Optimizing machine learning-based prediction of terrestrial dissolved organic matter in the ocean using fluorescence and LC-FTMS data

Marlo Bareth,[*,†,‡] Boris P. Koch,[*,†,¶] Gabriel Zachmann,[‡] Xianyu Kong,[†] Oliver J. Lechtenfeld,[§] and Sebastian Maneth[‡]

†*Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, Ecological chemistry department, Am Handelshafen 12, 27570 Bremerhaven, Germany*

‡*University of Bremen, Faculty 3 – Mathematics and Computer Science, Bibliothekstr. 5, 28359 Bremen, Germany*

¶*University of applied sciences Bremerhaven, An d. Karlstadt 8, 27568 Bremerhaven, Germany*

§*Helmholtz Centre for Environmental Research – UFZ, Department Environmental Analytical Chemistry, Research group BioGeoOmics, Permoserstr. 15, 04318 Leipzig, Germany*

E-mail: Marlo.Bareth@awi.de; Boris.Koch@awi.de

Abstract

Marine dissolved organic matter (DOM) is an extremely complex mixture of organic compounds that plays a crucial role in the global carbon cycle. In the Arctic, climate change is accelerating the release of terrestrial organic carbon. Since chemical information is the only way to track DOM sources and fate, it is essential to improve analytical and data science approaches to assess DOM composition. Here, we compare

random forest (RF), support vector machines, and generalized linear models (GLM) to predict a fluorescence-derived proxy for terrestrial DOM based on molecular formula data from liquid chromatography coupled with Fourier transform mass spectrometry (LC-FTMS). We systematically evaluate different data preprocessing, normalization, and ML techniques to optimize prediction accuracy and computational efficiency. Our results show that a generalized linear model (GLM) with sum normalization provides the most accurate and efficient predictions, achieving a normalized root mean square error (NRMSE) of 5.7%—close to the precision of the fluorescence measurement. The prediction based on RF regression was slightly less accurate and required significantly more computation time compared to GLM, but was most robust against data preprocessing and independent of linear correlations. Feature selection significantly improved the performance of all models, with robust predictions obtained using only ca. 2,000 of the ca. 70,000 molecular features per sample. Additionally, we assessed the impact of chromatographic retention time on prediction accuracy and explored the key molecular features contributing to terrestrial DOM signatures using Shapley values and permutation importance (for RFs). Our study is a blueprint for the application of ML to enhance the analysis of high-resolution mass spectrometry data, offering a scalable approach for predicting information important for the understanding of marine DOM chemistry.

1 Introduction

The large reservoir of marine dissolved organic matter (DOM) is an extremely complex mixture of organic compounds.[1,2] In the dissolved phase, chemical information on individual constituents allows reconstruction of DOM sources, transformations, and its role in the global carbon cycle. The immense chemical complexity that is explored with high-end analytical tools requires new data science approaches to extract relevant chemical information.

DOM results from a plethora of different organic matter sources and biological and chem-

ical alteration processes. It can be categorized into terrestrial (decomposed biomass mainly exported by rivers) and marine origin (mainly marine microalgae). In the context of climate change, understanding and monitoring DOM fluxes in the oceans is crucial for an accurate assessment of the global carbon cycle.[3] Particularly, the Arctic region is experiencing a larger than average increase in temperatures,[4] resulting in the massive release of terrestrial organic carbon stored in permafrost[5] that is exported through the Siberian rivers into the Arctic Ocean. The accelerating erosion of the Arctic coastline is another, more direct source of terrestrial DOM.[6] It is therefore likely that the composition of DOM in the Arctic Ocean will change with yet unknown consequences for the carbon cycle. To track these changes, it is crucial to improve data science approaches to evaluate the enormous amount of data generated by state-of-the-art analytical techniques[7,8] such as fluorescence spectroscopy and mass spectrometry.

Between 20% and 70% of the DOM can absorb visible and ultraviolet light[9] and is called chromophoric dissolved organic matter (CDOM). Some of this CDOM emits light as fluorescence (FDOM) when excited with light. Recording a range of intensities of the emission wavelengths for different excitation wavelengths is called excitation-emission matrices (EEMs).[10] A common approach for evaluating EEMs is parallel factor analysis (PARAFAC),[11,12] which reduces the multi-dimensional data into several linear components that can be used for the assessment of DOM sources (e.g. terrestrial input) and biological activity.[13]

Fourier transform ion cyclotron resonance or Orbitrap mass spectrometry (FTMS) can also be used to acquire chemical information about DOM. In FTMS, the DOM components are ionized and their exact molecular weight and intensity are measured, from which a molecular formula can be calculated. In direct infusion FTMS, a DOM extract is directly ionized and a single mass spectrum is generated. By coupling reversed-phase liquid chromatography and FTMS (LC-FTMS), the DOM is further separated according to its chemical polarity so that several mass spectra are recorded along the chromatographic retention time. LC-FTMS has recently been applied to measure unprocessed filtered seawater.[7] The main advantage of

the new approach is that it avoids bias caused by solid-phase extraction.[14–16] The resulting LC-FTMS raw data is large (ca. 25 GB per sample) and its efficient evaluation requires modern data analysis methods that are able to detect non-linear relationships for making predictions using this data.

Several recent studies applied machine learning (ML) for DOM data evaluation, for example, the classification of DOM bioavailability and reactivity based on molecular formula data.[17,18] Regression tasks have been approached with ML algorithms, such as random forest (RF), gradient boosting, linear regression, and support vector machine, for predicting stable carbon isotope ratios.[16] LC-FTMS data was used to predict Spearman's rank correlation coefficients of environmental factors (e.g. land use, sodium or magnesium concentration) and Shapley Additive Explanation values (SHAP values) were calculated to find trends of molecular descriptors .[19] RF regression was also applied to predict how well molecular formulas correlate with chlorophyll concentration and solar radiation.[20] The resistance of DOM against UV irradiation was evaluated using multi label ML regression methods.[21] Riverine DOM was studied using RFs and SHAP values for key feature analysis and combined model features into chemical groups.[22]

Previous ML-based evaluation of FDOM was typically based on PARAFAC components as features, which were used to predict origins of pollutants[23,24] and reduction in oxidant exposure.[25] Since it is now possible to measure original seawater for its DOM composition using LC-FTMS,[7] we can predict for the first time PARAFAC components based on non-extracted marine DOM data.

In our study, we aimed to lay the groundwork for analyzing FTMS and LC-FTMS data sets with ML methods by exploring the feasibility and challenges of using molecular formulas directly as features. For this, we used mass spectrometric and fluorescence data that were acquired from samples taken during a unique full-year ship-based campaign in the central Arctic Ocean (MOSAiC; Multidisciplinary drifting Observatory for the Study of Arctic Climate;.[26] We validated different ML methods for their ability to efficiently predict an

exemplary environmental parameter (contribution of terrestrial DOM in the Central Arctic) based on molecular formula data acquired by mass spectrometry. In this case study, we tested multiple methods for data preparation, reduction, and normalization: (i) sum and (ii) ubiquitous sum normalization, (iii) normalization by the dissolved organic carbon (DOC) concentration of a sample, and (iv) log ratio transformation. All of these combinations were optimized using hyperparameter tuning via grid search. Finally, we explored chemical characteristics of the PARAFAC predictor variable.[27] We regard our methodological study as a basis for similar applications of ML for LC-FTMS data in the future. Our main research questions for our case study are:

1. How well can ML algorithms predict a fluorescence proxy for terrestrial DOM based on LC-FTMS measurements?

2. Which preprocessing, normalization and ML method yields the best, most efficient, and most robust predictions?

3. Does the consideration of the chromatographic retention time in LC-FTMS improve the prediction?

4. Which features are most important for good predictions?

The best performing model in our study was a generalized linear model (GLM) with a root mean square error that was only 5.7% of the original scale of the terrestrial component. This model also had the fastest running times for training and tuning. RF regression models were least prone to changes in the preprocessing and normalization and also covered non-linear relationships. Fewer features generally led to improved performance in all machine learning approaches, and precise predictions were achieved using only 2,000 features instead of the entirety of ca. 70,000 features (molecular formula and retention time combination) per sample of the LC-FTMS data.

2 Methods

2.1 Origin of the water samples

95 water samples were collected during the "Multidisciplinary Drifting Observatory for the Study of Arctic Climate" (MOSAiC) expedition by the research vessel Polarstern, a drift study from October 2019 to July 2020.[26] The vessel passively drifted with the ice floes from the Amundsen Basin via the western Nansen Basin and Yermak Plateau, to the Fram Strait. A second drift period continued from the end position of the first drift, while the third drift started again from the Amundsen bay.[28] During the three drifts, the water column below the vessel was sampled from surface to bottom water.

