Development and application of the Branched and Isoprenoid GDGT Machine learning Classification algorithm (BIGMaC) for paleoenvironmental reconstruction

Pablo Martínez-Sosa[1], Jessica E. Tierney[1], Lina C. Pérez-Angel[2], Ioana C. Stefanescu[3], Jingjing Guo[4], Frédérique Kirkels[4], Julio Sepúlveda[5], Francien Peterse[4], Bryan N. Shuman[3], Alberto V. Reyes[6]

[1]Department of Geosciences, The University of Arizona, Tucson, Arizona, USA
[2]Institute at Brown for Environment and Society (IBES), Brown University, Rhode Island, USA
[3]Department of Geology and Geophysics, University of Wyoming, Wyoming, USA
[4]Department of Earth Sciences, Utrecht University, Utrecht, Netherlands
[5]Department of Geological Sciences and Institute of Arctic and Alpine Research (INSTAAR), University of Colorado, Colorado, USA
[6]Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, Alberta, Canada

Key Points:

- The distribution of GDGTs is particular to each depositional environment and has unique responses to environmental factors.
- The BIGMaC algorithm captures the relationship between both branched and isoprenoid GDGTs with depositional environments.
- Our approach can provide paleoclimatological and paleoenvironmental information based only on GDGTs.

Corresponding author: Pablo Martínez-Sosa, pmartoza@arizona.edu

Abstract

Glycerol dialkyl glycerol tetraethers (GDGTs), both archaeal isoprenoid GDGTs (isoGDGTs) and bacterial branched GDGTs (brGDGTs), have been used in paleoclimate studies to reconstruct environmental conditions. Since GDGTs are produced in many types of environments, their relative abundances also depend on the depositional setting. This suggests that the distribution of GDGTs also preserves useful information that can be used more broadly to infer these depositional environments in the geological past. Here, we combined existing iso- and brGDGT relative abundance data with newly analyzed samples to generate a database of 1153 samples from several modern sedimentary settings. We observed a robust relationship between the depositional environment and the relative abundances of GDGTs in our samples. This dataset was used to train and test the **B**ranched and **I**soprenoid **G**DGT **Ma**chine learning **C**lassification (BIGMaC) algorithm, which identifies the environment a sample comes from based on the distribution of GDGTs with high precision and recall (F1 = 0.95). We tested the model on the sedimentary record from the Giraffe kimberlite pipe, an Eocene maar in subantarctic Canada, and found that the BIGMaC reconstruction agrees with independent stratigraphic and palynological information, provides new information about the paleoenvironment of this site, and helps improve its paleotemperature reconstruction. In contrast, we also include an example from the PETM-aged Cobham lignite as a cautionary example that illustrates the limitations of the algorithm. We propose that in cases where paleoenvironments are unknown or are changing, BIGMaC can be applied in concert with other proxies to generate more refined paleoclimate records.

1 Introduction

Glycerol dialkyl glycerol tetraethers (GDGTs) are membrane-spanning lipids found in sediments and soils around the world. There are two main types of these molecules, branched and isoprenoid. Branched glycerol dialkyl glycerol tetraethers (brGDGTs) are characterized by their branched alkyl chains, with a differing number (4 – 6) and position (5-methyl or 6-methyl) of methyl groups and cyclopentane moieties (0 – 2). This unique structure defies the classical evolutionary dichotomy of the lipid divide by combining traits of Bacteria and Archaeal cell membranes (Weijers et al., 2006). Based on evidence including the presence of alkyl chains, the stereochemistry of the glycerol group (Weijers et al., 2006), and most importantly, microbial culture studies (Y. Chen et al., 2022; Halamka et al., 2022, 2021; Sinninghe Damsté et al., 2011), brGDGTs have a bacterial source.

In contrast, isoprenoid glycerol dibiphytanyl glycerol tetraether GDGTs (isoGDGTs) are produced by Archaea (de Rosa et al., 1977; Sinninghe Damsté et al., 2002). Their structures contain two phytane chains (Langworthy, 1977) and vary in the number of cyclopentane moieties (0 – 8) (De Rosa et al., 1983). Crenarchaeol is a member of this group of particular importance as it has been shown to be specifically produced by Thaumarchaeota (Sinninghe Damsté et al., 2002). Crenarchaeol contains four cyclopentane rings, one cyclohexane ring, and has one identified stereoisomer known as crenarchaeol' (Sinninghe Damsté et al., 2002, 2018).

Both isoprenoid and branched GDGTs are used in paleoclimate studies as their distribution is correlated with variables such as temperature and pH, and these molecules are relatively stable through the geological record. In marine sediments, the degree of cyclization of isoGDGTs is related to overlying water temperature, forming the basis of the TetraEther indeX of 86 carbons (TEX_{86}) proxy (Schouten et al., 2002, 2013). Similarly, the methylation, cyclization, and isomerization of brGDGTs have been shown to respond to temperature and pH in terrestrial environments, such as peats, soils, lakes, and rivers (Raberg et al., 2022; Martínez-Sosa et al., 2020; Dang et al., 2018; De Jonge, Stadnitskaia, et al., 2014; Tierney et al., 2010; Weijers et al., 2007). The Methylation

index of Branched Tetraethers (MBT'_{5Me}) proxy isolates the relationship between the methylation of brGDGTs and temperature (Weijers et al., 2007; De Jonge, Hopmans, et al., 2014) and has been widely used for terrestrial paleoclimate reconstructions (De Jonge, Hopmans, et al., 2014; Zheng et al., 2017; Naafs et al., 2018; Lauretano et al., 2021; Zhao et al., 2022).

Across environments, GDGT distributions broadly reflect the microbial community present. This is, for example, the basis of the Methane Index, which measures the contribution of methanotrophic archaea relative to marine Thaumarchaeota in the sedimentary isoGDGT pool (Zhang et al., 2011). Likewise, the distribution of isoGDGTs in marine systems reflects not only sea-surface temperature (captured by the TEX_{86} index), but also the water depth (and potentially, different archaeal communities) from which the isoGDGTs derive (Rattanasriampaipong et al., 2022; Taylor et al., 2013). Furthermore, previous work has found that the ratio of crenarchaeol/crenarchaeol' can indicate which *Thaumarchaeota* group (I.1a or I.1b) is responsible for the production of these lipids in lake sediments (Li et al., 2016). In terrestrial settings, De Jonge et al. (2019) proposed the Community Index for brGDGTs, which is based on the inference that brGDGTs are produced by different communities of bacteria, each with a unique response to soil temperature. The combined use of some of the GDGTs, through the Branched and Isoprenoid Tetraether (BIT) index, is used to broadly discriminate between marine and terrestrial environments based on the dominance of brGDGT-producing bacteria in most terrestrial settings and crenarchaeol-producing Thaumarchaeota in marine settings (Hopmans et al., 2004). However, BIT values in soils, lakes, and peats all tend to be high, which limits the ability of this index to reliably distinguish between these different types of terrestrial settings.

Building on these observations, we posit that the full range of archaeal and bacterial GDGTs (isoprenoidal and branched) contains information about their biological precursors and the overall composition of the microbial community. This information, based on ecological interactions rather than a physiological response, can in turn be used to discriminate between sediments deposited in terrestrial or marine environments, as well as whether terrestrial sediments are derived from freshwater, soil, or peatland environments. This would provide an additional tool for the identification of ancient depositional conditions in instances when it is not clear what the environment was, and therefore could inform the reconstruction of environmental variables, *i.e*, which GDGT-based temperature proxy and calibration is most appropriate to use. Machine learning provides a way to model highly dimensional and nonlinear data with complex interactions and missing values (El Bouchefry & de Souza, 2020). In this case it allows us to investigate drivers of the abundance of 19 GDGT structures (high dimensionality) which have a complex and non-linear relationship to the environment, making this an ideal approach to extract the environmental information present in the distribution of GDGTs.

Machine learning has previously been used in the Geosciences to discriminate between magma (Ueki et al., 2018) as well as identifying the source of water from oil wells (Engle & Brunner, 2019). In the field of biomarker-based paleoclimatology, classification algorithms have been applied to identify sources of alkenones (Zheng et al., 2019), as well as plant waxes (Peaple et al., 2021). Machine learning regression algorithms, as well, as deep neural network applications, have also been applied to GDGTs as in order to generate temperature calibrations (Dunkley Jones et al., 2020; Véquaud et al., 2022; Zheng et al., 2022). Here, we use a compilation of GDGT distributions in 1153 globally distributed soils, peats, and sediments from diverse depositional environments to train a classification algorithm which is capable of identifying the environment in which a sample was formed based on the distribution of both branched and isoprenoid GDGTs. We then demonstrate the application of this algorithm by using it to interpret the paleoenvironment and the paleotemperature in a Paleogene deposit that records a transition from a lacustrine to a peatland environment. We also highlight the limitations of this approach

in an application to a peatland dataset that spans the onset of the Paleocene-Eocene Thermal Maximum (PETM).

