peak intensities were calculated based on the sum-normalized intensities of all assigned peaks in each sample. A total of 8686 unique MFs were detected before or after the four photoirradiation experiments and utilized to construct the ML model, which were set as the learning data set. We referred to them as "learned MFs" in this study if not specified otherwise. A total of 7590 unique MFs from five estuaries could not be matched by the detected MFs in irradiation experiments and set as the prediction data set. Their photochemical reactivity would be predicted after the development of the model. We referred to them as "predicted MFs" if not specified otherwise (Figure S2).

Multiple molecular parameters such as the elemental ratios (O/C, H/C, N/C, and S/C), modified aromatic index (AI_{mod}), equivalent double bond number (DBE), and nominal oxidation state of carbon (NOSC) were calculated based on previous literature.[8] The MFs identified were operationally classified into different compounds according to parameter ranges, including polycyclic condensed aromatics (PCAs), polyphenols, highly unsaturated compounds (HU), unsaturated aliphatic compounds (UA), peptides, and carboxyl-rich alicyclic molecules (CRAM;[29,30] detailed in Supporting Information 1.2).

2.2. Data Set Preparation for Machine Learning. The MFs were preliminarily classified as photoresistant, photolabile, and photoproduct MFs according to their occurrence before

https://doi.org/10.1021/acs.est.3c00199
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

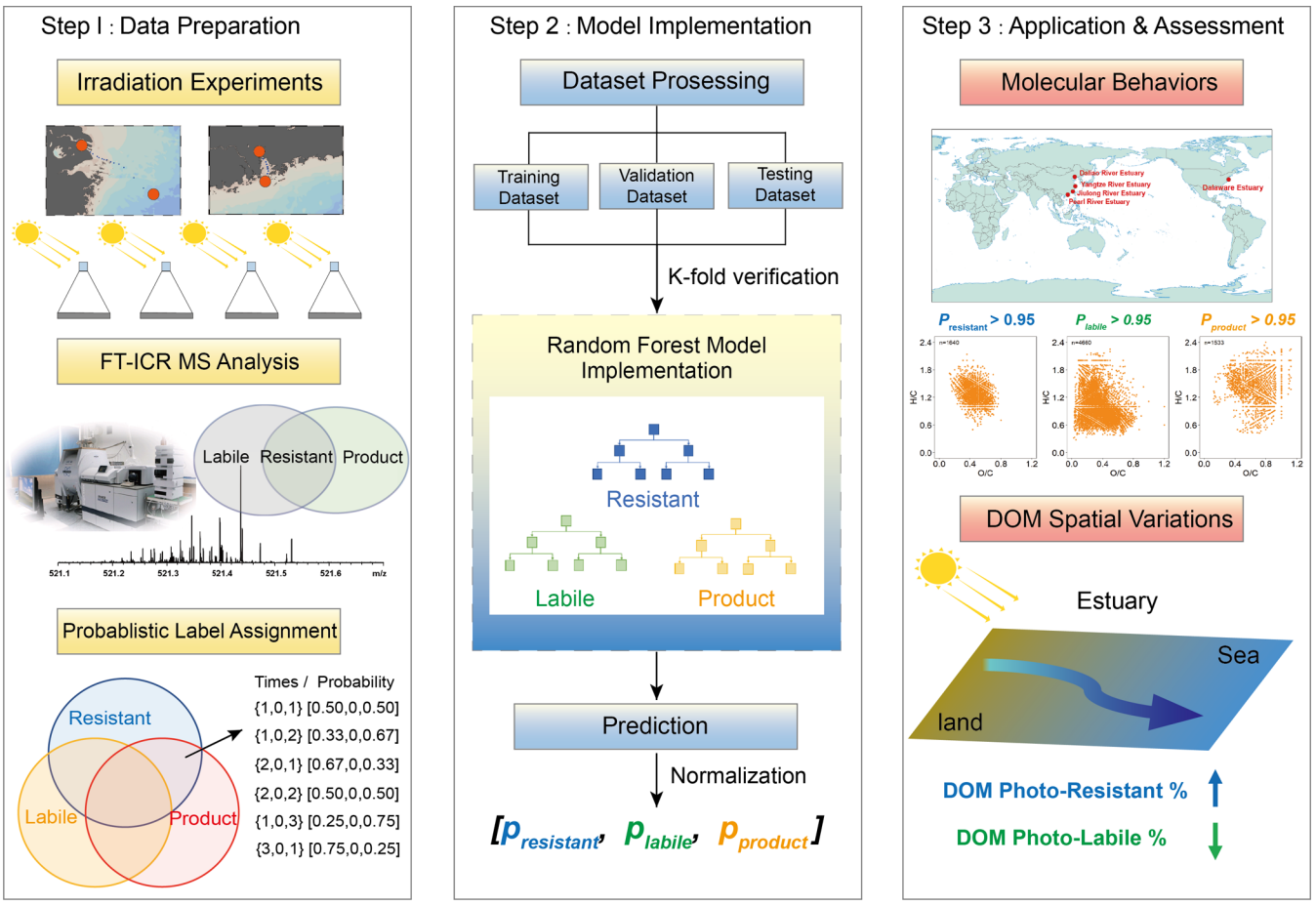


Figure 1. Flowchart of this research.

and after each irradiation experiment (detailed in Supporting Information 1.2 and Figure S3). In this step, it is noted that an MF can be assigned as more than one type. To overcome the "label conflict" problem, we alternatively assigned it a three-element vector label by calculating the probability that they occurred as three types (photoresistant, photolabile, and photoproduct) in the four irradiation experiments instead of assigning each MF a hard label (Figure S4). For instance, when an MF was classified into photoresistant, photolabile, and photoproduct types for two, one, and one times, respectively, its target label was denoted as [0.5, 0.25, 0.25]. Consequently, there would be seven potential labels assigned to learned MFs for each photochemical reactivity type: 0, 0.25, 0.33, 0.5, 0.67, 0.75, and 1 (Figure S4). Fifteen classically established parameters (the number of C, N, O, N, S atoms; O/C, H/C, N/C, S/C, AI_{mod}, DBE, DBE/C, DBE/O, and NOSC) depicting the fundamental molecular properties were selected for the training features. The selected 15 classic parameters cover the molecular information in different dimensions (Supporting Information 1.2), including the elemental composition (the number of atoms), van Krevelen (V−K) plot parameters (O/C and H/C), heteroatomic compounds composition (S/C and N/C), molecular mass (m/z), aromaticity degree (AI_{mod} and DBE), oxidation state (NOSC), and integrated information (DBE/C and DBE/O). The parameter ranges of learned MFs are shown in Table S1.

Briefly, $\mathbf{X} \subset \mathbb{R}_d$ (d = 15) is the input feature values; an instance could be represented as a vector of d feature values $\mathbf{x} = [x_1,...,x_d]$. For each instance $\mathbf{x}$, it has a target label $\mathbf{y} = [y_1,y_2,y_3]$ (in which $y_1 + y_2 + y_3 = 1$, where y_j represents photoresistant, photolabile, and photoproduct class, respectively). The training data contains $\mathbf{N}$ labeled molecules, where y^i is a probabilistic label of the ith instance; y^i_j is the probability of the ith instance belonging to the jth label.