2.2 LC-FTMS measurements of DOM in water samples

Liquid chromatography (LC) was applied before detection with Fourier transform ion cyclotron resonance mass spectrometry (FTMS). The LC (UltiMate 3000RS, Thermo Fischer Scientific, Waltham, U.S.A.) was carried out using a reversed-phase column (ACQUITY HSS T3, 1.8 µm, 100 Å, 150 x 3 mm, Waters,Milford, U.S.A.) and a water-methanol gradient[7] so that the chromatographic retention time represented a decreasing polarity of DOM molecules (see Fig. 1 top left going front to back). In FTMS (solariX XR, BrukerDaltonics, Billerica, U.S.A.) the molecules were ionized by electrospray ionization (Apollo II, Bruker Daltonics, Billerica, U.S.A., capillary voltage: 4.3 kV) and a unitless intensity for each ion as molecular mass (in Dalton) per charge (m/z) was recorded in a range from 150 m/z to 1000 m/z. To improve the signal-to-noise ratio, the mass spectra were summed in one-minute retention time segments. For our study, we used mass spectra from ten segments for each sample with start times between 12.3 and 22.3 minutes by using a custom script in DataAnalysis (Version 6, Bruker). Molecular formulas were assigned using the R tool UltraMassExplorer[8] as described in Kong et al. (submitted).

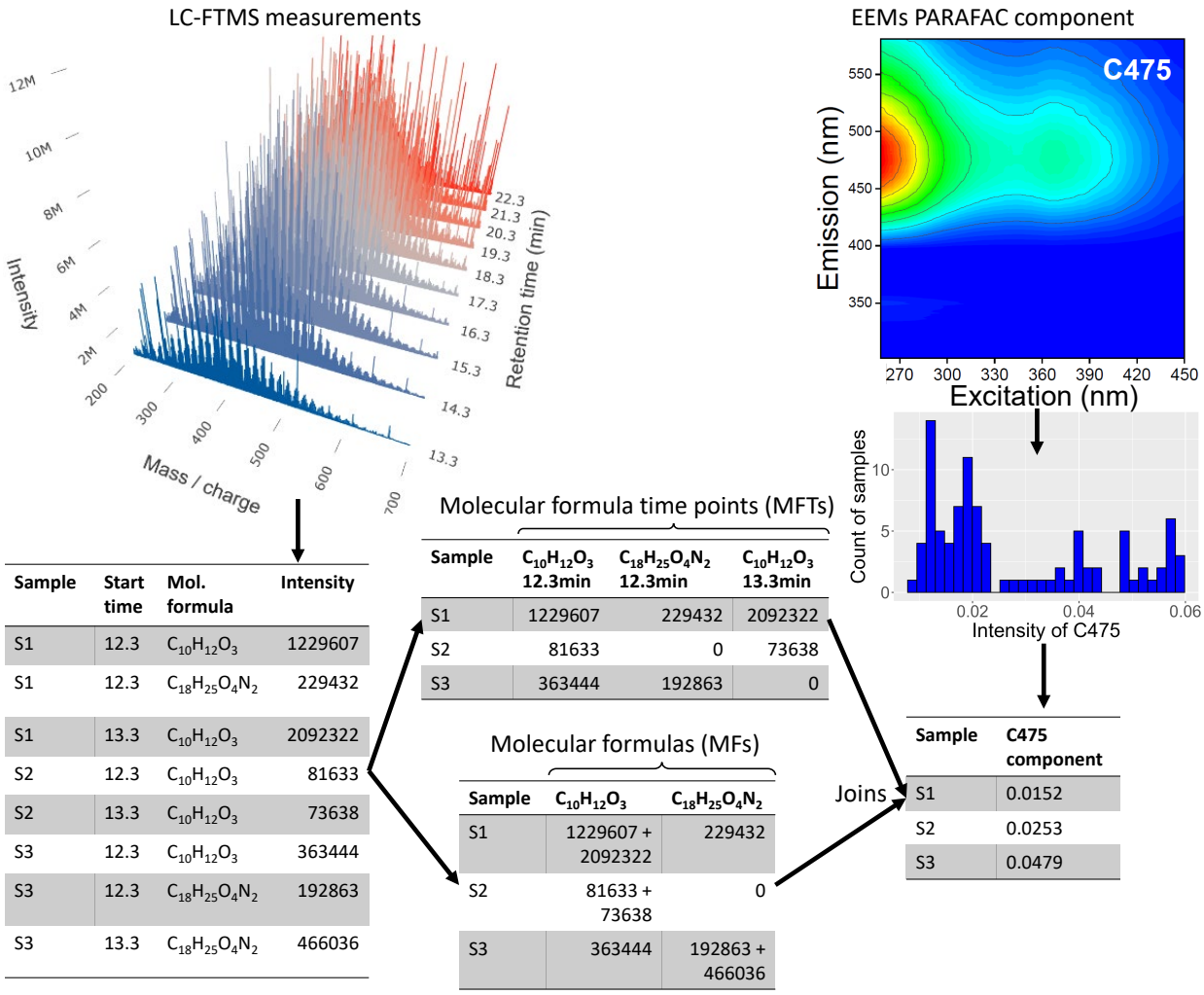


Figure 1: Overview of the data preprocessing: LC-FTMS data as mass spectra (top left) and as table of assigned molecular formulas (bottom left); tables of molecular formula data including the retention time (MFRTs, "time aware"; top center) and molecular formulas (MFs) for which the intensity of all retention times were summed ("time agnostic data"; bottom center); the PARAFAC component C475 (right side) was derived from EEM fluorescence analysis (top right) and represents terrestrial dissolved organic matter. Adapted in part (top right EEM) with permission from Kong, 2022.[29] Copyright 2022 Xianyu Kong (CC BY 4.0)

2.3 Fluorescence data

Three-dimensional excitation-emission matrix spectra (EEMs) were acquired and analyzed (for details cf. Kong et al. 2024[28]) using a parallel factor analysis (PARAFAC),[12,30,31] which is a generalized principal component analysis (PCA).[32] The PARAFAC component that had its maximum emission at 475 nm was named C475, and was shown to have terrestrial

characteristics.[28,33] It is used as a terrestrial proxy and hence as the predicted variable in our work.

2.4 Dataset

The data table for our study consisted of ca. 1.48 million rows, in which the first four columns were the longitude, latitude, depth, and date time. Other columns specified the start time of the retention time segment (each one minute) from the liquid chromatography, the measured (neutral) mass in Dalton (Da), the corresponding molecular formula of the mass,[34] the measured mass peak intensity, the intensity of the PARAFAC component (ranging from 0.00927 to 0.0624; (Fig. 1), and the dissolved organic carbon (DOC) concentration in µmol carbon per kilogram water (µmol/kg; used for DOC normalization).

2.5 Wide table format

To be a suitable input for the machine learning methods we applied, the data table described before was transformed into a wider format, where each sample was represented by exactly one row, resulting in a table of 95 rows. We distinguish two different wide tables obtained from the original table in two different ways:

1. the *time agnostic* table
2. the *time aware* table.

The time agnostic table had one column for each different molecular formula of the narrow table. If a molecular formula appeared in a given sample, then the entry for this molecular formula was the mean of all intensities for that sample and formula. If the molecular formula is not present in this sample, the entry is set to zero. The time aware table has one column for each combination of *molecular formula* (MF) and segment retention time (MFRT). The entry is the corresponding intensity if it existed or zero otherwise 1.

2.6 Feature elimination

Many molecular formulas were only present in a few samples (or retention time segements). If molecular formula feature was not present, it was assigned an intensity of zero in the two wide tables. These entries accounted for an average of 79.3% of the data and to check the influence of the zero value contribution, two approaches of feature elimination were applied: (i) Elimination of each feature that contained at least one zero value; remaining features were called *ubiquitous*; (ii) elimination of each feature that contained more than 90% zero values; remaining features were called *no low variance*.

2.7 Normalization

Different ranges of values, e.g., in DOC concentration or numbers of MFs can lead to problems in our comparisons of samples. Data normalization can help overcome such problems. We tested three different normalization methods, as well as one transformation:

(i) We expect higher intensities when more DOM is present in the sample. For DOC normalization (DOC-N) each mass peak intensity was divided by the DOC concentration of the sample.

(ii) Sum normalization (SUM) tries to balance out higher intensities in some samples. Instead of relying on other measurements (e.g. DOC concentration), we divide each intensity by the sum of all intensities of a spectrum.

(iii) Ubiquitous sum normalization (UBISUM) divides each intensity by the summed intensities of those molecular formulas that occur in *all* samples. This gives a common ground between samples and avoids including possible sample contaminations.

(iv) Log ratio transformation (ALR) was calculated by taking the logarithm of the ratio between each intensity and the ubiquitous feature with the lowest variance in intensity

over all samples. To avoid taking the logarithm of zero assigned intensities, one third of the lowest intensity is used as a zero replacement.

The normalizations were applied to the two wide table formats, which were filtered by combinations of the feature elimination. The two wide table formats yielded 32 different pre-processing combinations for four feature elimination combinations (none, ubiquitous filtering, low variances excluding filtering, or both), and four normalizations. The C475 component was normalized using the Z-score, setting the mean as zero and converting the scaling above or below in units of standard deviations.As the Z score takes the difference to the mean C475 value, positive values can be considered to indicate terrestrial influence, while negative normalized C475 values indicate a marine origin, assuming that only marine and terrestrial DOM sources predominate in the Arctic Ocean.

2.8 Machine learning methods

To train the machine learning models, we took a data set and trained the model to make predictions on the used data. We split the data set into two sets, *training* (80 %) and *testing* (20 %). The former was used to train the model and the latter to evaluate the models' performance on unseen data. In our study, we applied three machine learning methods: (i) generalized linear model (GLM) as a linear approach,[35] (ii) random forest regression (RF),[36] and (iii) support vector regression (SVR)[37] as non-linear approaches.