2 Materials and Methods

2.1 Global Dataset

We compiled a total of 1153 globally distributed (Fig. 1 and Table SI1) samples from different depositional environments: marine, lake, peat, river, and soil. These samples all have quantified relative abundances for the full suite of the most commonly used isoGDGTs (GDGT-0, GDGT-1, GDGT-2, GDGT-3, crenarchaeol, and crenarchaeol') and brGDGTs (IIIa, IIIa', IIIb, IIIb', IIa, IIa', IIb, IIb', IIc, IIc', Ia, Ib, and Ic) in paleoenvironmental reconstructions, and were all analyzed with the updated High Performance Liquid Chromatography-Mass Spectrometry (HPLC-MS) method of Hopmans et al. (2016). From the 1153 samples, 475 are peat (Naafs et al., 2018), 215 are marine sediments (this study), 196 are soil (Guo, Ma, et al., 2022; Dearing Crampton-Flood et al., 2020; Guo et al., 2020; Pérez-Angel et al., 2020), 162 are lake sediments (Martínez-Sosa et al., 2021; Guo et al., 2020), and 105 are riverbed sediments (Kirkels, Usman, & Peterse, 2022). For the Colombian and Inner Mongolia soil samples (Guo, Ma, et al., 2022; Pérez-Angel et al., 2020) we include here newly reported isoGDGT values not part of the original dataset.

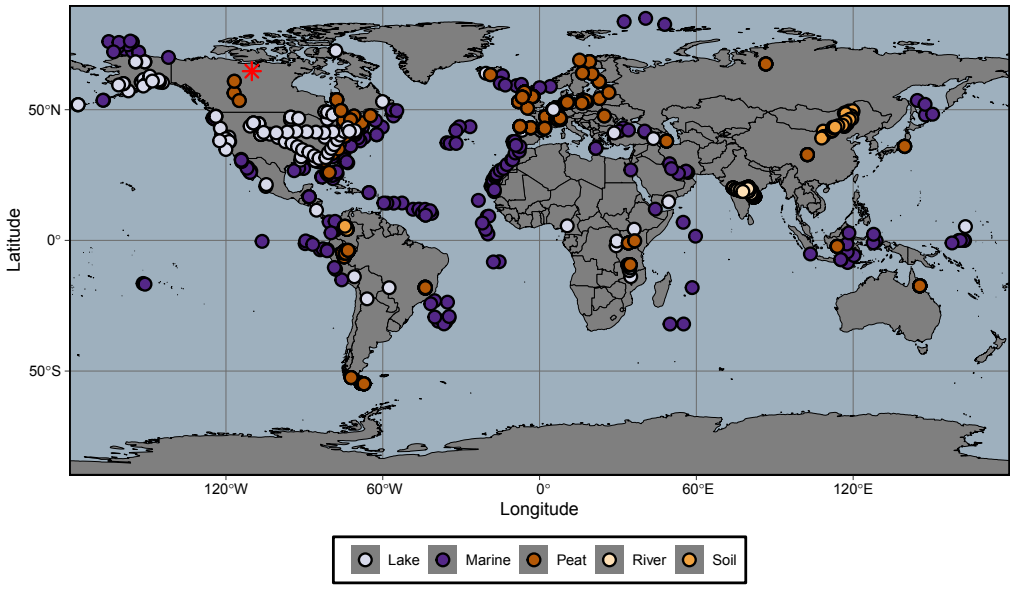

Figure 1. World map showing the distribution of the samples included in this work. Color code reflects the depositional environment which these samples were collected from. Red asterisk shows the modern location of the Giraffe pipe.

All marine sediments were processed at the University of Arizona following the method used in Martínez-Sosa et al. (2021). Briefly, sediments were freeze-dried, homogenized, and spiked with a C_{46} internal standard before being extracted with an Accelerated Solvent Extraction (ASE) system (run at 1500 psi, 100°C, with dichloromethane:methanol (DCM: MeOH, 9:1)). Total Lipid Extracts (TLEs) were eluted through a deactivated SiO_2 column with hexane:ethyl acetate (1:2) to obtain the polar fraction, and the elutant was dried under a N_2 stream. Polar fractions were redissolved in hexane:isopropanol (99:1), and then passed through a 0.45 μm PTFE filter prior to being analyzed by HPLC-

MS. GDGTs, isoprenoid and branched, were analyzed on an Agilent 1260 Infinity HPLC coupled to an Agilent 6120 single quadrupole mass spectrometer using two BEH HILIC silica columns (2.1×150 mm, 1.7 μm; Waters) following the methodology of Hopmans et al. (2016). We calculated peak areas using the MATLAB package ORIGAmI (Fleming & Tierney, 2016) and quantified brGDGTs by comparing the obtained peaks with the internal standard (Huguet et al., 2006).

For all samples in this dataset we calculated the relative abundance of all brGDGTs (except IIIc and IIIc', due to their general low abundance), as well as isoGDGTs 0–3, crenarchaeol, and its isomer. For all the analyses we used the fractional abundance of each compound relative to the total sum of GDGTs (branched + isoprenoid), to account for the difference in the relative abundances of both GDGT types among the depositional environments. Although it is known that the ionization of isoGDGTs and brGDGTs in the MS might be different between laboratories (Schouten et al., 2013), the potential impact of this is minimized in our statistical approach because the data are normalized before applying the machine learning techniques (see Section 2.2).

We collected the environmental parameters associated with the samples using the data available in the source datasets (Table S1). For the marine sediments analyzed for this study, we obtained mean annual temperature of the top 200m of the water column from the World Ocean Atlas 2018 (Locarnini et al., 2018).

2.2 Unsupervised Machine Learning

For the unsupervised machine learning analysis we centered and scaled the fractional abundances of GDGTs across the whole dataset. The optimal number of clusters for this dataset was calculated through a silhouette analysis, which calculates how similar a data point is within-cluster compared to other clusters. This analysis was performed by using the the Partitioning Around Medoids method from the *cluster* R package (Maechler et al., 2019).

Samples were separated into clusters by applying the fuzzy version of the k-means clustering algorithm using the *e1071* R package (Meyer et al., 2020). This method calculates the degree to which each sample belongs to each of the clusters (membership value) instead of assigning a single classification; we consider this a useful tool to classify depositional environments, which have diffuse boundaries. The fuzzy k-means analysis was performed using the best performing number of clusters from the silhouette analysis, with all other parameters of the function at default values.

Following the cluster analysis, we compared the cluster assignment of each sample to the available information on their environmental data and depositional environment. Combining both the statistical results and the observation-based information we assigned one of four new labels to each of the samples.

2.3 Supervised Machine Learning

For the supervised machine learning we worked in the *tidymodels* and *tidyverse* R environments (Kuhn & Wickham, 2020; Wickham et al., 2019), where we used the fractional abundances of GDGTs as predictor variables and the statistical and observation based labels as the response variable. The dataset was split into a training and testing set in a 3:1 ratio. To avoid subsampling the dataset in a biased manner, we preserved the distribution of sample types in both sets. We tested the performance of four different algorithms commonly used for classification applications: Random Forest, eXtreme Gradient Boosting (XGBoost), K-Nearest Neighbour and Naive Bayes plus a control non-informative (null) model. Below we present a brief overview of each of them.

Random Forest is an ensemble classification algorithm, where classification of a sample is based on the voting results of an ensemble of independent decision trees. Importantly, each tree is presented with a randomly selected subset of samples and predictor variables, ensuring that the trees generate independent results. This enables the result from the voting to account for biases present in each individual decision tree. Random Forest algorithms are considered reliable and fast (Parmar et al., 2019).

XGBoost is a gradient tree boosting algorithm. In this case, similar to Random Forest, the algorithm uses a series of trees to classify samples. However, XGBoost iteratively improves the performance of this process by improving on the results of each previous iteration. This algorithm is particularly good for biased data and has been a prefered algorithm for diverse applications due to its scalability (T. Chen & Guestrin, 2016).

Naive Bayes is a statistical algorithm, in which Bayes Theorem is applied to calculate the posterior probability of a new sample belonging to each possible categorical classification. The algorithm determines the classification by choosing that with the highest probability. Although this is a fast classification algorithm it is a relatively bad estimator (Sen et al., 2020).

K-nearest neighbor is another statistics-based method, where samples are classified as the voting result of their k closest neighbors. Neighbors are given priority based on their closeness to the new data. While this is an effective algorithm for large datasets, it can be computationally expensive as all the distances from the neighbors need to be calculated for all new data (Sen et al., 2020).

For all algorithms the hyperparameter values—parameters whose values control the learning process but are not part of the final model (*i.e.* number of independent trees used for a Random Forest)—were selected (tuned) as those with the best performance from a distribution of possible combination of values for each hyperparameter. Due to the required computing power for the hyperparameter tuning step, this was run using a High-Performance Computing cluster. Finally, the best hyperparameter values were selected by comparing their Receiver Operating Characteristic Area Under the Curve (ROC-AUC) score on the validation set (Table S1). ROC-AUC is a metric that measures the tradeoff between the true positive rate and false positive rate of the model, for this parameter higher values on the (0,1) range are desirable. An additional metric, F1, was applied to evaluate the performance of all the trained algorithms. F1 is calculated as the mean of the precision and recall—the ability to identify true positives—of the model. For reproducibility, a detailed script with the packages and parameter values used for this process is available on GitHub (Martínez-Sosa et al., 2023). Finally we can identify which GDGTs contribute the most to the classification process through the the importance metric. This value is calculated based on how much each GDGT contributes to decreasing the probability of incorrectly classifying a sample across all decision trees (Gini importance) (Wright et al., 2019; Greenwell et al., 2020). This metric shows how often a GDGT was selected to split the data and what was its discriminative power (Menze et al., 2009), larger values indicate a better variable to separate the data.

For this work, all analyses were performed in `R` (v. 4.1.3) (R Core Team, 2022). Additional dimensionality reduction analyses were done over the fractional abundances of all GDGTs using Principal Component Analysis (PCA) through the `princomp()` function from base R. For the PCA analyses the loadings of variables were visually inspected and corroborated by obtaining the eigenvector values. Correlation between environmental parameters for each sample and their scores on the principal components was performed by applying the `cor.test()` function from the *psych* R package using a Spearman correlation (Revelle & Revelle, 2015).