2.3. Construction and Comparison of Multiple Machine Learning Models. Multilabel regression was selected to address this issue. It selected the same type of ML model, conducted the regression separately for each label, and then combined the results of multiple labels. The learning data set was divided into the training data set (80%) and the testing data set (20%). We employed sophisticated autosearching technique Hyperopt-Sklearn[31] to fine-tune our models. We conducted the hyperparameter search on five types of ML models and use three evaluation indices (R^2, MAE, MSE) to evaluate the performance of the models. Random forest (RF), linear regression (LR), K-nearest neighbors regression (KNN), ridge regression (RR), and support vector machine regression (SVR) were the five machine learning models employed (see Supporting Information 1.3 for details). See Figure 1 for the flowchart of our ML implementation process. Through the hyperparameter search (including K-fold verification, K = 5), the model with the highest score for each label was identified, and the regression results were obtained by predicting on our test data set. Each model was evaluated using three indicators, R^2, MAE, and RMSE (Supporting Information 1.3). We normalized (each probability was divided by the sum of the three types) the probabilities of three types of photochemical types at the end to ensure that the sum of the probabilities of

https://doi.org/10.1021/acs.est.3c00199
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

the output is equal to 1. We used the SHAP approach to evaluate the contributions of input features in constructing each model (see Supporting Information 1.3 for details).[32,33]

2.4. Estimation of the Photochemical Reactivity of DOM Samples. There would be two kinds of probabilistic patterns of MFs after completing the ML prediction. The probabilistic labels of 8686 learned MFs were discrete values derived from the prior knowledge of irradiation experiments. In contrast, the probabilistic labels of the 7590 predicted MFs were continuous values forecasted by the RF model, ranging from 0 to 1. After obtaining the probabilities of all MFs, we further calculated the relative intensity of each type in DOM samples to semiquantitatively assess their photochemical reactivity in case studies. To be specific, supposing that N MFs were detected in a sample, and for the ith MF, its relative intensity within a sample is I^i, and its probability label is $\mathbf{y}^i = [y_1^i, y_2^i, y_3^i]$, the relative intensity $I_{\text{photochemistry-type}}$ of this sample would be calculated as in the following equations:

$$I_{\text{Photoresistant}} = \sum_{i=1}^{N} y_1^i I^i$$

$$I_{\text{Photolabile}} = \sum_{i=1}^{N} y_2^i I^i$$

$$I_{\text{Photoproduct}} = \sum_{i=1}^{N} y_3^i I^i$$

And considering that our strategy to define the $\mathbf{y}$ labels, the sum of y_1, y_2, and y_3 for MFs in both the learning and prediction data set equals to 1, we can know that:

$$I_{\text{Photoresistant}} + I_{\text{Photolabile}} + I_{\text{Photoproduct}} = 1$$

3. RESULTS AND DISCUSSION

3.1. Model Performance. The performance of five ML models is shown in Figure S5 and Table S2. The performance of different models varies greatly. The nonlinear models (RF, SVR, and KNN) basically outperform linear models (LR and RR), suggesting the complex constraints of molecular composition to DOM photochemical reactivity. RF achieved the best results on the three photochemical reactivity types with the highest R^2 and lowest MAE and RMSE on the testing data set. Therefore, RF was selected as the final prediction model in this study. As an ensemble model involving a network of decision trees by a bootstrap technique, RF maintains the interpretability of the decision tree for the total multiple regressors, which have a high level of accuracy.[34] This increases the confidence of users in the model's decisions. It has significant benefits, including the ability to reduce the overfitting issue in the original decision tree, automatically process missing values in the data, and eliminate the need to normalize the data. The R^2 values for the photoresistant, photolabile, and photoproduct RF models were 0.81, 0.75, and 0.68, respectively.

3.2. Suitability and Sustainability of the Approach. Traditionally, machine learning tasks have been separated into supervised and unsupervised learning. Generally, labeled data tasks belong to supervised learning problems. Supervised learning can be subdivided into classification and regression tasks.[35] Supervised approaches have been applied to the MFs classification in drinking water reservoirs' DOM.[36] They first utilized the Spearman's rank correlation to link the molecular intensities and the environmental parameter (e.g., chlorophyll concentration or solar irradiation dose) to obtain the coefficients. By constructing the ML model where the features were also the molecular parameters, and the target labels were coefficients, the coefficients of MFs with insufficient environmental parameters were predicted, and the MFs were further classified into photochemical or microbial products based on their coefficients. This approach has strengths in distinguishing the impacts induced by photochemical or microbial degradation. Nevertheless, we must note that the correlation analyses, including the popular Pearson correlation or Spearman's rank correlation, may not suggest a direct causal link. Providing more professional prior knowledge can be more reasonable to investigate the behaviors of MFs during degradation processes. In this study, we acquired the photoresistant, photolabile, and photoproduct types on the basis of the irradiation experiments, which provides relatively direct evidence to further disentangle the complex photochemical reactivity of MFs and bulk DOM.

According to the prior knowledge of irradiation experiments, individual irradiation experiments can provide operationally unambiguous classification (photoresistant, photolabile, and photoproduct) of MF. The task to predict the photochemical reactivity of unmatched MFs is supposed to be the classification problem from an intuitive thought. However, multiple factors control the fate of MFs in irradiation experiments (e.g., irradiation dose, incubation days, initial DOM chemistry, and molecular interactions), suggesting that the "labels conflicts" problem will inevitably emerge when more experiments are conducted. One plausible solution is to learn the common MFs that reserved the same photochemical reactivity in all experiments instead. But we can observe from the results of four experiments that only five and 42 common MFs for photolabile and photoproduct types can be found (Figure S3), respectively, suggesting that extremely limited MFs are available to learn. Another option is to introduce the idea of the "probabilistic label,"[37,38] which differs from the conventional hard label in that it provides a probability for each target class. When the experiments are very limited, the multiclass classification (Supporting Information 1.3) based on the probabilistic labels for each class can address part of the concerns. For instance, we can assign the MFs into six classes and classify unknown MFs when only two experiments are conducted (Figure S6). Unfortunately, we can expect that the number of classes will explode when we conduct numerous experiments (e.g., for only four experiments, there will be at most 22 classes), and the number of MFs within some classes would be very limited, which will weaken the performance and robustness of the classification algorithms. Therefore, considering both the suitability and sustainability of the potential approaches, we believe that the hard-label classification protocols are not capable of effectively resolving the predictions based on the irradiation experiments.