GLMs extend linear regression, which gives a prediction by building the sum of the products of a weight (*beta values*) and the model feature (e.g. the intensity of a molecular formula). One extension is a *random component*, which is a class of probability distribution that the response is assumed to follow. The other extension is the *link function*; it describes the relationship between the linear predictor and the prediction mean. We use a modified version of GLMs, which includes a regularization that tries to prevent overfitting by limiting the magnitude of the weights of the linear terms of the GLM. This "elastic net regularization" has two parameters: *Lambda*, the overall strength of the regularization and *alpha*, which

controls the norm that is used for the regularization by taking a value between zero and one. When alpha is one, it only applies the L1 norm (Manhattan distance) and is known as *Lasso* regression. When comparing two highly correlated features, it reduces one weight to zero. When alpha is zero, the complete regularization is based on the *Ridge* regression, which uses the L2 norm (Euclidian distance) and reduces the predictors but keeps all features.

The GLM method was tested due to its linear approach. Linear regression was previously used in DOM analysis but was outperformed by the nonlinear approaches.[21] Yet, we included the model, as its simplicity has two useful benefits. The GLMs can be trained quickly and provide simple access to a feature importance metric. To avoid overfitting and to improve the performance, we regularized the features with the elastic net.

RF regression models consist of multiple decision trees that form an average of different predictions. For the construction of each tree, only a subset of the data is available to create independent trees. This means not all features are available and not all samples. The decision tree consists of nodes, where the data is split into two subsets and directed into nodes further down in the tree. The splitting is done by trying numerous split candidates, which is called *mtry*, where the data is split according to one feature and a threshold. How well a split is considered at dividing the data into subset is determined by a *split rule*. The end of the repeated splitting is reached when a minimum number of samples is remaining in any child node, which is referred to as the *min node size*. Such leaf nodes are assigned the average of the remaining sample values. RFs were chosen in this study, because can show nonlinear relationships and are resistant to overfitting.[36,38] RFs were previously used in DOM analysis using mass spectrometry[18,20,21] and remote sensing data.[39]

SVR models try to find the best fit for the data within a margin of tolerance. This can be visualized in three dimensions as a tube or long cylinder that is fitted to minimize the error of points outside the tube. The errors inside the tube are not considered. How strictly the error is penalized, is defined by the *cost* parameter. *Sigma* steers how influential single samples are on the model. To make the model more suitable for different data sets, a kernel

can be applied. They map the input into a high-dimensional feature space. This allows the SVM to handle nonlinear data. The kernel functions can be based on different functions like polynomial or radial basis functions. SVMs were chosen in our study as a second non-linear ML method that was previously applied to DOM [18,21] and mass spectrometry data. [40,41] With GLMs already covering linear relationships, we chose the polynomial and the radial basis function kernel in our test.

2.9 Evaluation metrics

The model performance was measured using the root-mean-square error (RMSE) between the true PARAFAC component intensity and the predicted PARAFAC component intensity across all samples of the test set. With the predicted variable being a unitless PARAFAC component, we decided to convert the RMSE to be easier understandable. The normalized RMSE (NRMSE) was calculated by normalizing RMSE by the original scale of C475 values in the data set (see Fig 1, box plot on the right). This means the NRMSE is the RMSE divided by the difference between maximum and minimum of C475 values. During the training of models on the training set with different hyperparameters combinations (e.g. RFs with varying *mtry* values) RMSE was used to compare the models. We utilized NRMSE for reporting the results of the model that was best performing during training, based on it predicting the samples from the test set. To check whether a RMSE was an outlier, we utilized the standard error of the mean, which is the standard deviation divided by the square root of the number of repeated model trainings. For the significance tests of experiment comparisons, we use the two-sided rank test with continuity correction. A p-value of less than 0.05 is considered significant.

After tuning the models, we aimed to find the features that affected the model predictions the most, which we consider important. To evaluate these key features in RF models, the SHAP values and permutation importance determined the feature importance. For GLMs, the beta values were used, as they weigh the feature in a direction and thus can be considered

a scale of importance. This only works when the intercept of the linear model is zero. This way, a positive weight indicates a terrestrial feature. Sets of these features were compared base on the Jaccard similarity, which measures the similarity between two sets and has a range from zero, for sets without overlaps, to one, for equal sets.

Recursive feature elimination (RFE) is performed to check the model performances with different numbers of features. It is done by constructing a random forest, assessing the importance of the features, and removing the least important features from the dataset. Then a new random forest was built, and the next iteration of RFs were calculated until no features were left. For this, the permutation importance of the out-of-bag-error was used. This means that the samples that were not included for the building of the trees were used to evaluate the features by shuffling the feature value between these samples and comparing the corresponding predictions. We used a step size of 10% of the currently remaining features to be removed, giving us a good resolution at small feature numbers but not needing many models for high feature counts. To make the results more reliable, a cross validation with 10 folds was performed in each step. Each fold was used once as evaluation set to avoid over- or underscoring features for a single data split and the model only being able to predict on the training data.

2.9.1 SHAP values

The permutation importance provides insight into how important a feature was to find a good prediction. It does not grant insight into the direction, that a high, or low value of a feature, shifted the prediction. For this, SHAP (Shapley additive explanation) values can be used. The original Shapley values are based on game theory and how different players of a team contributed to a profit. The contribution is calculated by building coalitions of different players and player numbers, then predicting their profit and then each participants' contribution can be calculated. This is done for the features of ML predictions as well, by using SHAP values. They do simulate numerous possible feature coalitions instead of cal-

culating all coalitions, as the number of coalitions is exponentially increased by the number of features. Adding the SHAP values together results in the difference from the mean predicted value. The drawback of SHAP values is that they are calculated for each sample, and therefore are different for a feature when comparing different samples.[42]

2.10 Experimental setup

The programming language R (Version: 4.3.1[43]) was used for implementation, as it allows a seamless interaction with the UltraMassExplorer package.[8] The models were constructed using the *caret* package. Our experiments started with splitting the data tables into the commonly used 80% training set and 20% test set by using the *createDataPartition* method from caret, which is designed to build similar sets based on the C475 intensity. This avoided bias in the distribution of C475 inteensities in each set, but due to computational limitations, a nested cross validation where the test/training split would have been segmented as well was not feasible. The complete model training was only performed on the training set, and we used a grid search with ten-fold cross validation and ten repeats. The grid search builds models with different combinations of hyperparameters that influence the models. After initial tests, we selected some hyperparameters to have fewer values in the tuning grid to reduce the computation time of the training process. The ten-fold cross validation split the training set into ten parts, called folds. Nine of the folds were used to train the model and the other part was used to validate the model performance. For each fold this was performed for a total of ten iterations. Subsequently, the best performing models were selected and based evaluated on the test set. This test set was never used in training and was used to check if the model only performed well on the training data and not on new data (overfitting). We calculated the RMSE and NRMSE was only based on the test set evaluation, where overfitting was detected. For a few experiments, we repeated the cross validation with different random seeds to estimate the impact of the training and test data split.

For the GLM, we utilized the *glmnet* package (version 4.1-8). We assumed the error to be normally distributed and used the linear link function. The grid search contained the two parameters of an elastic net *alpha* and *lambda*. *alpha* was increased in 0.05 increments during the hyperparameter tuning. *lambda* took values from 0 to 2 in 0.05 increments. As mentioned in the previous chapter, we use an intercept of zero

For RF *ranger* method from caret was used. 1000 trees per random forest was found an acceptable compromise between performance and time required based on initial tests. The split rule was left as *variance* for all experiments. Values for mtry were tested in one-sixth increments of the number of features, up to the whole. The minimum node size was tested for values of 3, 5, 7, 10 and 15, having bigger steps between the higher values. As RF does not yield the same result every time a model is trained, 500 RFs were built after the hyperparameter tuning with the selected optimal tuning parameters.

For the polynomial kernel SVR, we tested *cost* hyperparameters of 0.01, 0.1, 0.25, 0.5, 0.75 and 1, as well as *polynomial degrees* from one to five degrees at a fixed *scale* parameter of one. The grid for the radial basis function kernel was built from *cost* and *sigma* values. *Cost* parameter values of 0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99 and 1 were used. We evaluated *sigma* values ranging from 10^{-6} to 1, including 10^{-6} up to 10^{-5} in factor 2.5 increments, $2 * 10^{-5}$, a coarser grid of $5 * 10^{-5}$ up to $5 * 10^{-4}$ in factor 5 increments and a sparse upper scale of 10^{-3} increasing by a factor of 10, up to one.

For selected models we also checked if the NRMSE was an outlier. For this, we repeated the tuning 1000 times and using different splits for the train and test set.

The model tuning and repeated model trainings were run on a server node that had 256 GB RAM and two AMD Rome Epyc 7702 processors with 128 cores total (albedo server, AWI). We used 120 threads in the experiments to allow for background processes to execute. All other preprocessing and experiments were conducted on a workstation with an Intel i7-1165G7 processor, an Nvidia T500 graphics card and 32 GB RAM.

3 Results

The preprocessing of the molecular formula data resulted in 32 data tables, each of which was used to train four ML methods, leading to 64 models without (Fig. 2) and 64 models with low variance filtering (Fig. 3).