2.4 Giraffe Kimberlite Pipe

We analyzed GDGTs from 83 samples from diamond exploration drill core BHP 99-01 from the Giraffe kimberlite pipe (paleolatitude $\sim 63°$N) (Wolfe et al., 2017). This core is stored at the Geological Survey of Canada core repository (Calgary), and it contains ≥ 50 vertical-equivalent meters of lacustrine sediment topped with ~ 32 m of peat. The sediments were dated to 37.84±1.99 Ma by glass fission-track dated rhyolitic tephra beds (Wolfe et al., 2017). Our dataset spans 83.5 vertical-equivalent meters and includes 19 samples from the peat section and 64 from the lacustrine section. For each sample, between 0.5 and 1 g of sediment was processed to obtain TLEs in the same manner as for the marine samples. For these samples, the GDGTs were isolated using a two-layer chromatography column filled with a 1:1 mix of LC-NH_2 (bottom layer) and 5% deactivated silica (top layer) gels as the solid phase (Windler et al., 2019). The GDGTs were recovered using dichloromethane:isopropanol (2:1) as the solvent. Branched and isoprenoid GDGTs were analyzed in all samples using the same HPLC-MS method described for the marine samples in section 2.1.

2.5 Cobham Lignite Bed

The Cobham lignite bed, Kent, UK ($\sim 48°$N palaeolatitude) is composed of a sand and mud unit at the base, overlain, in succession, by a charcoal-rich lower laminated lignite, a charcoal-poor upper laminated lignite, a middle clay layer, and a charcoal-poor blocky lignite. The Woolwich Shell Beds overly the Cobham Lignite (Collinson et al., 2009). A carbon isotope excursion is present near the top of the charcoal-poor upper laminated lignite, which is interpreted as being the characteristic excursion from the Paleocene Eocene Thermal Maximum (PETM, ~ 56 million years ago). Collinson et al. (2009) interpreted the units above this as representing the early part of the PETM. We tested our algorithm on the 27 samples obtained from this site previously analyzed by Inglis et al. (2019) and publicly available at the PANGAEA data repository (Inglis et al., 2019).

3 Results

3.1 Fuzzy K-means Classification

Our silhouette analysis showed that the global GDGT data is best separated into four clusters, a value which was then used to perform a fuzzy k-means classification. The four identified groups consist of between 219 and 465 samples each. When we compare the composition of each cluster using PCA, there is a clear differences between depositional environments (Fig. 2a and b, and Table 1). 86.9% of the peat samples fall within Group 1, while 84.6% of the lacustrine samples are assigned to Group 2. In turn, 92.4% of the river samples are assigned to Group 3, and 91.6% of the marine samples are assigned to Group 4 (Fig. 2a and b). Soil samples are more spread across the different groups, with the majority assigned to Group 3 (43.9%).

3.2 Within-Group Analyses

We analyzed the GDGT distribution within each of the clusters identified in the unsupervised machine learning step to assess its influence on the clustering results and how well it correlated with environmental parameters.

3.2.1 GDGT Distribution

Across the entire dataset crenarchaeol', GDGT-1–GDGT-3, Ib, Ic, IIc, IIc', IIIb, and IIIb' have the smallest proportion (< 0.1 fractional abundance) of all GDGTs (Fig. 3). There are, however, characteristic patterns associated with the different clusters. Samples from cluster 4 have a higher proportion of crenarchaeol, GDGT-0, and to a lesser

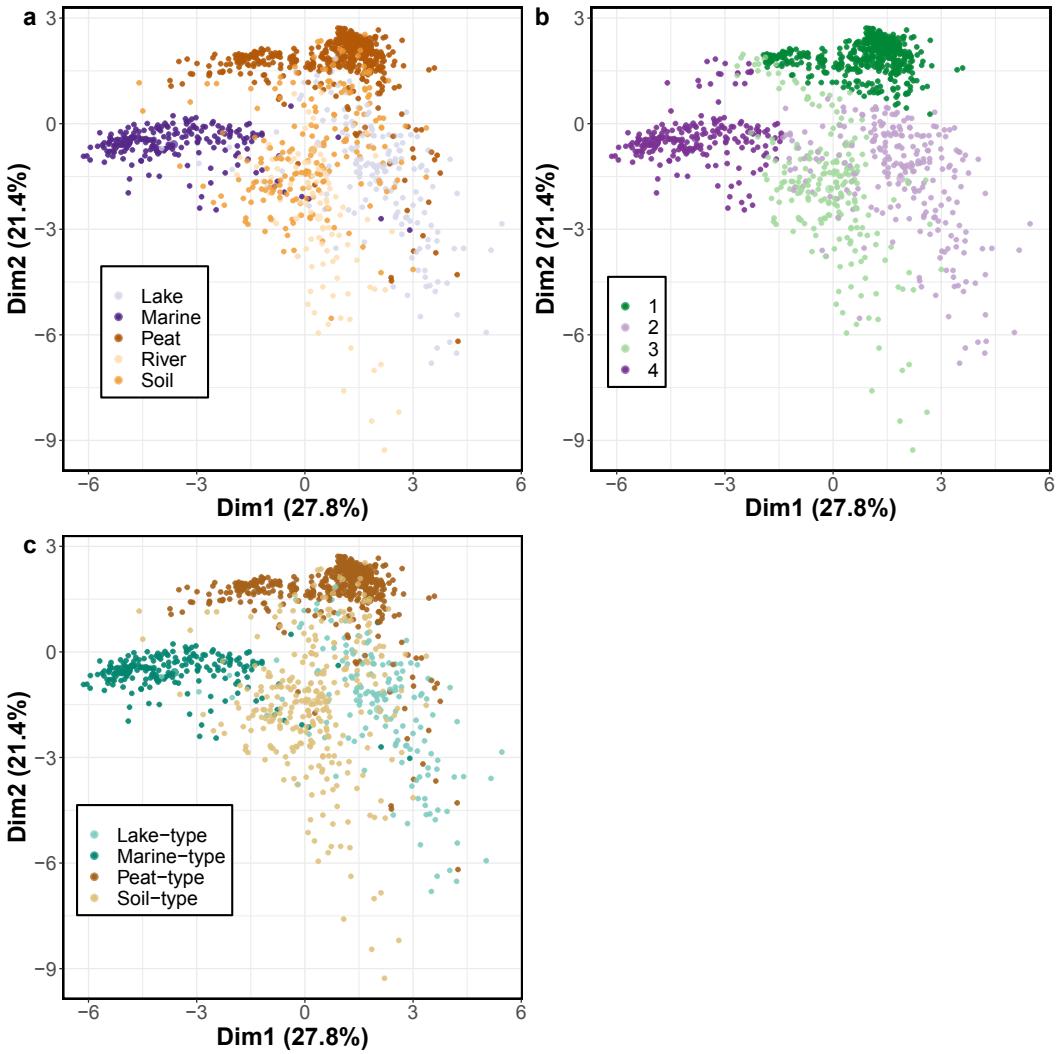

Figure 2. Samples from the dataset plotted in reduced dimensional space based on the fractional abundance of GDGTs. Plots show the same analysis with samples colored based on the depositional environment (a), their assigned group based on the fuzzy k-means analysis (b), and the curated clusters (c)

extent GDGT-1 and GDGT-2 compared with the other clusters (Fig. 3a). For crenarchaeol, cluster 3 is the next group with the highest proportion, while cluster 2 has the second highest proportion of GDGT-0. Cluster 1 consistently shows the lowest isoGDGT values. On the other hand clusters 1, 2, and 3 have a higher proportion of brGDGTs than cluster 4. Cluster 1 has the highest proportion of brGDGTs Ia and IIa, while cluster 2 has the highest proportion of IIIa and IIIa'. Although cluster 3 has lower proportion values than cluster 1, it is also dominated by the penta- and tetramethylated brGDGTs, and it shows the highest proportion of GDGT Ib. Both clusters 2 and 3 show comparable levels of IIa', the most abundant 6-methyl brGDGDT.

3.2.2 Environmental influence on GDGTs

To better understand the effect that environmental parameters (such as temperature, elevation, and pH) might have had on the classification of the samples, we per-

Table 1. Percentage of the total of each sample type assigned to each of the four clusters determined by fuzzy k-means analysis (left) as well as the four manually curated clusters (right). At the right is the total number of samples from each type. The highest percentage for each type of sample in the fuzzy k-means clusters is indicated in bold.

Type	C. 1	C. 2	C. 3	C. 4	*Peat-type*	*Lake-type*	*Soil-type*	*Marine-type*	Total
Lake	7.4	**84.6**	5.6	2.5	0.6	97.5	1.2	0.6	162
Marine	0	5.6	2.8	**91.6**	0	0	0	100	215
Peat	**86.9**	5.7	4.4	2.9	100	0	0	0	475
River	0	7.6	**92.4**	0	0	0	100	0	105
Soil	20.4	30.6	**43.9**	5.1	0	0	100	0	196

formed a principal component analysis for each of the sample types separately (Fig. 4, Fig. SI1, and Table SI2).

For peats (Fig. 4a), primarily classified in cluster 1, the main component (64.6% of variability) is positively associated with GDGT Ia, while negatively related to GDGT IIa. This component is strongly associated with MAAT ($\rho = 0.77$, Spearman's correlation and Fig. SI1a). Peats classified in cluster 1 plot throughout the first component, however, those peats classified as part of cluster 3 and 4 are associated with higher temperatures, while those classified as part of cluster 2 are associated with lower MAATs (Fig. 4a).