To overcome the above difficulties and ensure the sustainability of the approach, we comprehensively took advantage of the ideas of the regression tasks and probabilistic labels. With a prerequisite of the probabilistic labels acquired from experiments, we predicted the probabilities of undetected MFs (not detected in experiments) for each type by regressing algorithms. Although only seven possible probabilistic labels are available from four experiments, the quantity of labels would increase with increasing prior knowledge (i.e., more irradiation experiments), meaning more accurate labels

https://doi.org/10.1021/acs.est.3c00199
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

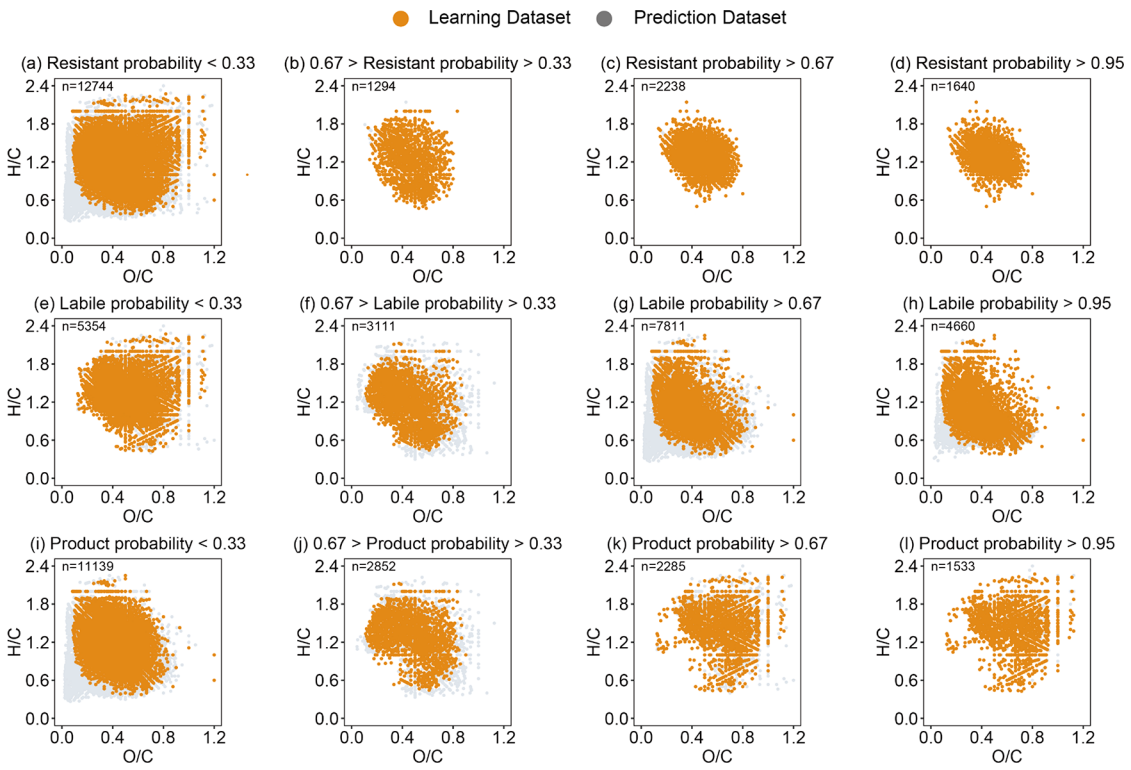


Figure 2. Comparisons between the learned and predicted molecular formulas with the increasing probability of three types of photoreactivity.

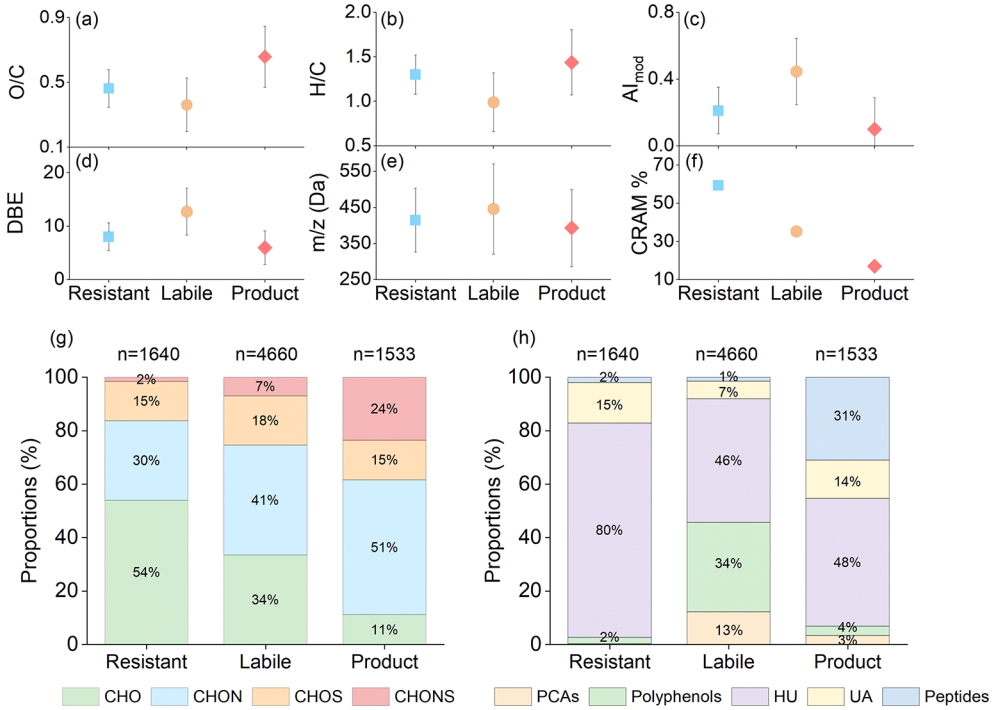


Figure 3. Comparison of three molecular pools with explicit (probability > 0.95) photochemical reactivity.

assigned to MFs, and thus more experiments now can be compatibly involved in our approach to promote the predicting capability.

3.3. The Photochemical Reactivity of MFs in Five Estuaries Worldwide. As specified in the Methods, the final output of the predicted MFs would be a three-element vector where each element represents the probability that MFs tend to be classified into a photochemical type (photoresistant, photolabile, and photoproduct). The threshold 0.33 and 0.67 probabilities were operationally selected as criteria to visualize the photochemical reactivity of MFs. We plotted all 16 276 MFs from five estuaries worldwide, including 8686 learned MFs and 7590 predicted MFs, into the *V*−*K* plots according to their probabilities of three types of photochemical reactivity (Figure 2 and Figure S7) to exhibit the transitions of MFs with increasing probability for each type. We further lent the

https://doi.org/10.1021/acs.est.3c00199
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

concept of the *95% confidence level* from the probability, and a *0.95 probability* threshold was operationally set to acquire the MFs with "explicitly" photochemical properties. Consequently, 1640, 4660, and 1533 MFs with explicit (probability > 0.95) photochemical properties of photoresistant, photolabile, and photoproduct were determined, respectively, and they made up three explicitly photochemical molecular pools.

The average molecular parameters and the proportions of the compound group in the three explicit pools were further calculated (Figure 3). The difference of parameters in three photochemical pools was tested using one-way ANOVA. Significant differences ($p < 0.01$) in O/C, H/C, AI_{mod}, DBE, and m/z were observed among three pools. The average AI_{mod} and DBE, indicative of aromatic and humification degree, were greater (0.45 and 12.73) in the photolabile pool than the other two pools, whereas the photoproducts were characterized as higher average H/C (1.44) and O/C (0.66) ratios among the three pools. The highest proportions (by number) of CHO, CHON, and CHOS compounds (30%, 51%, and 18%) were observed in the photoresistant, photoproduct, and photolabile pools, respectively. Regarding the compound groups, PCAs and polyphenols together accounted for nearly 50% of the photolabile pool, contrasting that less than 10% were found from the other two pools. The photoproduct pool had a significantly higher proportion of peptide-like groups (31%) than the photoresistant pool (2%) and photolabile pool (1%). Likewise, the HU represented over 75% of the photoresistant pool. The CRAM was also dominant (59%) in the photoresistant pool, compared with less than 40% in the others.