3.1 Prediction performance using all features

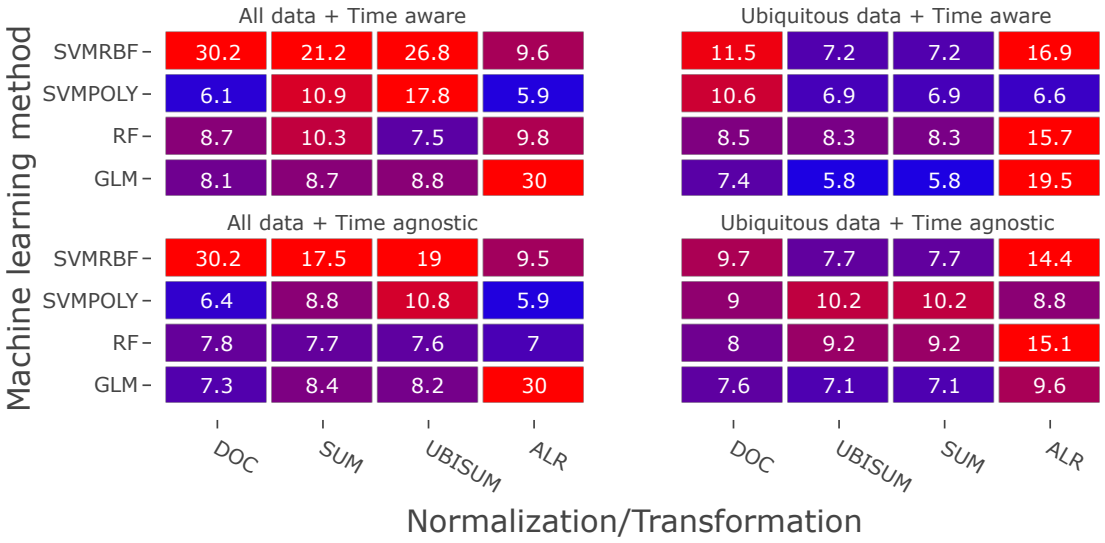


Figure 2: Model performance based on all molecular formula features: Normalized root-mean-square errors (NRMSE) in percent of the original scale of C475.

When comparing the NRMSE values, the best prediction was achieved with a GLM using ubiquitous data, the retention time dimension, and SUM or UBISUM normalization (Fig. 2, top right). GLMs performed best or second best (13 out of 16 variants), except for the models using ALR transformed data. The SVM models performed worst when using all data, with the exception of the models that used ALR transformed data. GLM was the only method that consistently yielded an NRMSE below 10%, when excluding ALR transformation. RF performance was best for the models that included all time aware data using SUM and

UBISUM normalization (lower left subplot; Fig. 2), whereas in the ubiquitous data and time aware experiments (top right subplot; Fig. 2), RF had the highest error. Generally, the NRMSE was similar or smaller, when using only the ubiquitous data, compared to the use of all data. The performance of models based on time aware data (top of Fig. 2) did not show significant differences ($p = 0.144$) compared to experiments with time agnostic data (bottom of Fig. 2). It should be noted that for ubiquitous data, the sum of intensities (SUM) was the same as the sum of ubiquitous features (UBISUM), leading to identical errors for models except RFs. The other metrics, such as R^2 and Mean Absolute error showed similar trends for all entries (Fig. S6, S8), but RFs did have the higest or second highest R^2 in 11 out of 16 cases. For the high error occurring in ALR transformed GLMs of unfiltered, time aware data, we exemplary performed 1000 repeated model trainings of the hyperparameter search with different train-test splits. The STEM was 0.547% NRMSE for 100 repeated splits and 0.184% NRMSE for 1000 repeated splits. For quality control, the repeated cross validation was performed using the unfiltered time aware data with DOC normalization that was used to train a GLM (Fig. 2 top left, bottom row first column). It was selected because the model had the highest feature count possible and the otherwise best performing normalization. The models showed an average NRMSE over 1000 models trained of 8.87%, compared to 8.1% in the heatmap.

3.2 Prediction performance after removal of low variance features

Removing the low variance features led to a 61–62% reduction in the number of features and a significantly ($p = 0.0417$) better performance (Fig. 3).

In contrast to all data experiments (Fig 2), the removal of low variance features overall led to an improved performance. The NRMSE values were consistently below 10%, except for five experiments, when excluding ALR transformation. A significant change in NRMSE was found over all experiments ($p = 0.0417$) when excluding the ALR transformed models. The highest NRMSE with 16.9% is around half as high as the worst case in the non-filtered

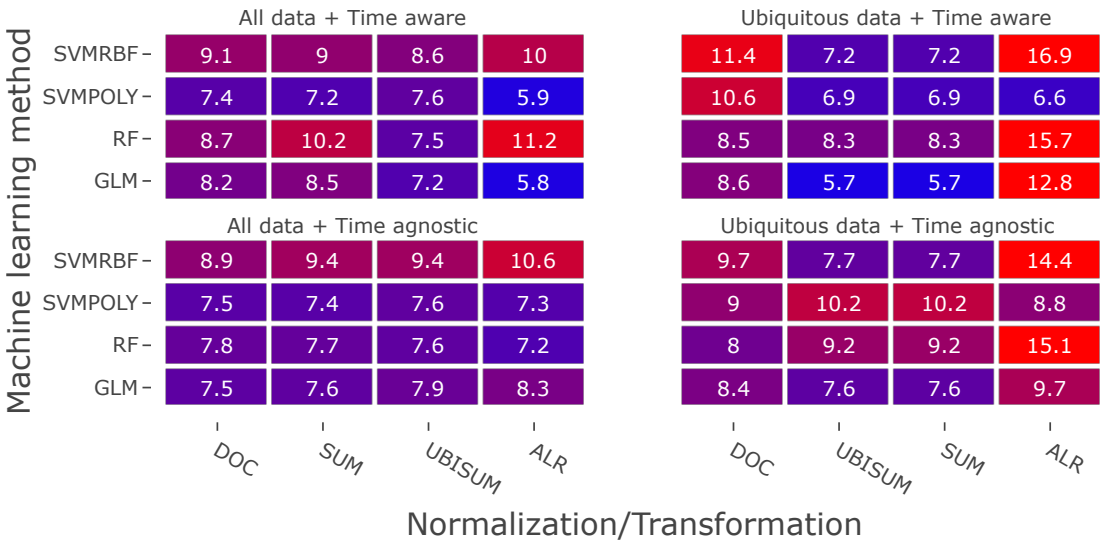


Figure 3: Model performance after removal of low variance features: Normalized root-mean-square errors (NRMSE) in percent of the original scale of C475.

experiments (30.2%). However, no significant ($p = 0.702$) difference was found between the ubiquitous filtered data without (Fig. 2 right) and with low variance filtering (Fig. 3 right). The difference between model performance using only low variance filtered data (Fig. 3 left) and those with additionally ubiquitous filtered data (Fig. 3 right) was not significant when excluding ALR transformation ($p = 0.643$). A paired Wilcoxon signed rank test with continuity correction between the time aware and time agnostic models found no significant difference ($p = 0.417$) between the two pairs of 64 models and also not significant between the low variance filtered data ($p = 0.660$). GLMs based on the time aware and SUM/UBISUM normalization (Fig. 3 top right) yielded the best predictions, similar to the results of the low variance including data sets (Fig. 2). SVM with polynomial kernel performed best or second best in most (seven out of eight) preprocessings for data, which was not further filtered by ubiquitous contents (Fig. 3 left). GLMs performed equally good in 14 out of 16 experiments. SVM with polynomial Kernel outperformed GLM or performed equal in half of the 16 experiments, where five out of these eight used time agnostic data. Similar trends

were observed for the R^2 and Mean Absolute Error (see Fig. S7, S9).

3.3 Random forest regression

Although the RF models did not yield the overall best performance when focusing on NRMSE, the RFs did have high R^2 values and didn't require linear relationships, and performed well across all tested normalization (excluding ALR transformation) and filtering approaches (Figs. 2 & 3). Since RF regression was the only method that was not deterministic, we assessed the prediction error introduced by randomness. The performances of 500 repeated runs were evaluated for RFs between variants that were not filtered to exclude low variances. The NRMSE introduced by RF uncertainty was much smaller than the errors between the models (details in Fig. S1) and the RF uncertainty was significantly lower than the NRMSE differences between time aware and time agnostic models ($p = 0.0078$).

3.3.1 Recursive feature elimination: Evaluation of random forest model performance

We tested recursive feature elimination to validate how the model performance changes with the number of features used in each model where we used the complete time aware data with UBISUM normalized. The importance is based on the permutation importance obtained from the out-of-bag error that is accessible in RF models from caret and ranger.

In each iteration of the recursive feature elimination, the 10% least important features were removed (Fig. 4) and a 10-fold cross validation and recalculation of the importance was performed. For less than ca. 2,000 features, the error increased from 10% to about 12% NRMSE for less than 300 features. A plateau of about 30% NRMSE was reached at 20 or fewer features.