For lake sediments (Fig. 4b), which are mostly classified as part of cluster 2, the first component (36.6% of variability) is associated negatively with GDGT Ia, while positively related to GDGT IIIa. The second component (32.4% variability) is positively associated with GDGT-0. Both components have a strong correlation with MAAT ($\rho = -0.65$ and 0.51, respectively, and Fig. SI1b). While the majority of the distribution cloud is classified in cluster 2, samples with more negative values on PC1, which have higher MAAT values, are classified as cluster 3 and 1 (Fig. 4b and Fig. SI1b). Additionally, four samples are classified in cluster 4, which were identified as sediments from Lake Kivu, Mono Lake, and two from Lake Malawi.

Soil samples are classified into a wide range of clusters (Fig. 4c), with 44% of the samples labeled as part of cluster 3, 31% as cluster 2, 20% as cluster 1, and 10% as cluster 4. The first component of the PCA for this sample type (47% of variability) is negatively associated with brGDGT Ia, while the second component is positively related with GDGT IIa and negatively with crenarchaeol. The environmental parameters show strong correlations to both the first and second components: pH (ρ =0.47 and -0.74), MAAT (ρ =-0.45 and -0.42), and elevation (ρ =-0.21 and -0.56, respectively). Additionally, the assigned cluster strongly correlates with the sample location (Fig. SI1c). Most samples from the Godavari river catchment (98%) and Inner Mongolia (69%) were classified as cluster 3, while 70% of samples from Carminowe creek are classified as cluster 2. Finally, 47% of samples from Colombia are classified as cluster 1.

Similar to soils, the river sediments are mostly classified as cluster 3 (Fig. 4d). For these samples the first principal component explains 51.4% of the variance and is negatively associated with brGDGT Ia. For this sample type, the most determinant variable for their cluster classification was location site (Fig. SI1d), with all but one out of

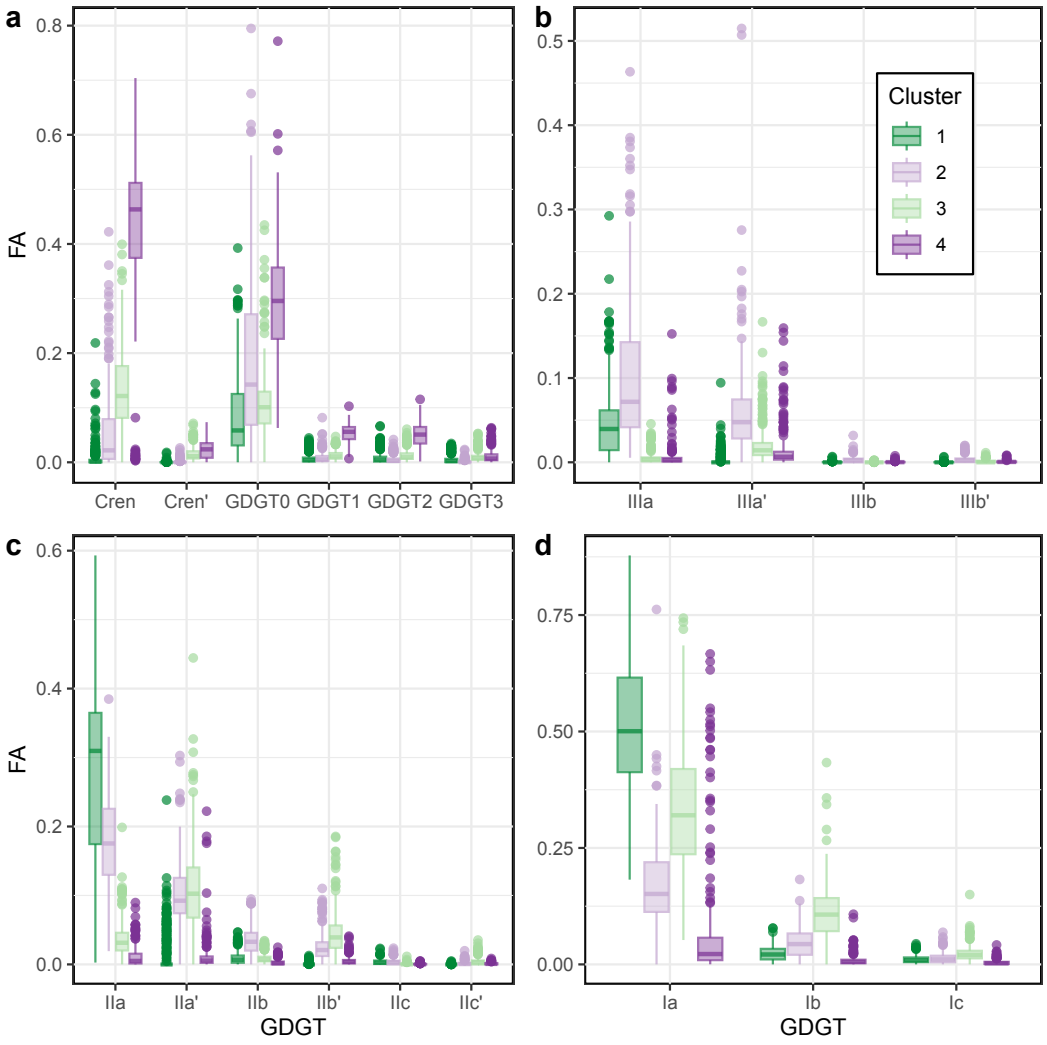

Figure 3. Box plots showing the distribution of the fractional abundance (FA) of all GDGTs in each of the fuzzy k-means clusters, following the color code of Figure 2b. GDGTs separated by isoGDGTs (a), hexamethylated brGDGTs (b), pentamethylated brGDGTs (c), and tetramethylated brGDGTs (d).

98 samples from the Godavari classified in cluster 3, while all samples from Carminowe Creek are classified as cluster 2.

An additional PCA with both river and soil samples (data not shown) showed that none of the first three components (88% of combined variance) separated these samples by type, but rather soil and river samples cluster together based on the location (Godavari or Carminowe).

Finally, for the marine sediments, primarily classified as cluster 4 (Fig. 4e), the first component of the PCA (52.6% of variability) is positively associated with GDGT Ia and negatively associated with crenarchaeol. The second component (37.2%) is positively associated with GDGT-0. The first component is associated with the sample's distance to the coast ($\rho = 0.43$ and Fig. SI1e), while the second component shows a strong correlation with the mixed layer temperature ($\rho = -0.79$). We, however, do not find a strong

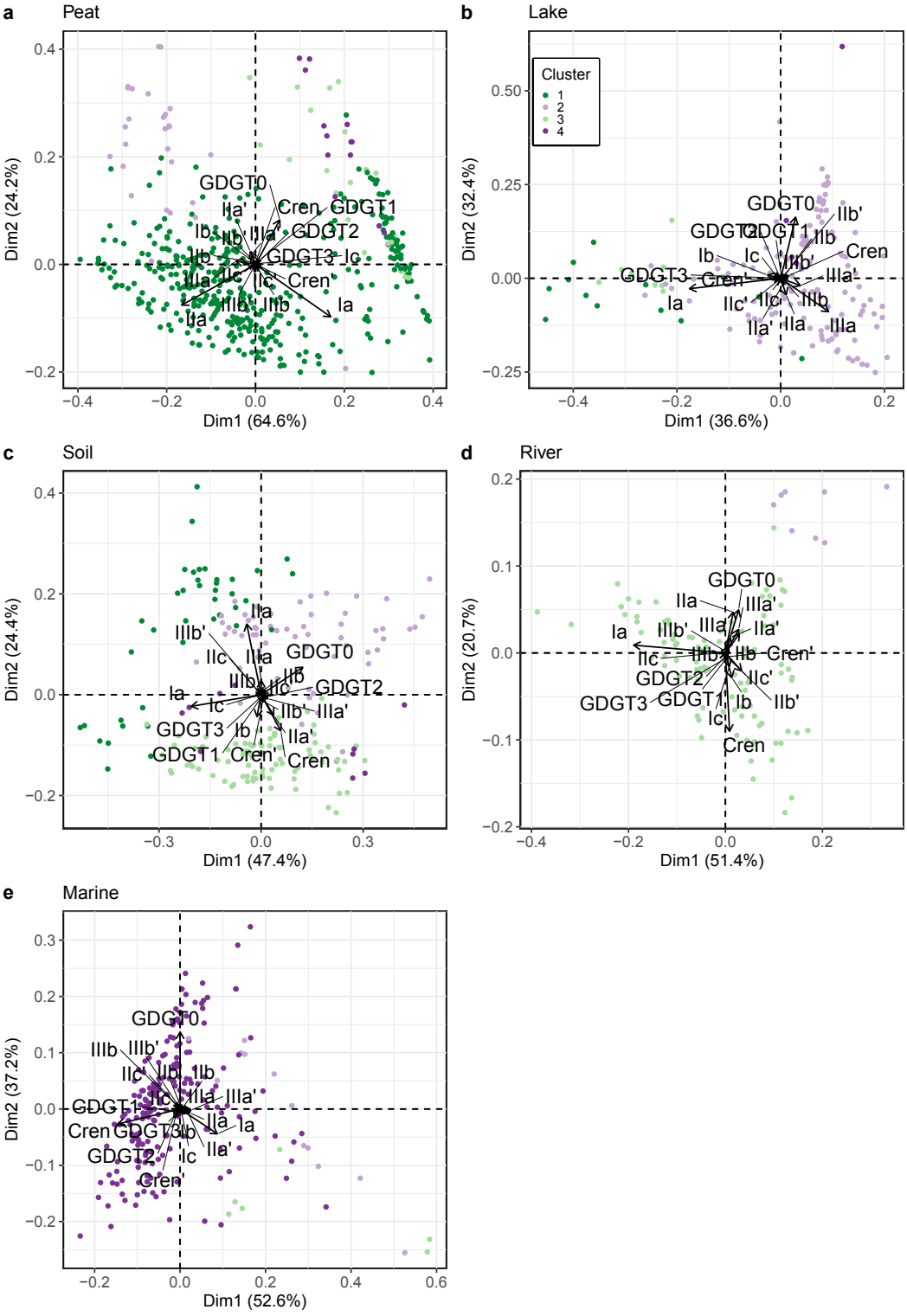

Figure 4. Principal Component Analysis of each sample type with samples colored based on their assigned K-means cluster. Each panel shows the PCA for peat (a), lake sediments (b), soils (c), river sediments (d), and marine sediments (e).

association between samples classified as cluster 2 or 3 and environmental parameters considered here.