The DOM chemistry is expected to control its reactivity to photodegradation. For instance, for aromatic DOM, characterized by the molecular signatures of low O/C and H/C ratios,[39] its absorption of sunlight is likely responsible for the degradation of aromatic, high-molecular-weight compounds into aliphatic and lower-molecular-weight compounds.[39−41] The fate of the irradiated DOM can be either completely oxidized to CO_2 or partially oxidized to stimulate bacterial respiration,[40,42,43] where the partial processing can incorporate the oxygen into DOM through carboxylation with the conversion of ketone and aldehyde to carboxyl and consequently increase its O/C ratio.[44] Analysis of three explicitly photochemical molecular pools (probability > 0.95) indicated that explicitly photolabile MFs exhibited higher aromatic and humifaction degree (Figure 3), which is consistent with previous irradiation-based studies,[14,15,45] also suggesting that the partially oxidized processes are probably predominant in estuarine ecosystems. Moreover, the photolabile pool had the highest CHOS% among the three pools (Figure 3 and Figure 4), supporting its facile photochemical degradation as reported from the deep sea,[14,46] porewater,[18] and acid mine drainage,[47] with the potential products of climate-impacting gases carbonyl sulfide, dimethyl sulfide, or methanesulfonic acid. Significant amounts of N-containing compounds produced after photoirradiation were also found in samples from both natural and anthropogenic sources and were interpreted by the photoinduced incorporation of dissolved organic nitrogen into DOM.[18,48] The photoresistant pool was dominated by the HU group (Figure 3) that is also known to be biologically recalcitrant to heterotrophic microbes.[30] Consistently, a substantial fraction of the CRAM found in the photoresistant pool, which is believed to be a significant component of the marine refractory dissolved organic carbon pool,[31] implies that most of the photochemically recalcitrant DOM can be simultaneously biologically recalcitrant.

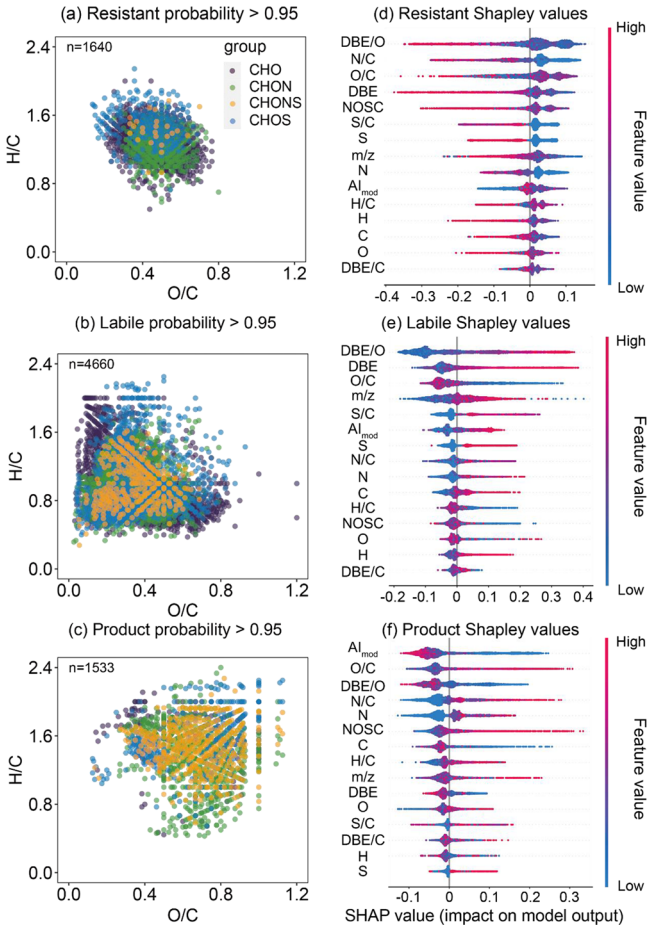


Figure 4. (a−c) The *V*−*K* plots for the molecules with explicit (probability > 0.95) photochemical reactivity. (d−f) The Shapley values of MFs with explicit photochemical reactivity (probability >0.95) for the photoresistant, photolabile, and photoproduct models.

We selected the MFs with probability > 0.95 to conduct SHAP analysis to eliminate potential effects of less strong data points (i.e., MFs with ambiguous photochemical reactivity). The SHAP revealed that the AI_{mod}, DBE, DBE/O, O/C, and N/C were tightly associated with the molecular photochemical reactivity upon the construction of models (Figure 4). As discussed above, the AI_{mod}, DBE, and O/C have been widely used to evaluate structural or oxidation degree conversions induced by photochemical reactions.[13−15] It is worthy to note that the partial photo-oxidation has an opposite impact on the aromaticity degree and oxidation degree of MFs, and thus the DBE/O (i.e., the ratio of aromaticity degree to the oxygen atom number) is probably more sensitive to the photoinduced effects, which exerted a more significant influence on the model construction (Figure 4). Meanwhile, the elemental composition N/C is a bulk and classic indicator to discriminate the terrestrial (lower N/C) and marine source (higher N/C) of DOM. The high importance of N/C confirms the hypothesis that the source of DOM controls its fate during irradiation. The Spearman's correlation analysis also showed that most of the important features of each model significantly ($p < 0.01$) correlated with model outputs in the testing data set (Table S3). Therefore, the importance of these molecular

https://doi.org/10.1021/acs.est.3c00199
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

features collectively reconfirms the reliability of our models from a geochemical view.

3.4. The Comparisons between the Learned and Predicted MFs. Significant differences among different photochemical molecular pools were observed when comparing the learned and predicted MFs (Figure 2 and Figure S8). Due to our strategy to assign labels, only seven possible probabilities for each type existed for the learned MFs. Therefore, the probability over 0.95 for the learned MFs means the probability was actually 1, suggesting that they only occurred as one type at least once in the four irradiation experiments. There were 1640, 2626, and 1340 learned MFs with a probability of 1 for the photoresistant, photolabile, and photoproduct types, respectively. Meanwhile, 2034 and 193 MFs were predicted as explicitly ($probability > 0.95$) photolabile or photoproduct types, respectively, whereas no predicted MFs possessed an over 0.95 probability for the photoresistant type. The statistical histograms compared the frequency distributions of probability between learned and predicted MFs, also indicating that predicted MFs had greater (e.g., over 0.5) probability for the photolabile type.

We compared the learned and predicted MFs with explicit photochemical reactivity (i.e., probability over 0.95) in terms of their molecular parameters and compound groups (Figure S9). Results showed that, for the explicitly photolabile pool, predicted MFs had higher average AI_{mod}, DBE, and m/z; higher PCAs % and polyphenols %; lower average H/C and O/C; and lower UA % and peptides % than learned MFs. Adversely, the predicted MFs in the explicit photoproduct pool were imprinted as higher average O/C and H/C; higher UA % and peptides %; lower average AI_{mod}, DBE, and m/z; and fewer PCAs % and polyphenols %. As a result, the molecular disparity between the photolabile pool and the photoproduct pool for learned MFs was strengthened for predicted MFs (Figure S9).