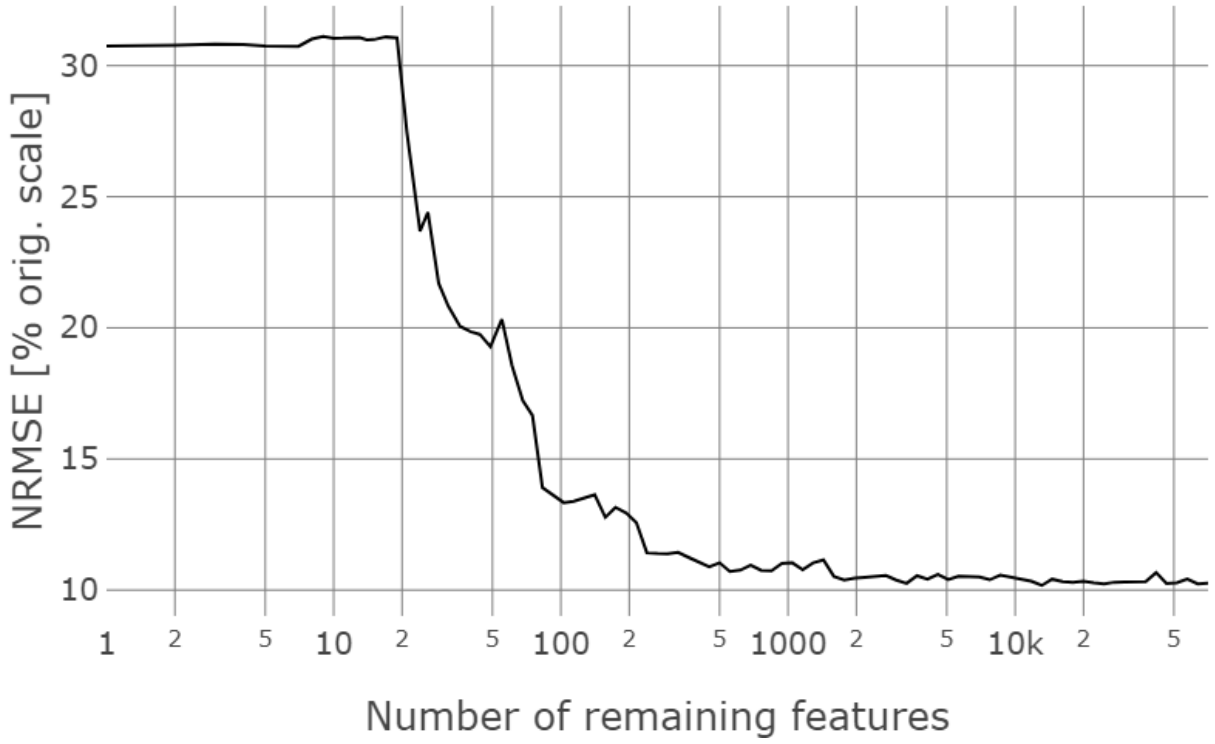


Figure 4: Recursive feature elimination: normalized root-mean-square error (NRMSE) of the random forest experiment based on time aware DOC normalized and non-filtered data; the initial number of features is $n = 70,683$.

3.4 Feature importances

As a first step for key feature analysis, we used RF permutation importance and SHAP values to identify the most important molecular formulas and MFRTs, respectively, that represented the C475 fluorescence component.

3.4.1 Random forest permutation importance

The permutation importance of RFs is a commonly applied parameter to identify the most relevant features. To check how reliable the found importance values were, we created 500 forests. The hyperparameter optimization was only performed once at the beginning and then used for the RF creation. From these, we selected a number of top ranking features from each forest, increasing in number (1 to 100). Then we built a set that contained all unique formulas that were included in these 500 forests. The RFs reached 592 different MFRTs in

the top 100 ranked important features by permutation importance, using a non-filtered data set, of which ALR transformation resulted in the best model (see Fig. S2). If time agnostic data was used, the number of different features reached around 300 MFs. Similar numbers of varying features were observed from ranking the SHAP values of the built RFs (see Fig. S3 for time aware and S4 for time agnostic data).

3.4.2 SHAP values of random forest models

To identify the importance of the features in the RFs, we utilized SHAP values. We associated positive SHAP values with terrestrial features, as our prediction variable was high for terrestrial samples. Therefore, negative SHAP values were assumed to represent marine features. SHAP values for the features were calculated on a sample-wise basis and were calculated from the RF model that excluded low variance features, and used DOC-N and time agnostic data. To get an overview of different samples, we selected three example samples from the test set. We selected one sample that was close to the average C475 of the test set, as well as the most terrestrial and marine samples. To evaluate the composition of the features we scaled the positive SHAP values from one to zero and computed a weighted average composition for these samples. For the average terrestrial sample, the positive SHAP values accumulated in the lower mass region of 200–400 m/z. When scaling these SHAP values to between zero and one and multiplying them with the composition of the feature, the average composition is $C_{18.1}H_{21.5}N_1O_{9.6}S_{0.18}$. This means a sulfur to carbon atom (S/C) ratio of 0.009 and a carbon to nitrogen atom (N/C) ratio of 0.06. The very marine sided sample had a composition of $C_{18.3}H_{24}N_{1.4}O_{10}S_{0.24}$. The S/C ratio was 0.013 and the N/C ratio was 0.08. For the very terrestrial sample, the composition of a MF with positive SHAP values was $C_{18.2}H_{21.1}N_1O_{9.6}S_{0.18}$, yielding a S/C ratio of 0.01 and a N/C ratio of 0.05, which were similar to the average terrestrial sample.

From the average terrestrial sample from the test set, we found 3,546 positive SHAP values, which indicated terrestrial content. A previous analysis of this data, via linear re-

gression, found 1,450 MFs to be terrestrial.[28,29] The overlap of the two sets was 759 MFs and the Jaccard similarity between the sets was 0.18. To compare the permutation importance as well, we use the 1,000 repeated RF runs with DOC normalization, where we selected the top 100 features by permutation importance. Around 592 unique MFRTs were identified in the time aware data. For time agnostic data, the number is 316 MFs. The latter was compared to the 1,450 terrestrial features[28,29] and yielded a 0.16 Jaccard similarity between the sets.

Lastly, we compare our sets from SHAP values and from permutation importance, to each other. We found a Jaccard similarity of 0.70 for the time-aware data (MFRTs) and 0.74 for the time agnostic data (MFs).

3.5 Running Times

The computation time for the grid search was strongly dependent on the machine learning methods (Fig. 5). GLMs were quickest for each preprocessing combination, whereas the RF took the longest time to compute, especially when no feature reduction and time aware data was used. In this case, the hyperparameter search for the Random forest took close to 200 minutes, with only small variations between the normalization types (191 minutes minimum and 201 minutes maximum). It is important to note that our experiments were executed on

Method	Filter	Time (min)
GLM	agnostic	0.06
GLM	aware	0.07
RF	agnostic	3.1
RF	aware	6.6
SVMPOLY	agnostic	1.1
SVMPOLY	aware	1.5
SVMRBF	agnostic	3.7
SVMRBF	aware	5.13

Method	Filter	Time (min)
GLM	agnostic	0.5
GLM	aware	1.1
RF	agnostic	59.5
RF	aware	194.5
SVMPOLY	agnostic	5.1
SVMPOLY	aware	11.3
SVMRBF	agnostic	19.0
SVMRBF	aware	48.3

Figure 5: Running times for the grid search, averaged over all different normalizations for ubiquitous data (left) and all data (right).

a high performance computer employing 120 cores that run in parallel (exemplary models trained on the laptop took 15–20 times longer). For ubiquitous data, GLMs were faster than RFs by a factor of between 53 (time agnostic) and 91 times (time aware), when we compare GLMs to SVMPOLY, then they are still faster by a factor of about 19 (Fig. 5). When considering all data, GLMs are faster than RFs by factors between 125 and 170 and GLMs are faster than SVMPOLY by a factor of around 10.

4 Discussion

4.1 How well can ML algorithms predict the fluorescence proxy for terrestrial DOM based on LC-FTMS measurements?

The most accurate, molecular-formula based prediction for the fluorescence component C475, a proxy for terrestrial organic matter, was achieved from the GLM. Depending on the choice of feature type (molecular formulas or molecular formula time points) and preprocessing, the lowest prediction error showed an NRMSE of 5.7% of the original range of C475. This value was only slightly above the precision of the fluorescence method itself (below a factor of two), which we estimated to have an NRMSE of 3.5% of the total range of intensities for all samples. This NRMSE estimation was calculated based on the repeatability of C475 values in the deep Arctic Ocean (400 m or below), which generally showed very similar fluorescence signals (cf. Kong et al. 2024). In a previous study, five ML models were tested for their ability to predict the stable carbon isotope ratio $\delta^{13}C$ from molecular formula data.[16] Based on solid-phase extracted DOM, the linear SVM kernel achieved the best predictions, in contrast to our study where GLMs and RF regression showed the best performance. The authors report a prediction error for $\delta^{13}C$ of 0.3‰. To make this value comparable to our results, we scaled this error to the entire range of $\delta^{13}C$ values (-27.7 to -21.9‰). This resulted in an NRMSE of 5.1%, which is comparable to our best value of an NRMSE of 5.7% for the GLM prediction of terrestrial fluorescence. Our best-performing SVM (NRMSE of

5.9%) used a polynomial kernel of first degree and was comparable to the results by Yi et al.[16] who used a linear SVM kernel.