Following the aim of this study, we considered the description of each sample, the effects of environmental parameters in it, as well as the K-means classification. With this information we generated new classification labels, which we prioritized in cases where the fuzzy K-means classification and these labels disagreed. The informed labels are named according to the dominant depositional environment. Group 1 was renamed as *Peat-type*, Group 2 as *Lake-type*, Group 3 as *Soil-type*, and finally Group 4 as *Marine-type* (Fig. 2c). While *Soil-type* combines samples from rivers and soils, we name it as such since we have a much larger representation of soils in the dataset, but also mechanistically it is more likely that GDGTs from the soils are influencing nearby rivers rather than the other way around.

3.3 Supervised Machine Learning

The new informed labels generated through the unsupervised machine learning phase were used for the supervised classification. We tested the performance of all four classification algorithms against each other and compared them with the null model using both the F1 and ROC-AUC parameters. Our results suggest that overall, all methods performed significantly better than the noninformative control and relatively similar to each other. For the F1 scores, Random Forest performed the best (0.95), followed by XGBoost (0.94), K-Nearest Neighbour (0.91), and Naive Bayes (0.87). In contrast, the null model had a score of 0.58. Similarly, for the ROC-AUC parameter Random Forest, XGBoost, and K-Nearest Neighbour had the same performance (0.99), followed by Naive Bayes (0.96), and the null model had a value of only 0.5. Based on these results we chose the Random Forest algorithm for our study. The performance of this algorithm in the test set is similar to the one obtained for the training set (0.94 and 0.99 for F1 and ROC-AUC respectively, Fig. 5).

Finally, we diagnose the importance that each predictor variable has on the trained classification algorithm. This analysis shows that brGDGT IIa' and crenarchaeol have the highest importance scores (> 90), followed by IIb', IIIa', IIIb, Ia, and crenarchaeol' (> 30). All other variables had importance values < 30 (Fig. 5b).

The finalized model, named **B**ranched and **I**soprenoid **G**DGT **Ma**chine learning **C**lassification algorithm (BIGMaC), is available on Github as an R object (Martínez-Sosa et al., 2023).

4 Discussion

4.1 Unsupervised Machine Learning

While our fuzzy k-means clusters show strong patterns that reflect relationships with depositional environments (Fig. 2a), some samples whose context was unequivocally documented cluster in with samples unrelated to their depositional environment (*i.e.* soils plotting as peats).

For both peats and lakes, temperature has a strong influence on the GDGT distribution, causing samples at either ends to be clustered with other sample types (Fig. 4 a, b, and Fig. SI1 a and b). At high temperatures, samples from these depositional environments tend to have a higher proportion of brGDGT Ia, and lower proportion of IIa and IIIa (Weijers et al., 2007), which causes the lake sediments to be classified as cluster 1 and 3 (high Ia), or peats as cluster 3 and 4 (low IIa and IIIa). At lower temperatures, the opposite effect causes some of the peats to be classified as cluster 2. In addition, particularly deep (*i.e.* Lake Malawi) or hypersaline alkaline lakes also showed an increased proportion of isoGDGTs compared with other lakes, causing them to be clas-

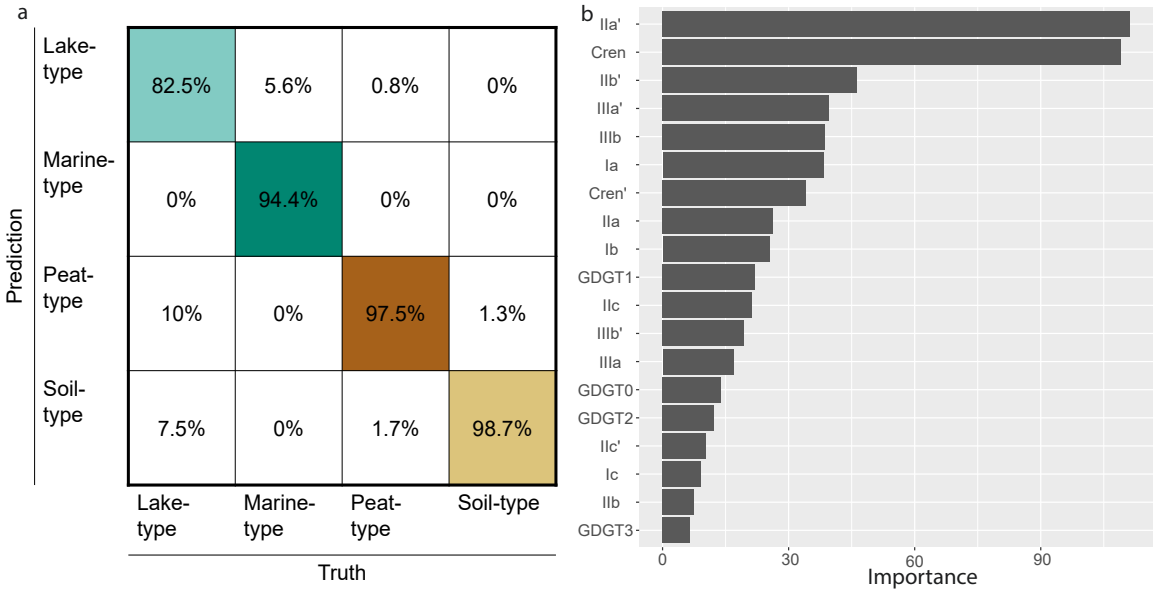

Figure 5. Confusion matrix and importance value for all the GDGTs considered in the classification model. The confussion matrix (a) shows the performance of the BIGMaC Random Forest algorithm in the test dataset. Columns show the true label of the samples and rows show the predicted label. Diagonal cells are color-coded based on Fig. 2. The bar plot (b) shows the importance value of each GDGT, calculated based on Gini impurity, used in the BIGMaC classification algorithm.

sified with marine samples. These distinct distributions could be due to specific niches in the water column associated with water chemistry, stratification, and/or nutrient content, as previous work has suggested (Sinninghe Damsté et al., 2022; Baxter et al., 2021; Kumar et al., 2019).

Soil samples cluster into several different groups, which may reflect the fact that soils are highly diverse environments with diffuse boundaries and are often in contact with other depositional settings. Furthermore, studies have shown that chemical properties of soils (i.e. pH, metal concentrations) have great spatial heterogeneity even at small scales (Yavitt et al., 2009). Indeed, the main factor that separated soils in this study is their location (Fig. 4c and Fig. SI1c), which overrides any other environmental signal. However, this may simply be a feature of the limited existence of soil datasets with both branched and isoprenoid GDGTs (represented by only four locations). It is clear from our results that soils require a more in-depth analysis, with the use of more extensive datasets.

While there is some debate regarding the relative influence that soil input and in situ production have on the GDGT pool in river organic matter (Kirkels et al., 2020; Zell et al., 2013; De Jonge, Stadnitskaia, et al., 2014), our analysis shows that the river sediments more closely resemble soils rather than peats or lake sediments, and similarly to soils, the sample location shows the strongest correlation with how these samples are classified by fuzzy k-means (Fig. 4d and Fig. SI1d). We do find a difference in the 5-methyl/6-methyl proportion in these groups, where soils have a relatively similar proportion of these isomers, while river sediments contain relatively more 6-methyl brGDGTs (Fig. SI2 b and c). Although this could be interpreted as soil-derived GDGTs dominating river inputs, with some autochthonous production of 6-methyl brGDGTs in rivers, our river data come from only two locations and primarily from only one system (the Godavari river),

so this interpretation could be particular to this watershed. Notably, within the Godavari River, the membership value for the samples, which measures the degree of belonging to each cluster, varies with their location and collection season (Fig. 6). Membership to the soil-dominated Group 3 is higher in the lower Godavari basin, as well as from the wet (post-monsoon) season (Fig. 6 c and d). In contrast, membership to the lake-dominated Group 2 is overall higher in the dry season, and in the upper basin year-round (Fig. 6 a and b). These results are in line with those presented in the original study by Kirkels, Zwart, et al. (2022), where it was noted that GDGTs from soils have a stronger influence on the river's GDGT content during the wet season and within the lower basin, which experiences higher precipitation. In contrast, in-situ production of brGDGTs, characterized by a high proportion of 6-methyl isomers, has a stronger influence on the GDGT content of samples from the dry season as well as those from the upper basin.