A high proportion (2034 out of 4660) of predicted photolabile MFs with a probability over 0.95 suggests that four irradiation experiments conducted in the YRE and PRE can only capture limited photolabile MFs. As shown in the lower left corner of $V-K$ plots, numerous MFs with lower O/C and H/C ratios were predicted as the photolabile type (Figure 3). Similar evidence was observed from the PCA and polyphenol compound groups with the primary origin of incomplete combustion of biomass and terrestrial inputs,[30,49,50] where the sum of their proportions in the predicted photolabile pool represents over 60% compared with the 30% in the learned photolabile pool. As such, the predictions of photolabile MFs are reasonable regarding the domain knowledge (e.g., previous irradiation experiments based study[14,15]). Further examination reveals that the highly aromatic MFs in the prediction data set were mainly from DWE and JRE (Table S4), where corresponding samples have a higher photolabile relative intensity (Figure S10). Therefore, the heterogeneity of the photolabile MFs among estuaries is likely due to varying land-use situations, upstream inputs, and anthropogenic activities,[51,52] which also largely lead to the strengthened disparity between the photolabile and photoproduct pool for predicted MFs. Meanwhile, we would expect an underestimation of photolabile relative intensity if only directly matching MFs from an irradiation experiment to other estuaries.

By contrast, the presence of few or even no predicted MFs into explicitly photoproduct or photoresistant pool indicates the relative homogeneity of these two kinds of MFs in estuarine environments. Compared with photolabile MFs, photoresistant and photolabile MFs could be easily captured by our four irradiation experiments for YRE and PRE samples. In addition to the spatial distributions of explicitly predicted photolabile MFs, predicted photoproduct MFs were mainly from DLE and DWE (Table S4). Although samples from DWE, DLR, and JRE were not implemented for irradiation experiments, the photochemical reactivity of their MFs can be successfully predicted by our ML models.

In this study, the parameter range of a fraction of predicted MFs could be outside the range of learned MFs, while the parameter ranges of these out-of-range MFs are on the same order of magnitude as the learned MFs. We can use the range of learned MFs to define the model's applicability domain (Table S1). However, after carefully examining the predictions of out-of-range MFs by using well-recognized domain knowledge, it is revealed that their predictions are reasonable regarding the previous irradiation experiment-based research (details in Table S5). Therefore, we believe that the predictions for out-of-range MFs are reliable in this research, and we recommend that future applications of this approach also carefully examine the predictions of out-of-range predicted MFs to ensure validity.

3.5. Spatial Variations in Photochemical Reactivity of DOM among Estuaries. As defined in the methods, we calculated the relative intensity of three types of MFs within a sample, although MFs were no longer regarded to have unique photochemical reactivity due to the probabilistic labels. In the 102 samples from five river-dominated estuaries across the world, the average relative intensity of photoresistant, photolabile, and photoproduct MFs were 79 ± 10%, 14 ± 7%, and 7 ± 3%, respectively (Figure S10). Significant differences among estuaries were also observed. The average relative intensity of five estuaries varies from 67% to 89%, 7% to 22%, and 4% to 11% for the photoresistant, photolabile, and photoproduct types, respectively. The regression slope, intercept, r square, and p value are shown in Table S6, and the extrapolated relative intensities at salinity 35 were also calculated to evaluate the photochemical reactivity of DOM samples under nearby seawater environments.

Previous research preliminarily evaluated the spatial variability of photochemical reactions in the Amazon River plume by principal component analysis.[21] However, dimension reduction likely leads to the loss of original information, and the interpretation to principal components is commonly speculative. The results of principle component analysis are also difficult to directly apply when conducting cross-system research. By assigning MFs a probabilistic label, we can semiquantitatively assess the spatial variations of samples upon photochemical reactivity along the salinity gradient from the same estuary and among estuaries. The significantly ($p < 0.01$) negative correlations between the photolabile relative intensity and salinity prevailed for all five estuaries, suggesting that the photolability of DOM significantly decreases when the organic matter is transported seaward (Figure S10). This could be caused by both the source variations and the photochemical transformation. With the addition of seawater, the source of DOM would be gradually dominated by fresh in situ productions which have less photolability as reported in the Amazon River plume and the YRE.[21,53] Meanwhile, we plotted the correlations between the relative intensity of photochemical reactivity types and the water turbidity in the DWE

https://doi.org/10.1021/acs.est.3c00199
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

(Figure S11). Results showed that the relative intensity of photolabile MFs increased with turbidity while an inverse correlation was observed for the photoproduct and photoresistant type, supporting that photoinduced degradation will decrease the relative intensity of the photolabile type with the improving water transparency (decrease of turbidity) from the river to the coastal seas.

By contrast, the photoresistant relative intensity positively correlated with the salinity significantly ($p < 0.01$). Considering an explicitly photoresistant pool includes a large portion of CRAM (Figure 4); this fraction of DOM could be further transported to the open oceans and constitute a part of the long-term sequestrated organic carbon in the ocean interior.[1,54] Similarly, a large fraction of CRAM was also found in the photoresistant pool from the North Atlantic deep water (NADW) or saltmarsh porewater irradiation experiment,[14,18] which partly supports the carbon sink role of this fraction of photo- and bio-recalcitrant DOM. Compared with the photolabile and photoresistant types, the variations of photoproduct relative intensity along the salinity gradient are relatively small. The correlation between the latitudes of estuaries (assuming a gradient of photoirradiation intensity) and their photoproduct relative intensity was not found (Figures S1 and S10), hinting that other factors on the regionally spatial scale likely play a more significant role in the photoproducts. Moreover, the order of the relative intensity for the same type in five estuaries has little change from the freshwater to the seawater end member (Figure S10 and Table S6), suggesting that the initial chemical composition of the riverine end member largely determines their DOM spatial variations upon photochemical reactivity in estuaries. Although DOM must also experience other physical and biogeochemical transformation processes than photochemistry during transportation, the riverine inputs have major impacts on shaping the differences of photochemical reactivity among estuaries.

3.6. Limitations and Further Directions. In recent years, deep learning has garnered special attention within the field of machine learning. Typically, it needs to use the linear layer as the output layer of the model and output the probability of multiple labels using the activation function. Deep belief networks[55] have been utilized for multilabel classification, but research on regression tasks is rare. However, the black-box nature of the deep learning model makes it difficult for individuals to discern which features the model is acquiring. Moreover, in the absence of sufficient data for the deep learning, the under-fitting phenomenon likely will occur. Therefore, we opted to experiment with conventional machine learning models in this study. We expect further development of deep learning algorithms preserving more robustness to deal with the probabilistic outputs.

The generalization and specificity of the ML model compete to some extent, which largely depends on the selections and matching degree of the learning and prediction data set. In this study, we conducted irradiation experiments with samples collected from end members of estuaries with contrasting physical and biogeochemical conditions (i.e., freshwater and seawater) to cover the estuarine environment, achieving a better generalization capability to predict the MFs in other estuaries. On the other side, to get a better specificity, i.e., more precise constraint of MFs from a specific environment instead, we recommend that the potential practices consider keeping the consistency for the conditions under which the MFs are obtained, meaning that initial samples for the irradiation experiments and predicted samples are preferably collected from similar environments. In other words, although we could perform irradiation experiments in other aquatic environments (e.g., inland water, deep ocean), they are not suitable for this study, which was specially designed to elucidate the roles of photochemical processing in estuarine carbon cycling.