4.2 Which combination of preprocessing, normalization and ML method yields the best predictions?

Generally, the best predictions for C475 were achieved with data that was filtered for ubiquitous formulas and normalized by SUM (equivalent to UBISUM normalization). The sum normalization was also used by Yi et al.[16], but the molecular formula assignment differed slightly: While we validated and filtered the molecular formula set by limiting the double-bond equivalents minus oxygen (DBE-O[44]) to a maximum of ten, Yi et al. (2003) filtered formulas by H/C < 3 and O/C < 1.5.[16] As expected, the choice of preprocessing and type of machine learning model had a large influence on the performance of the different models. Somewhat surprisingly, a reduction of the number of features led to an improved prediction. The performance of most models was improved by removing low variance features. Alternatively, filtering for ubiquitous formulas, if the low variance filtering was not applied. The only exception were the RF experiments, which delivered robust performance irrespective of the number of features when excluding the ALR transformed data.

The different normalization methods did influence the model performance substantially. The ALR transformation prevented high error rates for the polynomial kernel SVMs, when using unfiltered data, but increased the error for all other combinations. We discuss this in a later paragraph.

The time-aware data that included only the ubiquitous features and normalized by the sum of intensities of each segment performed best, with an NRMSE of 5.7% for GLMs. This was nearly (0.1% NRMSE) independent of low variance filtering, which could be due to the ubiquitous filter and excluding low variance features, since this way every sample measured a signal and the low variance features mostly are those with many cases of zero filling.

We clearly found that some model setups were particularly sensitive to higher numbers

of features in combination with the low number of samples (e.g., SVMs with RBF kernels, Fig. 2). RF regression and GLMs did handle the unfiltered data consistently well for most cases (Fig. 2 and 3). We assumed that RFs cope better with a relatively large number of features, as the large number of split candidates for each tree filter out the low variance features as bad splits.[38,45] Several experiments in our study yielded prediction performances with high errors (NRMSE of greater than 10%). The maximum NRMSE of 30.2% is a factor of more than eight above the precision of the fluorescence measurement. High uncertainty was observed particularly for those experiments that were based on a large number of features (molecular formula time points and no removal of low variance features; Fig. 2). We suspect that the situational large prediction errors for SVMs were caused by failing to find a solution to the equation system from having too few samples per feature.

We hypthesize that the high test set errors of ALR transformed models were caused by overfitting, meaning that the model adepted too much to the trainign data and does not generalize well. Evidence for that is that the model has high errors on the test set while showing low training set errors, where the RMSE was up to three times as high in the test data evaluation than in the chosen model from the grid search. Our theory for the overfitting was induced by collinearities arising either from the original data. They could ba a coincidence, or induced by the ALR transformation. We have no other hypothesis why the overfitting only affected ALR models to such a degree.

One general problem could also be the splitting of the test set. With only 95 samples, the left out samples possibly were not well-balanced for the features with high weights that were found by the model. This would explain why this behavior was not seen in the low variance filtered set with no ubiquitous filtering. Similar problems with a small test set have been observed and were approached via a pooled test set, which could be an approach for our future work.[46]

In our study, the reduction of the number of features led to an improved prediction performance for most experiments and faster processing time, particularly for RFs. Three

processes reduced the number of features: (i) creating time-agnostic data from the original retention time data (molecular formula time points, MFRTs), (ii) filtering for formulas or MFRTs that occurred in all samples (ubiquitous), and (iii) removing features with low variance. Utilizing all available features led to a larger variability of the prediction performance and increased the influence of the normalization method (see Fig. 2). Applying low variance feature filtering (ca. 61%) minimized the differences in performance between the models. It also improved the performance of the models with previously situational high errors like SVMs and the GLMs with ALR transformation in cases were ubiquitous filtering was not applied (Fig. 2 and 3 left). It also reduced the time for tuning the RF and SVMRBF models by circa three times (see Fig.5). Using only ubiquitous features, all normalizations except ALR normalization performed well over most experiments and reduced the calculation time from unfiltered data by 5–15 times (see Fig.5). The reduction of feature is generally beneficial, as each feature adds a new dimension that needs data to cover it.[47] The exclusion of features allows the data to fill more locations in the reduced dimensional space.[48] Only utilizing the ubiquitous features, the feature counts may reached areas of too strong filtering, where information was lost. This was supported by the RFE, where NRMSE rose quickly for data sets below 2,000 features. The time agnostic, ubiquitous data set had only ca. 1800 features, which may explain the increased error.

In our study, the number of samples is small, relative to the number of features. Therefore, some features are likely to be highly correlated. Using RFE, we found that ca. 2,000 out of ca. 70,000 features were sufficient to create RF model with comparable prediction power. The filtering here was only based on the feature importance, which shows the value of preprocessing by excluding low variance features or only keeping ubiquitous features to avoid unnecessary computational effort. The results from the RFE also relied on cross validation to protect the feature selection and model performance from overfitting. The reported error is based on the omitted folds and not on a separately kept test set. As the model performed similar to the other RFs with filtered data, we do not see proof of overfitting in the RFE.

We found a plateau in the required minimum number of features, an observation that was also seen in previous studies using RFE and random forests.[49] Our results using RFE agreed with the low NRMSE after low variance filtering. These findings lead us to advise using at least one form of filtering to counteract high-dimensionality. We did not analyze if the small number of features required is based on the sparse abundance or a lack of chemical relevance of molecular formulas was not analyzed. This will be the subject of a subsequent study.

4.3 Does the consideration of the chromatographic retention time improve the prediction?

In standard liquid chromatography, time-dependent data assumes that the time windows are independent of each other, i.e., the same molecular formula present at different retention times is considered two independent features. Our comparison of time-agnostic and time-aware experiments made it possible to validate the significance of chromatographic retention time. Time-aware features that included retention time improved the GLM performance (by around 0.24% NRMSE), but decreased the performance in most RF experiments. Overall, there was no statistical difference found between the prediction performance based on the time aware or time agnostic, although the retention time represents chemical information (polarity of the molecules) and allows a better separation of the complex organic material. When removing the time dimension, we traded better separation of compounds for an input matrix with less sparsity and an experimental setup with wider availability in the community. The adaptability of RFs to sparse data[38] could have reduced the observable differences between both data sets (time-agnostic and aware), but both were sparse. Still, predictions of RFs using on time-agnostic data were better, which we attribute to less overfitting due to the increased variance per feature and the general reduction in feature dimensions.

The signals of one structural formula have an elution window of around 30 seconds and a bell shaped intensity curve. We thus assume the majority of each structural formula to be in one retention time window, even though this is a simplification. Other approaches avoid

this by selecting specific retention time windows with gaps between them.[19]

4.4 Which key features are most important for a good prediction?

Computing SHAP values and permutation importance was suitable to isolate the important features in the complex mass spectrometry dataset. Our findings are particularly helpful for targeted analytical approaches that aim to identify structural information from the key features identified. Terrestrial MFs were characterized by lower masses compared to marine MFs, which matched the results of previous studies.[28,50] The average terrestrial N/C element ratio in previous work (Kong et al., submitted[51]) was lower than 0.001, in contrast to our results which showed an average N/C of 0.06 for terrestrial MFs identified by SHAP values in a sample with high terrestrial DOM contribution, and an average N/C of 0.08 for a sample with predominately marine DOM contribution. For S/C ratios a similar trend was observed.

The different ratios indicated vastly different molecular formulas compared to previous work (Kong et al., submitted[51]), even though the N/C and S/C ratios were consistently lower in terrestrial compared to marine DOM in both studies. If we compare the number of identified top 100 important features in 1,000 runs, the chance of a feature appearing in a model's top 100 features is low (less than 16%, ca. 600 unique features in the top 100 features). When we used time agnostic data, the number of identical, important features between models increased to 30%. One possible explanation is that time aware data can predict fluorescence well on each time slice but with different MFRTs. Another explanation is the reduced feature space. With roughly one third of the features in the time-agnostic data (23,835 compared to 70,683, around 33.7%), only around one third of the features are found in the 1000 repeated RF runs, indicating a similar spread of total features being deemed important. The similar trends between the permutation importance features and the SHAP value features was only partially supported, as they had a Jaccard similarity of 0.7 between the ca. 600 features of the top 100 ALR time-aware data. The time agnostic top features were even more similar, with a Jaccard similarity of 0.74. Previously reported stable groups

of similar features from rivers[52] were not discovered by our key feature search. Reasons for this are likely the longer distance in the ocean, as the samples from the MOSAiC cruise were not taken from the rivers directly but mostly from the Central Arctic Ocean. This likely allowed degradation of the MFs by degrading ultraviolet radiation[53] or microorganisms. The cruise was performed over a complete year and seasons were not a parameter we considered in our analysis, but changes in DOM were shown to be more dependent on region than on the season.[28]

All found key features only covered a small fraction of DOM. The direct measurement of ocean water with LC-FTMS reduced methodological bias of identified DOM but our models were also dependent on the PARAFAC component from the EEMs. As these EEMs measure the fluorescence, all found features and all trained models were blind to DOM that was not part of the small subset of fluorescent DOM. In a new study, C475 was correlated to $\delta^{13}C$ and salinity and indicated that 95% of the terrestrial MFs were also detected with the $\delta^{13}C$ and salinity approaches (Kong et al., in prep.). Supporting that C475 is a representative proxy for terrestrial DOM.