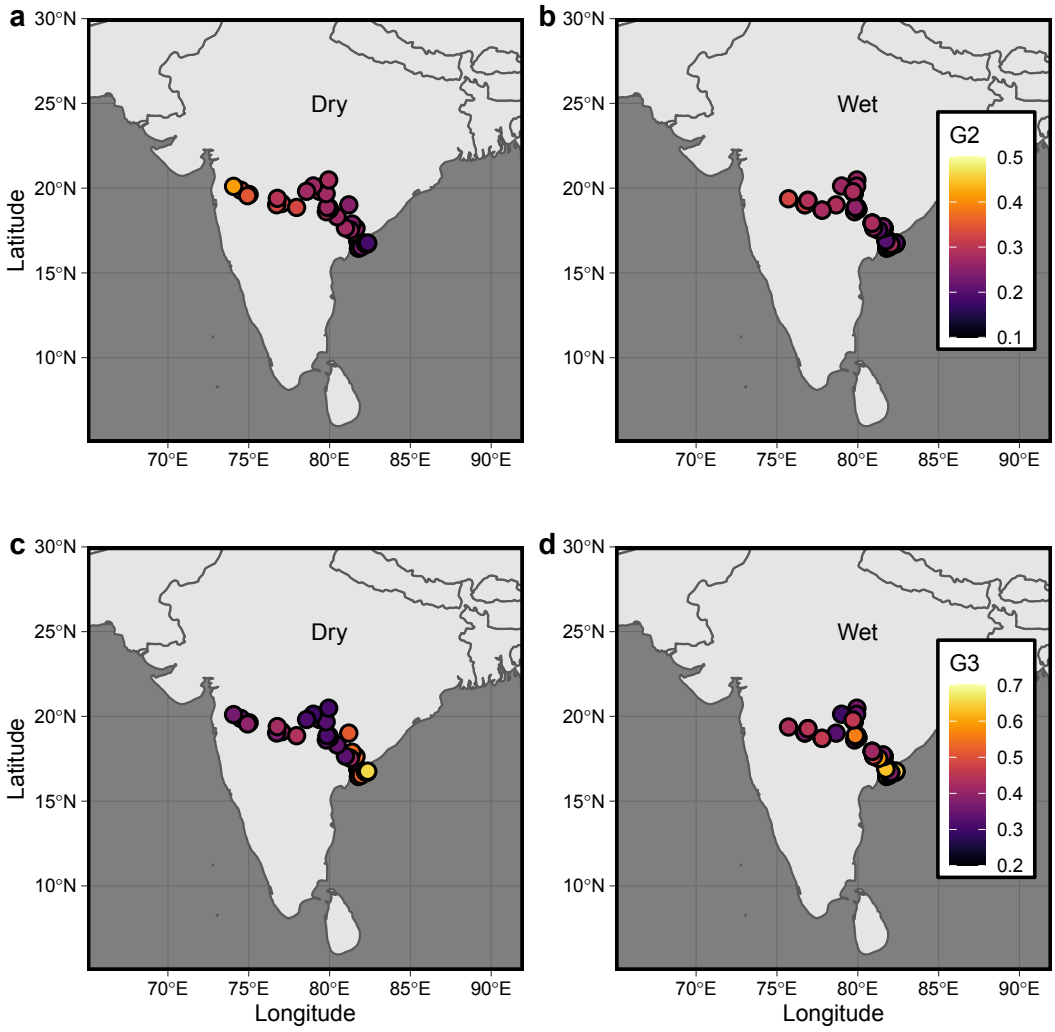

Figure 6. Maps for the Godavari River sample locations for the dry (left column) and wet (right column) seasons. Maps show the sample memberships, calculated through fuzzy k-means analysis, to the lake-dominated Group 2 (a and b), and to the soil-dominated Group 3 (c and d).

Notably, for marine sediments the first dimension of their PCA is dominated by a positive relation with brGDGT Ia and a negative one with crenarchaeol and it is as-

sociated with distance from the coast, although this is true only for samples more than 10 km away from the nearest coast, suggesting that more coastal samples may be affected by other GDGT sources (Fig. 4e and Fig. SI1e). The second dimension, which is positively related to GDGT-0, more closely follows the mixed layer temperature, as shown by the Spearman correlation between this variable and the principal component ($\rho = -0.79$). Although GDGT-0 is traditionally omitted from the TEX_{86} calculation because it is a generic isoGDGT produced by many types of archaea (including methanotrophs and methanogens) (Kim et al., 2010; Schouten et al., 2002), our analysis shows that it is strongly influenced by temperature. Furthermore, the PCA shows no relation between GDGT-0 and brGDGTs (Fig. 4e), which suggests that GDGT-0 is not influenced by terrestrial sources. Our results suggest that temperature strongly influences the abundance of GDGT-0 and, unlike previously thought (Guo, Yuan, et al., 2022; Kim et al., 2010), other environmental parameters may not be as important in open marine settings. This supports the observation of Cramwinckel et al. (2018) that, at higher temperatures, the ratio of crenarchaeol to GDGT-0 might be more sensitive to temperature changes than TEX_{86}.

Since our intention with the supervised machine learning was to test whether GDGT distributions can be used to identify the depositional environment, we generated four new groups which broadly follow the fuzzy k-means clusters, but considered the actual depositional environmental that the samples came from (Fig. 2 a to c). For example, although some of the peat samples fell into cluster 2 (lakes) or cluster 4 (marine), since they were derived from peatlands, we re-assigned them as such. As noted above, some of these false assignments appear to be associated with the effect of temperature on GDGT distributions in peats. Similarly, although soils fell into several different clusters, this seems to be because their GDGT distributions were influenced by the site location, so we assigned them all to a single soil group. By manually reassigning samples to match their true environment, we feed the classification algorithm a more realistic dataset that includes some of the uncertainties associated with the relationship between GDGT distributions and their depositional settings. While we cannot rule out that the effects observed here are due to artifacts of the clustering technique used, as we only tested the performance of fuzzy k-means based on Euclidian distances, which can be affected by high dimensionalities, the environmentally relevant results obtained give us confidence in the approach used.

4.2 Curated clusters

The manual curation of these clusters does not alter the general composition of the groups compared with the statistically-derived ones (Table 1). *Peat-type* and *Marine-type* are very similar in composition and size to Group 1 and 4 respectively. While Group 1, with 465 samples, had 87% of the peats and 20% of the soils; *Peat-type*, with 476 samples, has all of the peats and only one lake sediment. Similarly, Group 4, with 225 samples, had 92% of the marine sediments, while *Marine-type* includes all of them and has a total of 216 samples. The reduction in size from Group 4 to *Marine-type* is mostly due to the reassignment of lake sediments, peats and soils. The largest change observed is between Group 2 and *Lake-type* (86 sample difference), and Group 3 and *Soil-type* (84 sample difference). Most of this comes from the reassignment of 60 soils from Group 2 to *Soil-type*.

The sample reassignment also does not alter the general GDGT distribution patterns of the groups when we compare them before and after the reassignment (Fig. 3). It does, however, preserve the cluster selection for samples where the depositional environments may not be as clear (i.e., humic-rich lakes which are classified as *Peat-type*).

4.3 Supervised Classification

In general, all of the machine learning algorithms exhibited good performance in the training phase, with F1 and ROC-AUC scores above 0.85 and 0.95 respectively. Nevertheless, we chose the Random Forest algorithm since it was the best performing one across all parameters, in addition to being widely used in the field of Geosciences (Peaple et al., 2021; El Bouchefry & de Souza, 2020). This algorithm also performed well in the testing phase (0.94 and 0.99, for F1 and ROC-AUC respectively, and Fig. 5). This result suggests that the algorithm is not overfitting the data, as the algorithm's performance would have significantly decreased if it had been trained to only classify samples it had previously been exposed to.

When we apply the BIGMaC algorithm to the complete dataset, we can investigate the importance of each GDGT in the model. This analysis shows that the two compounds that contribute the most to the classification are IIa' and crenarchaeol, although all GDGTs contribute to some extent to the classification, and thus relying only on these two compounds may not be as effective as the complete model. While these compounds have not been specifically associated with particular environments, they have very specific distributions in the identified clusters (Fig. 3), and also could be associated with characteristics of the depositional environments. BrGDGT IIa' is the 6-methyl GDGT with the highest abundance in lakes and soils, however, these isomers are less abundant in peats although this could also be due to most of the peats in our dataset being acidic (Naafs et al., 2017) and the inclusion of less acidic samples could change the variables used. In addition, generally all brGDGTs have a lower abundance in marine environments (Hopmans et al., 2004). In contrast, crenarchaeol is generally the most abundant isoGDGT in marine sediments (Liu et al., 2011), and it has been shown that it is also present in relatively high proportion in soils, but not in lakes (Naeher et al., 2014), and it has generally a low proportion in peats (Naafs et al., 2017). These specific distributions form four distinct patterns, which likely explain their selection as the most informative variables: low IIa'—low crenarchaeol (*Peat-type*), high IIa'—low crenarchaeol (*Lake-type*), low IIa'—high crenarchaeol (*Marine-type*), and high IIa'—high crenarchaeol (*Soil-type*).

4.4 Applications

We demonstrate that our model can be successfully used to analyze changes in depositional environments through time by testing the BIGMaC algorithm on GDGTs measured in two different sites: the Eocene-aged post-eruption peat and lacustrine sediments recovered from the Giraffe kimberlite pipe in the subarctic; and the Cobham lignite bed, dated to the beginning of the PETM.