There are still some limitations in this study. First, we initially classified the MFs upon photochemical reactivity based on their occurrence before and after the irradiation experiments without considering the variations of their relative intensity, which also represent the photochemical reaction-induced impacts.[56] Second, although FT-ICR MS has the powerful resolution to identify MFs, the information on isomeric-level and structural conversions under photoirradiation cannot be probed and deserve further investigation by ion mobility mass spectrometry or tandem mass spectrometry.[57] Moreover, in addition to molecular composition, molecular interactions probably influence the fate of MFs during irradiation. Recent research is working on the molecular network analysis by using advanced algorithms, such as the "PMD-based reactomics,"[58] trying to estimate the molecular interactions.[59,60] However, to the best of our knowledge, current studies are still premature to provide well-established parameters describing the intermolecular interactions. We shall continue to work for it to make the model consider more aspects.

Even though the above limitations require further efforts, we have offered a compatible approach to integrate existing multiple irradiation experiments to overcome their inconsistency known as "label conflicts" and understand the photochemical reactivity of MFs from a novel perspective. We successfully predict the MFs which cannot be matched by limited irradiation experiments and further provide insight into the photodegradation processes of DOM. Moreover, the ML model can be further updated with the addition of more irradiation experiments, which will also increase the accuracy of the prediction results. The ML model will be incorporated into a development platform (www.dom-dream.com) for serving DOM researchers. We believe this approach can be expanded to biological incubation or adsorption/desorption experiments which also suffer from the "label conflicts" problem with increasing experiments.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.3c00199.

Study site and water chemistry details; experimental design details; quality control of FT-ICR MS; data set combination description; additional descriptions of machine learning algorithms; model performances in figures and tables; further analysis of predicted results; and additional references (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Ding He − *Department of Ocean Science and Center for Ocean Research in Hong Kong and Macau, The Hong Kong University of Science and Technology, Hong Kong 999077, China; State Key Laboratory of Marine Pollution, City*

https://doi.org/10.1021/acs.est.3c00199
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

University of Hong Kong, Hong Kong 999077, China; orcid.org/0000-0001-9620-6115; Email: dinghe@ust.hk

Authors

Chen Zhao − *Department of Ocean Science and Center for Ocean Research in Hong Kong and Macau, The Hong Kong University of Science and Technology, Hong Kong 999077, China*

Xinyue Xu − *Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong 999077, China*

Hongmei Chen − *State Key Laboratory for Marine Environmental Science, Institute of Marine Microbes and Ecospheres, College of Ocean and Earth Sciences, College of the Environment and Ecology, Xiamen University, Xiamen 361000, China;* orcid.org/0000-0003-2460-838X

Fengwen Wang − *State Key Laboratory of Coal Mine Disaster Dynamics and Control, Department of Environmental Science, Chongqing University, Chongqing 400030, China*

Penghui Li − *School of Marine Sciences, Sun Yat-sen University, Zhuhai 519082, China; Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519082, China; Guangdong Provincial Key Laboratory of Marine Resources and Coastal Engineering, Zhuhai 519082, China*

Chen He − *State Key Laboratory of Heavy Oil Processing, China University of Petroleum, Beijing 102249, China;* orcid.org/0000-0002-7529-6366

Quan Shi − *State Key Laboratory of Heavy Oil Processing, China University of Petroleum, Beijing 102249, China;* orcid.org/0000-0002-1363-1237

Yuanbi Yi − *Department of Ocean Science and Center for Ocean Research in Hong Kong and Macau, The Hong Kong University of Science and Technology, Hong Kong 999077, China;* orcid.org/0000-0001-8420-8025

Xiaomeng Li − *Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong 999077, China*

Siliang Li − *Institute of Surface-Earth System Science, School of Earth System Science, Tianjin University, Tianjin 300072, China;* orcid.org/0000-0002-0295-9675

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.3c00199

Author Contributions

[$]Contributed equally to this work

Author Contributions

The manuscript was written through the contributions of all authors. All authors have given approval for the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research is supported by the National Natural Science Foundation of China (42222061, 42188102, 41973070), the Research Grants Council of Hong Kong (ECS26300822), the State Key Laboratory of Heavy Oil Processing, China University of Petroleum, and funding support from the Center for Ocean Research in Hong Kong and Macau (CORE). CORE is a joint research center for ocean research between QNLM and HKUST. C.Z. and X.X. appreciate the financial support of the Hong Kong Ph.D. Fellowship Scheme (HKPFS) from Hong Kong RGC.

■ REFERENCES

(1) Jiao, N.; Herndl, G. J.; Hansell, D. A.; Benner, R.; Kattner, G.; Wilhelm, S. W.; Kirchman, D. L.; Weinbauer, M. G.; Luo, T.; Chen, F.; Azam, F. Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. *Nat. Rev. Microbiol* **2010**, *8* (8), 593−9.

(2) Hedges, J. I. Global biogeochemical cycles: progress and problems. *Mar. Chem.* **1992**, *39* (1), 67−93.

(3) Yamashita, Y.; Jaffe, R. Characterizing the interactions between trace metals and dissolved organic matter using excitation-emission matrix and parallel factor analysis. *Environ. Sci. Technol.* **2008**, *42* (19), 7374−7379.

(4) Bai, L.; Cao, C.; Wang, C.; Xu, H.; Zhang, H.; Slaveykova, V. I.; Jiang, H. Toward quantitative understanding of the bioavailability of dissolved organic matter in freshwater lake during cyanobacteria blooming. *Environ. Sci. Technol.* **2017**, *51* (11), 6018−6026.

(5) Remucal, C. K. The role of indirect photochemical degradation in the environmental fate of pesticides: A review. *Environ. Sci. Processes Impacts* **2014**, *16* (4), 628−653.

(6) Ward, C. P.; Nalven, S. G.; Crump, B. C.; Kling, G. W.; Cory, R. M. Photochemical alteration of organic carbon draining permafrost soils shifts microbial metabolic pathways and stimulates respiration. *Nat. Commun.* **2017**, *8* (1), 772.

(7) Liao, Z.; Wang, Y.; Xie, K.; Xie, N.; Cai, X.; Zhou, L.; Yuan, Y. Photochemistry of dissolved organic matter in water from the Pearl river (China): Seasonal patterns and predictive modelling. *Water Res.* **2022**, *208*, 117875.

(8) Maizel, A. C.; Li, J.; Remucal, C. K. Relationships Between Dissolved Organic Matter Composition and Photochemistry in Lakes of Diverse Trophic Status. *Environ. Sci. Technol.* **2017**, *51* (17), 9624−9632.

(9) Harfmann, J. L.; Avery, G. B.; Rainey, H. D.; Mead, R. N.; Skrabal, S. A.; Kieber, R. J.; Felix, J. D.; Helms, J. R.; Podgorski, D. C. Composition and lability of photochemically released dissolved organic matter from resuspended estuarine sediments. *Org. Geochem.* **2021**, *151*, 104164.

(10) D'Andrilli, J.; Cooper, W. T.; Foreman, C. M.; Marshall, A. G. An ultrahigh-resolution mass spectrometry index to estimate natural organic matter lability. *Rapid Commun. Mass Spectrom.* **2015**, *29* (24), 2385−401.

(11) Koch, B. P.; Dittmar, T. From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. *Rapid Commun. Mass Spectrom.* **2006**, *20* (5), 926−932.