4.5 Tuning grid decisions and their influence on model performance

The performance and the key features changed, when the preprocessing and hyperparameters of the experiment changed. This is especially true for the tuning grid that is searched during training of the model. Due to the computational cost, especially for the RFs, we decided not to use a nested cross-validation. We tried to counteract the possible bias by creating similar training and test sets. Repeated training of cross validation hyperparameter searches for the high error GLMs showed that the mean error did not vary much between preprocessing and machine learning method combinations, indicating a successful split of the training and test set. The two examples, on which repeated cross validation was performed, showed results similar to the results from the heat maps, indicating that overfitting in these cases was not induced by the one-time split of training and test data. While RFs can cope with heigh

feature numbers and the GLM is regularized by an elastic net, the SVMs do not have such mechanics. The SVMs where overfitting occurred on the unfiltered data, likely originated from a suboptimal choice of kernels. Our choice of polynomial and radial basis function kernels was driven by their widespread use and ability to capture non-linear relationships. A previous study showed that linear kernels perform better for cases where overfitting is a problem.[54]

One of the key aspects for RFs was the *mtry* parameter. A larger *mtry* value was suitable for datasets with large parameter counts. This led to more focused key features, as the few stronger predictors appeared more often in the selected subset.[55] Limiting the *mtry* values to a smaller percentage of features may have been more suitable to find the more subtle features. With MS data, we therefore may not want to tune *mtry* to very high values to avoid reducing the importance of many lesser obvious features. A strategy for future work could be to perform the model training with different *mtry* value limitations to compare the important features in the data. For generalized linear models, this could be tuned via the *alpha* parameter. With fewer feature weights being reduced to zero, the weights would be spread wider across the data. Instead of automated feature selection by the elastic net,[56] selecting a fixed *alpha* value and tuning *lambda* may be beneficial as a feature selection method.

We also considered GLMs, since they are the fastest to tune, allow easy usage of the weights as importance measurement and automatic feature selection by changing the elastic net which also help against overfitting. This was mostly achieved with the exception of the ALR tranformation, revealing the necessity to evaluate normalization methods. The repeated cross-validations also showed the cases of overfitting were no outliers and a stable performance in the case of SUM normalization was achieved. For the lasso regression in GLMs, all features should be of the same scale to avoid introducing a bias. This was not done by rescaling the features between the normalization and the model training step (*preprocess* in the package caret). The found key features were not the most intense features (Fig. S5)

as the selected weights are not from the features with the highest mean abundance.

5 Conclusions

Several state-of-the-art Machine Learning methods were applied in order to predict the PARAFAC component C475, which indicates terrestrial contribution in DOM, based on a combination of molecular formulas and their retention time. The MF and RT measurements were obtained via modern LC-FTMS, which most recently can be applied directly to saltwater samples. Using our methods, we were able to predict our proxy with an NRMSE of only 5.7% of the original scale of the proxy. Moreover, the MFs that were predicted as terrestrial were different from those that were identified with prior methods *not* taking into account the RT values. Our methods were thus able to produce new chemical insights show similar trends to the literature whilst giving a guideline for ML-based approaches to DOM problems using LC-FTMS: For an improved model performance, we generally suggest filtering the data at least for MFs with low variance. For normalization we recommend not using the ALR normalization and for time aware/time agnostic data, we found no preference. We suggest RFs for data that is not filtered at all or approaches that require non-linear assumptions and GLM for faster computation times.

In the future, it would be desirable to investigate more closely the chemical characteristics of those MFs that are the best predictors for terrestrial DOM. Is it possible, using our methods, to further narrow down the search for the *actual* molecules that are represented by the MFs? As it turns out, our methods are able to make good predictions using only around 2,000 (as compared to the original over 70,000 features). It would be interesting to further investigate these 2,000 important features. How can they be characterized? Is it possible to combine or reduce these features further to a much smaller set of features, while still being able to make good predictions?

Can our models be applied to water samples stemming from other regions of the ocean,

in particular to regions that are far less terrestrial than the Arctic ocean (e.g. the Antarctic ocean)? It is an intriguing open question whether the identical set of important features will be applicable to such samples, which will be addressed in a follow-up study.

6 Data and Software Availability

The used EEM data and PARAFAC components are available at: https://doi.pangaea.de/10.1594/PANGAEA.948019

The LC-FTMS data is part of another publication and will be available under the following link when that publication is accepted: https://doi.org/10.6084/m9.figshare.29210642

The code can be made available upon request.

Acknowledgement

The authors thank Prof. Dr. Christian L. Müller for his input regarding GLMs. The first author is funded through the Helmholtz School for Marine Data Science (MarDATA), Grant No. HIDSS-0005.

Supporting Information Available

The following files are available free of charge.

- Supporting information: Boxplot comparing the 500 RFs, three plots regarding the feature consistency in RFs over 1000 repeats, a Boxplot of the SHAP value distibution over the mass, as well as Heatmaps for the R^2 and mean absolute error for low variance including and excluding experiments.

References

(1) Coble, P. G.; Green, S. A.; Blough, N. V.; Gagosian, R. B. Characterization of dissolved organic matter in the Black Sea by fluorescence spectroscopy. *Natur* **1990**, *348*, 432–435.

(2) Huguet, A.; Vacher, L.; Relexans, S.; Saubusse, S.; Froidefond, J.; Parlanti, E. Properties of fluorescent dissolved organic matter in the Gironde Estuary. *Organic Geochemistry* **2009**, *40*, 706–719.

(3) Kalbitz, K.; Solinger, S.; Park, J. H.; Michalzik, B.; Matzner, E. Controls on the dynamics dissolved organic matter in soils: A review. *Soil Science* **2000**, *165*, 277–304.

(4) Pörtner, H. O.; Roberts, D. C.; Masson-Delmotte, V.; Zhai, P.; Tignor, M.; Poloczanska, E.; Mintenbeck, K.; Alegría, A.; Nicolai, M.; Okem, A.; Petzold, J.; Rama, B.; Weyer, N. M. The Ocean and Cryosphere in a Changing Climate: Special Report

of the Intergovernmental Panel on Climate Change. *The Ocean and Cryosphere in a Changing Climate: Special Report of the Intergovernmental Panel on Climate Change* **2022**, 1–756.

(5) Semiletov, I. P.; Pipko, I. I.; Shakhova, N. E.; Dudarev, O. V.; Pugach, S. P.; Charkin, A. N.; Mcroy, C. P.; Kosmach, D.; Gustafsson, Ö. Carbon transport by the Lena River from its headwaters to the Arctic Ocean, with emphasis on fluvial input of terrestrial particulate organic carbon vs. carbon transport by coastal erosion. *Biogeosciences* **2011**, *8*, 2407–2426.

(6) Thoman, R. L.; Moon, T. A.; M. L. Druckenmiller, E. NOAA Arctic Report Card 2023. 2023.

(7) Lechtenfeld, O. J.; Kaesler, J.; Jennings, E. K.; Koch, B. P. Direct Analysis of Marine Dissolved Organic Matter Using LC-FT-ICR MS. *Environmental Science & Technology* **2024**, *58*, 4637–4647.

(8) Leefmann, T.; Frickenhaus, S.; Koch, B. P. UltraMassExplorer: a browser-based application for the evaluation of high-resolution mass spectrometric data. *Rapid Communications in Mass Spectrometry* **2019**, *33*, 193–202.

(9) Laane, R.; Koole, L. The relation between fluorescence and dissolved organic carbon in the Ems-Dollart estuary and the Western Wadden Sea. *Netherlands Journal of Sea Research* **1982**, *15*, 217–227.

(10) Perdue, E. M.; Benner, R. *Biophysico-Chemical Processes Involving Natural Nonliving Organic Matter in Environmental Systems*; John Wiley & Sons, Ltd, 2009; Chapter 11, pp 407–449.

(11) Stedmon, C. A.; Markager, S.; Bro, R. Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy. *Marine Chemistry* **2003**, *82*, 239–254.

(12) Stedmon, C. A.; Bro, R. Characterizing dissolved organic matter fluorescence with parallel factor analysis: a tutorial. *Limnology and Oceanography: Methods* **2008**, *6*, 572–579.

(13) Parlanti, E.; Wörz, K.; Geoffroy, L.; Lamotte, M. Dissolved organic matter fluorescence spectroscopy as a tool to estimate biological activity in a coastal zone submitted to anthropogenic inputs. *Organic Geochemistry* **2000**, *31*, 1765–1781.

(14) Kong, X.; Jendrossek, T.; Ludwichowski, K.-U.; Marx, U.; Koch, B. P. Solid-Phase Extraction of Aquatic Organic Matter: Loading-Dependent Chemical Fractionation and Self-Assembly. *Environmental Science & Technology* **2021**, *55*, 15495–15504.

(15) Wünsch, U. J.; Geuer, J. K.; Lechtenfeld, O. J.; Koch, B. P.; Murphy, K. R.; Stedmon, C. A. Quantifying the impact of solid-phase extraction on chromophoric dissolved organic matter composition. *Marine Chemistry* **2018**, *207*, 33–41.

(16) Yi, Y.; Liu, T.; Merder, J.; He, C.; Bao, H.; Li, P.; Li, S.; Shi, Q.; He, D. Unraveling the Linkages between Molecular Abundance and Stable Carbon Isotope Ratio in Dissolved Organic Matter Using Machine Learning. *Environmental Science & Technology* **2023**, *57*, 17900–17909, PMID: 37079797.