4.4.1 Giraffe Kimberlite Pipe

When we apply the BIGMaC algorithm to the Giraffe kimberlite pipe core we see that the samples are generally correctly classified with the general stratigraphy previously described for the core (Wolfe et al., 2017; Hamblin et al., 2003) (Fig. 7 and GDGT distributions shown in Fig. SI4). All samples from the top peatland section are classified as *Peat-type*, and all samples from the lacustrine section below 85 m are classified as *Lake-type*. However, we also identified a section, between 76.5 and 85 m, within the lacustrine facies that is classified as *Peat-type*. Furthermore, the samples immediately above the excursion oscillate between *Lake-type* and *Soil-type* for at least one meter (Fig. 7), before returning to being classified as *Lake-type* for the last 4.8 m of the lake section. The results from our classification algorithm are in contrast to other approaches previously used to determine the origin of GDGTs in sediments. For example, applying the BIT index to this core shows that for the majority of the core, the mean value is 0.999± 0.001, suggesting a terrestrial setting. The BIT index record only deviates from these values in the one-meter section where the BIGMaC classifies samples as either *Lake-type*

or *Soil-type*, although even in this section the BIT index values are only reduced to 0.97. Since the BIT index is unable to distinguish between soil, peat and lake deposits, this showcases the advantage of using the classification algorithm where all GDGTs are considered over a singular index.

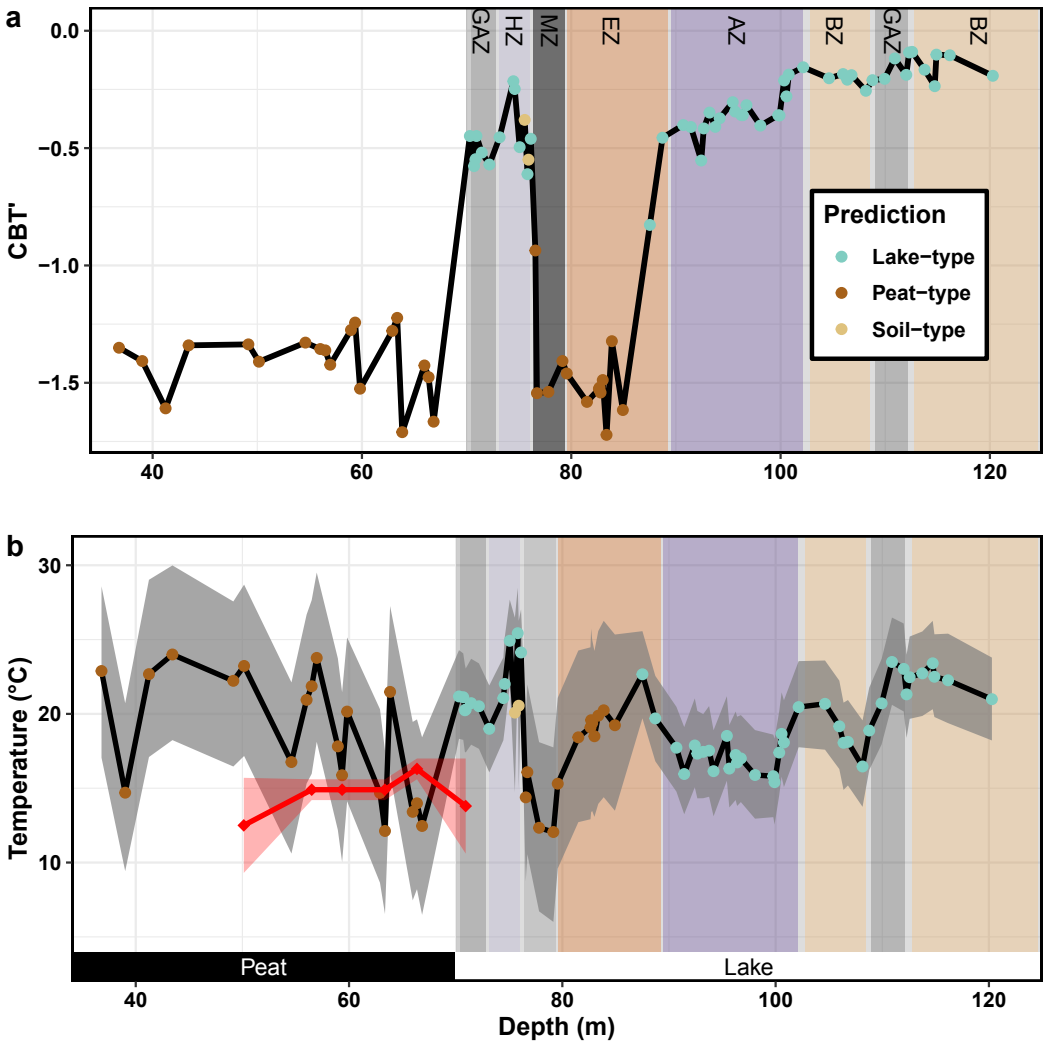

Figure 7. CBT' values (a) and inferred temperature (b) calculated from Giraffe core brGDGTs plotted against vertical-equivalent depth. The temperature reconstruction was generated by applying the Dearing Crampton-Flood et al. (2020) Bayesian calibration for *Peat* and *Soil-type* samples, and Martínez-Sosa et al. (2021) calibration for *Lake-type* samples. Palynological estimates of mean annual temperature with their associated error from Wolfe et al. (2017) are shown in red diamonds in (b). Samples are color-coded based on the predicted groups. White and gray shading indicates peat and lacustrine sediments in the core, respectively. Zones in the lake section as described by Siver and Lott (2023) are shown as colored areas: *Botryococcus* Zone (BZ) in light orange, Going Acidic Zone (GAZ) in light gray, *Aulacoseira* Zone (AZ) in purple, Eunotiod Zone (EZ) in orange, Mixing Zone (MZ) in gray, and Heterotrophic Zone (HZ) in light purple. The peat and lake sections of the core are indicated at the bottom of the plot with black and white rectangles, respectively

By using the CBT' index, which has been shown to be strongly associated with pH in peats (Naafs et al., 2017) and mildly correlated to pH in lakes (Martínez-Sosa et al., 2021), we estimate that in general the peat section has much lower CBT' values (associated with lower pH) than those in the lacustrine section. While this trend is maintained for most of the core, a marked decrease in CBT' values occurs in the section within the lacustrine facies that is classified as *Peat-type*.

Our results show a close relationship with the independently developed microfossil-based ecological reconstruction done by Siver and Lott (2023) for this site. While Hamblin et al. (2003) had previously speculated this site to be a shallow lacustrine setting with intermittent wet and dry periods, the microfossil ensemble suggest that the lake was initially a shallow ($\sim$ 1m) slightly acidic lake, this is represented by the *Botryococcus* Zone (BZ) between 102.8 and 124.6 m (Fig. 7). All samples in this section are classified as *Lake-type* with generally stable CBT' values. While Siver and Lott (2023) reported a shift in the microfossil ensemble within this region between 112.1 and 109 m, which was interpreted as an acidifying section, we do not find evidence of this in the CBT' values; however, this subsection does roughly align with an estimated reduction in MBT'_{5Me} values. Following this initial section, the microfossils suggest that the lake deepened to between 3 — 5 m in the *Aulacoseira* Zone (AZ) (Siver & Lott, 2023), between 102.1 and 89.5m. BIGMaC still classifies all samples from this section as *Lake-type*, however there is a declining trend in CBT' values throughout this region, shifting from vales of -0.2 to -0.55. The following section located between 89.2 to 79.6 m, identified by the microfossil ensemble as Eunotiod Zone (EZ), is thought to represent a period characterized by enhanced acidity and increase in dissolved humic material. This section corresponds to the large CBT' excursion in our record, with values from -0.5 to below -1.5. Microfossils in this section are associated with low pH environments, such as lakes and bogs. This is in agreement with our BIGMaC results, where all but one sample in this section are classified as *Peat-type*, with the exception being a sample right at the bottom that BIGMaC still classified as *Lake-type*. We note that the CBT' values are lower than those of previously reported lakes with a pH of 4.3 (Martínez-Sosa et al., 2021). In contrast, these values correlate to a pH between 4 and 5 in peats, which is more in line with what is associated with the described organisms in this section (Siver & Lott, 2023). Between 79.4 and 76.4 m the microfossil ensemble suggests the presence of a transitional zone that is not associated with any of the other sections. Our results classify all samples within this section as *Peat-type*, although the CBT' values have a steep increase, going from -1.5 to -0.5, which could suggest a transitory state towards a higher pH environment. The following section, between 76 and 73.1 m, identified as the Heterotrophic Zone (HZ) by the microfossil ensemble, is interpreted as a period with higher pH and in line with conditions of an eutrophic body of water. Our analysis shows a transitory period at the begining of this section where BIGMaC interprets two samples as *Soil-type*, while the rest of the samples are classified as *Lake-type*. The CBT' values are on par with those of the previous sections that were classified as *Lake-type*. Finally, the top-most part of the lake section is interpreted as a body of water with increased acidity and levels of dissolved humic matter. While the CBT' values do show a small decrease, the BIGMaC algorithm still classifies all samples in this section as *Lake-type*.

Overall, there is a good agreement in the interpretation of both independent proxy-based reconstructions for the lake section of the core. Differences between the interpretations could be due to constraints of the specific proxies. For example, while the microfossil ensemble suggests periods of increased acidification throughout the BZ region, this is not reflected in the CBT' values. In contrast, while the microfossil ensemble suggests a sudden transition from the AZ to the EZ sections, our results show a decreasing trend in CBT' values throughout AZ leading to the much steeper acidification in EZ. It is possible that the decreases in pH in both cases have different origins, which are only captured by one of the proxies. This underscores the advantage of combining independent proxies for paleoenvironmental reconstructions.