(12) Medeiros, P. M.; Seidel, M.; Powers, L. C.; Dittmar, T.; Hansell, D. A.; Miller, W. L. Dissolved organic matter composition and photochemical transformations in the northern North Pacific Ocean. *Geophys. Res. Lett.* **2015**, *42* (3), 863−870.

(13) Ward, C. P.; Cory, R. M. Complete and Partial Photo-oxidation of Dissolved Organic Matter Draining Permafrost Soils. *Environ. Sci. Technol.* **2016**, *50* (7), 3545−53.

(14) Stubbins, A.; Dittmar, T. Illuminating the deep: Molecular signatures of photochemical alteration of dissolved organic matter from North Atlantic Deep Water. *Mar. Chem.* **2015**, *177*, 318−324.

(15) Stubbins, A.; Spencer, R. G. M.; Chen, H.; Hatcher, P. G.; Mopper, K.; Hernes, P. J.; Mwamba, V. L.; Mangangu, A. M.; Wabakanghanzi, J. N.; Six, J. Illuminated darkness: Molecular signatures of Congo River dissolved organic matter and its photochemical alteration as revealed by ultrahigh precision mass spectrometry. *Limnol. Oceanogr.* **2010**, *55* (4), 1467−1477.

(16) Wang, K.; Li, P.; He, C.; Shi, Q.; He, D. Density currents affect the vertical evolution of dissolved organic matter chemistry in a large tributary of the Three Gorges Reservoir during the water-level rising period. *Water Res.* **2021**, *204*, 117609.

(17) Osterholz, H.; Kilgour, D. P. A.; Storey, D. S.; Lavik, G.; Ferdelman, T. G.; Niggemann, J.; Dittmar, T. Accumulation of DOC

in the South Pacific Subtropical Gyre from a molecular perspective. *Mar. Chem.* **2021**, *231*, 103955.
(18) Riekenberg, P. M.; Oakes, J. M.; Eyre, B. D. Shining Light on Priming in Euphotic Sediments: Nutrient Enrichment Stimulates Export of Stored Organic Matter. *Environ. Sci. Technol.* **2020**, *54* (18), 11165−11172.
(19) Medeiros, P. M.; Seidel, M.; Ward, N. D.; Carpenter, E. J.; Gomes, H. R.; Niggemann, J.; Krusche, A. V.; Richey, J. E.; Yager, P. L.; Dittmar, T. Fate of the Amazon River dissolved organic matter in the tropical Atlantic Ocean. *Global. Biogeochem. Cy.* **2015**, *29* (5), 677−690.
(20) Xue, M.; Zhu, C. A Study and Application on Machine Learning of Artificial Intellligence, 2009. *International Joint Conference on Artificial Intelligence*, April 25−26, 2009; pp 272−274.
(21) Lou, R.; Lv, Z.; Dang, S.; Su, T.; Li, X. Application of machine learning in ocean data. *Multimedia Systems* **2021**, 1−10.
(22) Palansooriya, K. N.; Li, J.; Dissanayake, P. D.; Suvarna, M.; Li, L.; Yuan, X.; Sarkar, B.; Tsang, D. C. W.; Rinklebe, J.; Wang, X.; Ok, Y. S. Prediction of Soil Heavy Metal Immobilization by Biochar Using Machine Learning. *Environ. Sci. Technol.* **2022**, *56* (7), 4187−4198.
(23) Huang, K.; Zhang, H. Classification and Regression Machine Learning Models for Predicting Aerobic Ready and Inherent Biodegradation of Organic Chemicals in Water. *Environ. Sci. Technol.* **2022**, *56* (17), 12755−12764.
(24) Zhao, W.; Bao, H.; Huang, D.; Niggemann, J.; Dittmar, T.; Kao, S. J. Evidence from molecular marker and FT-ICR-MS analyses for the source and transport of dissolved black carbon under variable water discharge of a subtropical Estuary. *Biogeochemistry* **2023**, *162*, 43.
(25) He, D.; He, C.; Li, P.; Zhang, X.; Shi, Q.; Sun, Y. Optical and Molecular Signatures of Dissolved Organic Matter Reflect Anthropogenic Influence in a Coastal River, Northeast China. *J. Environ. Qual.* **2019**, *48* (3), 603−613.
(26) Osterholz, H.; Kirchman, D. L.; Niggemann, J.; Dittmar, T. Environmental Drivers of Dissolved Organic Matter Molecular Composition in the Delaware Estuary. *Frontiers in Earth Science* **2016**, *4*, 95.
(27) Cai, R.; Zhou, W.; He, C.; Tang, K.; Guo, W.; Shi, Q.; Gonsior, M.; Jiao, N. Microbial processing of sediment-derived dissolved organic matter: Implications for its subsequent biogeochemical cycling in overlying seawater. *Journal of Geophysical Research: Biogeosciences* **2019**, *124* (11), 3479−3490.
(28) Zheng, X.; Cai, R.; Yao, H.; Zhuo, X.; He, C.; Zheng, Q.; Shi, Q.; Jiao, N. Experimental insight into the enigmatic persistence of marine refractory dissolved organic matter. *Environ. Sci. Technol.* **2022**, *56* (23), 17420−17429.
(29) Seidel, M.; Yager, P. L.; Ward, N. D.; Carpenter, E. J.; Gomes, H. R.; Krusche, A. V.; Richey, J. E.; Dittmar, T.; Medeiros, P. M. Molecular-level changes of dissolved organic matter along the Amazon River-to-ocean continuum. *Mar. Chem.* **2015**, *177*, 218−231.
(30) Hertkorn, N.; Benner, R.; Frommberger, M.; Schmitt-Kopplin, P.; Witt, M.; Kaiser, K.; Kettrup, A.; Hedges, J. I. Characterization of a major refractory component of marine dissolved organic matter. *Geochim. Cosmochim. Acta* **2006**, *70* (12), 2990−3010.
(31) Komer, B.; Bergstra, J.; Eliasmith, C. In Hyperopt-sklearn: automatic hyperparameter configuration for scikit-learn. *ICML Workshop on AutoML 2014*; Citeseer, 2014; p 50.
(32) Lundberg, S. M.; Lee, S. I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*; MIT Press, 2017; Vol *30*, p 4765.
(33) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S. I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56.
(34) Breiman, L. Random forests. *Machine learning*. **2001**, *45*, 5−32.
(35) Alpaydin, E. *Introduction to Machine Learning*, 3rd ed.; MIT Press, 2014.
(36) Herzsprung, P.; Wentzky, V.; Kamjunke, N.; von Tümpling, W.; Wilske, C.; Friese, K.; Boehrer, B.; Reemtsma, T.; Rinke, K.; Lechtenfeld, O. J. Improved Understanding of Dissolved Organic Matter Processing in Freshwater Using Complementary Experimental and Machine Learning Approaches. *Environ. Sci. Technol.* **2020**, *54* (21), 13556−13565.
(37) Peng, P.; Wong, R. C.W.; Yu, P. S. Learning on Probabilistic Labels. In *Proceedings of the 2014 SIAM International Conference on Data Mining (SDM)*; Society for Industrial and Applied Mathematics, 2014; pp 307−315.
(38) Ji, H. K.; Sun, Q. S.; Ji, Z.-X.; Yuan, Y.-H.; Zhang, G.-Q. Collaborative probabilistic labels for face recognition from single sample per person. *Pattern Recognition* **2017**, *62*, 125−134.
(39) Helms, J. R.; Stubbins, A.; Ritchie, J. D.; Minor, E. C.; Kieber, D. J.; Mopper, K. Absorption spectral slopes and slope ratios as indicators of molecular weight, source, and photobleaching of chromophoric dissolved organic matter. *Limnol. Oceanogr.* **2008**, *53* (3), 955−969.
(40) Cory, R. M.; McKnight, D. M.; Chin, Y.-P.; Miller, P.; Jaros, C. L. Chemical characteristics of fulvic acids from Arctic surface waters: Microbial contributions and photochemical transformations. *Journal of Geophysical Research: Biogeosciences* **2007**, *112* (G4), DOI: 10.1029/2006JG000343.
(41) Mann, P. J.; Davydova, A.; Zimov, N.; Spencer, R. G. M.; Davydov, S.; Bulygina, E.; Zimov, S.; Holmes, R. M. Controls on the composition and lability of dissolved organic matter in Siberia's Kolyma River basin. *Journal of Geophysical Research: Biogeosciences* **2012**, *117* (G1), DOI: 10.1029/2011JG001798.
(42) Judd, K. E.; Crump, B. C.; Kling, G. W. Bacterial responses in activity and community composition to photo-oxidation of dissolved organic matter from soil and surface waters. *Aquatic Sciences* **2007**, *69* (1), 96−107.
(43) Cory, R. M.; Ward, C. P.; Crump, B. C.; Kling, G. W. Carbon cycle. Sunlight controls water column processing of carbon in arctic fresh waters. *Science* **2014**, *345* (6199), 925−8.
(44) Cory, R. M.; McNeill, K.; Cotner, J. P.; Amado, A.; Purcell, J. M.; Marshall, A. G. Singlet oxygen in the coupled photochemical and biochemical oxidation of dissolved organic matter. *Environ. Sci. Technol.* **2010**, *44* (10), 3683−9.
(45) Gonsior, M.; Peake, B. M.; Cooper, W. T.; Podgorski, D.; D'Andrilli, J.; Cooper, W. J. Photochemically induced changes in dissolved organic matter identified by ultrahigh resolution fourier transform ion cyclotron resonance mass spectrometry. *Environ. Sci. Technol.* **2009**, *43* (3), 698−703.
(46) Stubbins, A.; Niggemann, J.; Dittmar, T. Photo-lability of deep ocean dissolved black carbon. *Biogeosciences* **2012**, *9* (5), 1661−1670.
(47) Herzsprung, P.; Hertkorn, N.; Friese, K.; Schmitt-Kopplin, P. Photochemical degradation of natural organic sulfur compounds (CHOS) from iron-rich mine pit lake pore waters–an initial understanding from evaluation of single-elemental formulae using ultra-high-resolution mass spectrometry. *Rapid Commun. Mass Spectrom.* **2010**, *24* (19), 2909−24.
(48) Mesfioui, R.; Abdulla, H. A.; Hatcher, P. G. Photochemical alterations of natural and anthropogenic dissolved organic nitrogen in the York River. *Environ. Sci. Technol.* **2015**, *49* (1), 159−67.
(49) Coppola, A. I.; Seidel, M.; Ward, N. D.; Viviroli, D.; Nascimento, G. S.; Haghipour, N.; Revels, B. N.; Abiven, S.; Jones, M. W.; Richey, J. E.; Eglinton, T. I.; Dittmar, T.; Schmidt, M. W. I. Marked isotopic variability within and between the Amazon River and marine dissolved black carbon pools. *Nat. Commun.* **2019**, *10* (1), 4018.
(50) Bao, H.; Niggemann, J.; Luo, L.; Dittmar, T.; Kao, S. J. Aerosols as a source of dissolved black carbon to the ocean. *Nat. Commun.* **2017**, *8* (1), 510.
(51) Kurek, M. R.; Stubbins, A.; Drake, T. W.; Dittmar, T.; M. S. Moura, J.; Holmes, R. M.; Osterholz, H.; Six, J.; Wabakanghanzi, J. N.; Dinga, B.; Mitsuya, M.; Spencer, R. G. M. Organic Molecular Signatures of the Congo River and Comparison to the Amazon. *Global. Biogeochem. Cy.* **2022**, *36* (6), e2022GB007301.
(52) Wagner, S.; Riedel, T.; Niggemann, J.; Vahatalo, A. V.; Dittmar, T.; Jaffe, R. Linking the Molecular Signature of Heteroatomic