(17) Liu, J.; Wang, C.; Zhou, J.; Dong, K.; Elsamadony, M.; Xu, Y.; Fujii, M.; Wei, Y.; Wang, D. Thermodynamics and explainable machine learning assist in interpreting biodegradability of dissolved organic matter in sludge anaerobic digestion with thermal hydrolysis. *Bioresource Technology* **2024**, *412*, 131382.

(18) Zhao, C.; Wang, K.; Jiao, Q.; Xu, X.; Yi, Y.; Li, P.; Merder, J.; He, D. Machine Learning Models for Evaluating Biological Reactivity Within Molecular Fingerprints of Dissolved Organic Matter Over Time. *Geophysical Research Letters* **2024**, *51*, e2024GL108794.

(19) Gad, M.; Khomami, N. T. S.; Krieg, R.; Schor, J.; Philippe, A.; Lechtenfeld, O. J. Environmental drivers of dissolved organic matter composition across central European

aquatic systems: A novel correlation-based machine learning and FT-ICR MS approach. *Water Research* **2025**, *273*, 123018.

(20) Herzsprung, P.; Wentzky, V.; Kamjunke, N.; Tümpling, W. V.; 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. *Environmental Science and Technology* **2020**, *54*, 13556–13565.

(21) Zhao, C.; Xu, X.; Chen, H.; Wang, F.; Li, P.; He, C.; Shi, Q.; Yi, Y.; Li, X.; Li, S.; He, D. Exploring the Complexities of Dissolved Organic Matter Photochemistry from the Molecular Level by Using Machine Learning Approaches. *Environmental Science & Technology* **2023**, *57*, 17889–17899.

(22) Müller, M.; D'Andrilli, J.; Silverman, V.; Bier, R. L.; Barnard, M. A.; Lee, M. C. M.; Richard, F.; Tanentzap, A. J.; Wang, J.; de Melo, M.; Lu, Y. Machine-learning based approach to examine ecological processes influencing the diversity of riverine dissolved organic matter composition. *Frontiers in Water* **2024**, *6*.

(23) Cuss, C. W.; McConnell, S. M.; Guéguen, C. Combining parallel factor analysis and machine learning for the classification of dissolved organic matter according to source using fluorescence signatures. *Chemosphere* **2016**, *155*, 283–291.

(24) Nguyen, X. C.; Seo, Y.; Park, H. Y.; Begum, M. S.; Lee, B. J.; Hur, J. Tracking the sources of dissolved organic matter under bio-and photo-transformation conditions using fluorescence spectrum-based machine learning techniques. *Environmental Technology & Innovation* **2023**, *31*, 103179.

(25) Cha, D.; Park, S.; Kim, M. S.; Kim, T.; Hong, S. W.; Cho, K. H.; Lee, C. Prediction of Oxidant Exposures and Micropollutant Abatement during Ozonation Using a Machine Learning Method. *Environmental Science and Technology* **2021**, *55*, 709–718.

(26) Fong, A. A. et al. Overview of the MOSAiC expedition: Ecosystem. *Elementa: Science of Anthropocene* **2024**, *12*.

(27) Stubbins, A.; Lapierre, J.-F.; Berggren, M.; Prairie, Y. T.; Dittmar, T.; del Giorgio, P. A. What's in an EEM? Molecular Signatures Associated with Dissolved Organic Fluorescence in Boreal Canada. *Environmental Science & Technology* **2014**, *48*, 10598–10606.

(28) Kong, X.; Granskog, M. A.; Hoppe, C. J. M.; Fong, A. A.; Stedmon, C. A.; Tippenhauer, S.; Ulfsbo, A.; Vredenborg, M.; Koch, B. P. Variability of Dissolved Organic Matter Sources in the Upper Eurasian Arctic Ocean. *Journal of Geophysical Research: Oceans* **2024**, *129*, e2023JC020844.

(29) Kong, X. Molecular and optical characterization of dissolved organic matter in the Central Arctic Ocean. Ph.D. thesis, University of Bremen, 2022.

(30) Bro, R. Chemometrics and intelligent laboratory systems Tutorial PARAFAC. Tutorial and applications. *Chemomemcs and Intelligent Laboratory Systems* **1997**, *38*, 149–171.

(31) Harshman, R. Foundations of the PARAFAC procedure: Model and conditions for an explanatory factor analysis. *Technical Report UCLA Working Papers in Phonetics 16* **1970**, 1–84.

(32) Jolliffe, I. T. *Principal Component Analysis*; Springer New York: New York, NY, 2002; pp 338–372.

(33) Coble, P. G. Characterization of marine and terrestrial DOM in seawater using excitation-emission matrix spectroscopy. *Marine Chemistry* **1996**, *51*, 325–346.

(34) Kind, T.; Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **2007**, *8*, 1–20.

(35) Nelder, J. A.; Wedderburn, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)* **1972**, *135*, 370–384.

(36) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

(37) Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. Advances in Neural Information Processing Systems. 1996; pp 155–161.

(38) Biau, G. Analysis of a Random Forests Model. *J. Mach. Learn. Res.* **2012**, *13*, 1063–1095.

(39) Ruescas, A. B.; Hieronymi, M.; Mateo-Garcia, G.; Koponen, S.; Kallio, K.; Camps-Valls, G. Machine Learning Regression Approaches for Colored Dissolved Organic Matter (CDOM) Retrieval with S2-MSI and S3-OLCI Simulated Data. *Remote Sensing* **2018**, *10*.

(40) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* **2019**, *16*, 299–302.

(41) Beck, A. G.; Muhoberac, M.; Randolph, C. E.; Beveridge, C. H.; Wijewardhane, P. R.; Kenttämaa, H. I.; Chopra, G. Recent Developments in Machine Learning for Mass Spectrometry. *ACS Measurement Science Au* **2024**, *4*, 233–246.

(42) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* **2017**, *2017-December*, 4766–4775.

(43) R Core Team R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2024.

(44) D'Andrilli, J.; Dittmar, T.; Koch, B. P.; Purcell, J. M.; Marshall, A. G.; Cooper, W. T. Comprehensive characterization of marine dissolved organic matter by Fourier trans-

form ion cyclotron resonance mass spectrometry with electrospray and atmospheric pressure photoionization. *Rapid Communications in Mass Spectrometry* **2010**, *24*, 643–650.

(45) Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning: data mining, inference and prediction*, 2nd ed.; Springer, 2009; pp 596–604.

(46) Collart, F.; Guisan, A. Small to train, small to test: Dealing with low sample size in model evaluation. *Ecological Informatics* **2023**, *75*, 102106.

(47) Verleysen, M.; François, D. The Curse of Dimensionality in Data Mining and Time Series Prediction. Computational Intelligence and Bioinspired Systems. Berlin, Heidelberg, 2005; pp 758–770.

(48) Altman, N.; Krzywinski, M. The curse(s) of dimensionality. *Nature Methods* **2018**, *15*, 399–400.

(49) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 1947–1958.

(50) Kujawinski, E. B.; Longnecker, K.; Blough, N. V.; Vecchio, R. D.; Finlay, L.; Kitner, J. B.; Giovannoni, S. J. Identification of possible source markers in marine dissolved organic matter using ultrahigh resolution mass spectrometry. *Geochimica et Cosmochimica Acta* **2009**, *73*, 4384–4399.

(51) et al., X. K. Terrestrial dissolved organic carbon budget in the Arctic Ocean.

(52) Behnke, M. I. et al. Pan-Arctic Riverine Dissolved Organic Matter: Synchronous Molecular Stability, Shifting Sources and Subsidies. *Global Biogeochemical Cycles* **2021**, *35*, e2020GB006871, e2020GB006871 2020GB006871.

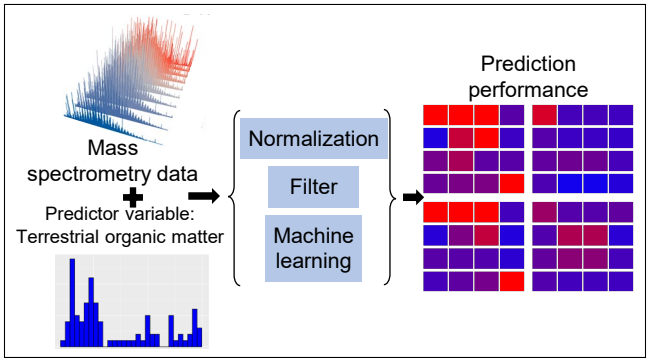
(53) Erickson, D. J.; Sulzberger, B.; Zepp, R. G.; Austin, A. T. Effects of stratospheric ozone depletion, solar UV radiation, and climate change on biogeochemical cycling: Interactions and feedbacks. *Photochemical and Photobiological Sciences* **2015**, *14*, 127–148.

(54) Han, H.; Jiang, X. Overcome Support Vector Machine Diagnosis Overfitting. *Cancer Informatics* **2014**, *13s1*, CIN.S13875, PMID: 25574125.

(55) Probst, P.; Wright, M. N.; Boulesteix, A. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery* **2019**, *9*.

(56) Zou, H.; Hastie, T. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **2005**, *67*, 301–320.

TOC Graphic


Mass
spectrometry data
Predictor variable:
Terrestrial organic matter
Normalization
Filter
Machine
learning
Prediction
performance