https://doi.org/10.1021/acs.est.3c00199
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Dissolved Organic Matter to Watershed Characteristics in World Rivers. *Environ. Sci. Technol.* **2015**, *49* (23), 13798−806.
(53) Zhou, Y.; He, D.; He, C.; Li, P.; Fan, D.; Wang, A.; Zhang, K.; Zhao, C.; Chen, B.; Wang, Y.; Shi, Q.; Sun, Y. Spatial changes in molecular composition of dissolved organic matter in the Yangtze River Estuary: implications for estuarine carbon cycling. *Sci. Total Environ.* **2021**, *759*, 143531.
(54) Jiao, N.; Robinson, C.; Azam, F.; Thomas, H.; Baltar, F.; Dang, H.; Hardman-Mountford, N. J.; Johnson, M.; Kirchman, D. L.; Koch, B. P.; Legendre, L.; Li, C.; Liu, J.; Luo, T.; Luo, Y. W.; Mitra, A.; Romanou, A.; Tang, K.; Wang, X.; Zhang, C.; Zhang, R. Mechanisms of microbial carbon sequestration in the ocean − future research directions. *Biogeosciences* **2014**, *11* (19), 5285−5306.
(55) Read, J.; Perez-Cruz, F. Deep learning for multi-label classification. *arXiv preprint* **2014**, arXiv:1502.05988.
(56) Stubbins, A.; Mann, P. J.; Powers, L.; Bittar, T. B.; Dittmar, T.; McIntyre, C. P.; Eglinton, T. I.; Zimov, N.; Spencer, R. G. M. Low photolability of yedoma permafrost dissolved organic carbon. *Journal of Geophysical Research: Biogeosciences* **2017**, *122* (1), 200−211.
(57) Lu, K.; Li, X.; Chen, H.; Liu, Z. Constraints on isomers of dissolved organic matter in aquatic environments: Insights from ion mobility mass spectrometry. *Geochim. Cosmochim. Acta* **2021**, *308*, 353−372.
(58) Yu, M.; Petrick, L. Untargeted high-resolution paired mass distance data mining or retrieving general chemical relationships. *Commun. Chem.* **2020**, *3*, 1−6.
(59) Liu, J.; Wang, C.; Hao, Z.; Kondo, G.; Fujii, M.; Fu, Q. L.; Wei, Y. Comprehensive understanding of DOM reactivity in anaerobic fermentation of persulfate-pretreated sewage sludge via FT-ICR mass spectrometry and reactomics analysis. *Water Res.* **2023**, *229*, 119488.
(60) Wu, G.; Wang, X.; Zhang, X.; Ren, H.; Wang, Y.; Yu, Q.; Wei, S.; Geng, J. Nontarget screening based on molecular networking strategy to identify transformation products of citalopram and sertraline in wastewater. *Water Res.* **2023**, *232*, 119509.

https://doi.org/10.1021/acs.est.3c00199
Environ. Sci. Technol. XXXX, XXX, XXX−XXX