To generate a temperature reconstruction for the Giraffe core, we applied the previously published BayMBT calibration for lakes or soil/peat, depending on the results of BIGMaC for each sample. These particular calibrations were chosen as they are consistent with each other and allow us to generate a continuous confidence interval for the reconstruction. However, we emphasize that the choice of temperature calibration is independent from BIGMaC and any GDGT-based calibration can be used. For example, we also applied the peat-specific calibration (Naafs, 2017) to the sections of the core classified as *Peat-type* but it generated only marginally different results than BayMBT, so we chose to use the latter.

Our reconstruction suggests a relatively stable climate with no clear trend (Fig. 7a). The mean temperature of our reconstruction (19±3.2°C) agrees with independent studies. A pollen reconstruction at this site (red diamonds in Fig. 7a), suggests a MAAT of 14.5 ± 1.3°C, with a warmest month mean temperature of 24.5 ± 0.8°C (Wolfe et al., 2017). In addition, Jahren and Sternberg (2003) estimated a mean annual temperature of 13.2±2°C for the middle Eocene Arctic based on oxygen isotopes measured in calcite preserved in fossil *Metasequoia*. While our estimate is at the upper end of both estimates, they fall within the confidence interval of our reconstruction (Fig. 7a). Moreover, both the peat/soil and lake calibrations predict mean annual temperatures above freezing (MAF) rather than strictly MAAT, so if there were freezing temperatures during the winter, the GDGT estimates are expected to be higher. Conversely, if we had used only the lakes or soil/peat calibration for the entire core, there would be large temperature swings of more than 6°C associated with changes in core lithology. In particular, the excursion to *Peat-type* samples within the lacustrine section would be estimated to be 5.7°C higher without the BIGMaC-based correction.

4.4.2 Cobham Lignite Bed

While the application of the BIGMaC algorithm in the Giraffe pipe showcases its strengths, our analysis of the Cobham lignite illustrates that there are some limitations of the approach. Inglis et al. (2019) previously showed that increased precipitation during the PETM in this area caused changes in the hydrology of the site, and that this potentially caused the brGDGTs to become unreliable as temperature proxies. Namely, while several lines of evidence suggest an increase in temperature during the PETM, the temperature reconstructions based on brGDGTs suggest cooling. We applied BIGMaC to this site to investigate whether changes in the depositional settings could explain the discrepancy (Fig. 8). Almost all samples preceding 54.15 cm, identified as the start of the PETM, are predicted to be *Peat-type*, with the exception of one sample from the upper laminated lignite unit that is classified as *Soil-type*. In contrast, there is a wider variation in the sample classification during the PETM, where samples are classified as *Peat-type* (10), *Soil-type* (3) and *Lake-type* (1). Besides one sample classified as *Peat-type* from the PETM upper laminated lignite, all other PETM samples are located in the blocky lignite unit. The variations in predicted depositional environments do not coincide with changes in either MBT'_{5Me} or CBT' values, nor are they organized in any evident pattern within the unit. Moreover, the PETM samples are primarily classified as *Peat-type* and *Soil-type*, suggesting that adjusting for the predicted depositional environment would not significantly increase the temperature reconstructed by Inglis et al. (2019), as would be the case if the samples were classified as *Lake-type*. Vegetation and charcoal records suggest that the Cobham site became waterlogged and may have even developed areas of open water during the PETM (Inglis et al., 2019). Given this perspective, the oscillating results from BIGMaC likely point to an unstable, dynamically changing depositional environment with mixed sources of brGDGTs. Since BIGMaC is a categorical classification algorithm, it cannot detect mixed signatures. This underlines the need to incorporate mixing models in studies where input from different sources is expected, and suggests that BIGMaC would benefit from incorporating this capability in future updates.

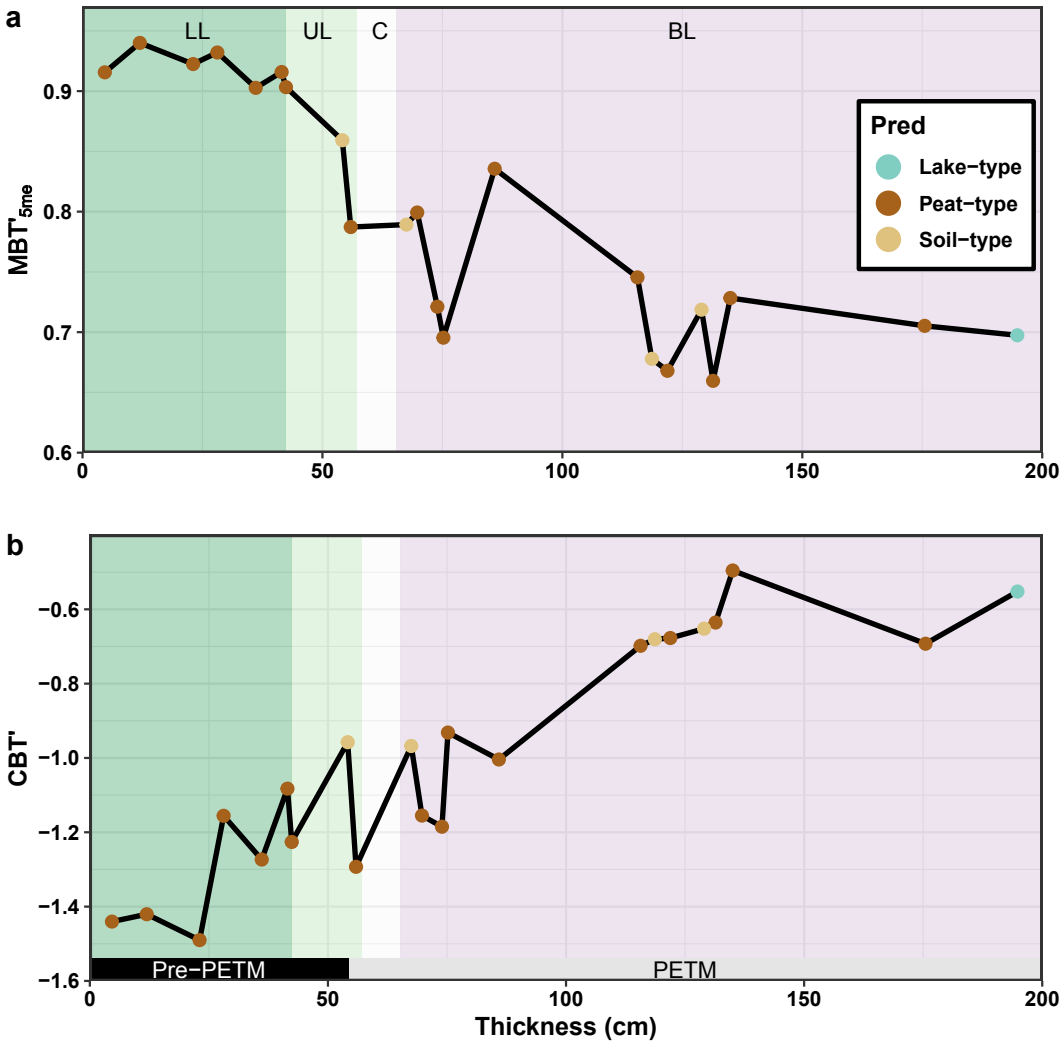

Figure 8. Calculated MBT'_{5Me} (a) and CBT' (b) values of the Cobham lignite bed across the site thickness (cm). Samples are color coded based on the BIGMaC predicted groups. Different units are colored and labeled on the top as: lower laminated lignite (LL, dark green), upper laminated lignite (UL, light green), clay (C, white), and blocky lignite (BL, purple).

Overall, this application suggests that other environmental parameters, not considered in this work, could still affect the distribution of GDGTs, and we include the Cobham lignite example as a reminder that applying the BIGMaC algorithm is not always a panacea and is best done in concert with other independent proxies to accurately interpret the results.

5 Conclusions

Our analyses of GDGTs in 1153 globally distributed samples from soils, lakes, rivers, and marine sediments show that the depositional environment from which samples were obtained has a significant and measurable impact on the combined distribution of isoprenoid and branched GDGTs, which allows us to cluster the samples from our dataset into environmentally relevant groups. Furthermore, we find that the distribution of GDGTs

in each cluster is uniquely impacted by the environment. There is a strong association between temperature and the *Lake-type* and *Peat-type* groups. *Marine-type* samples are also clearly influenced by temperature and their distance from the coast, at least for samples more than 10km away from the shore. The cause of this distinction is an observation that deserves further study. Although they are represented by a limited sample set, soils show variability that is strongly influenced by their location. Additionally, while our analysis groups soil and river samples together into the *Soil-type* cluster, river systems seem to have more 6-methyl brGDGTs and their GDGT distributions reflect local changes within the catchment.

We used the dataset presented here to train the Random Forest classification algorithm BIGMaC, which is capable of identifying the environment in which a sample was formed based on the distribution of GDGTs. Our results show that GDGTs IIa' and crenarchaeol are the most influential compounds in the classification algorithm, due to their combined unique changes between the four depositional groups. As a demonstration, we apply the BIGMaC model to an independent record from the Giraffe kimberlite, which was stratigraphically shown to record a transition from a lacustrine environment to peatland. Our BIGMaC algorithm is not only able to recreate the transition, but further suggests an excursion to peatland conditions within the upper lacustrine section of the core, which is consistent with independent evidence for more acidic conditions. This result is encouraging for the application of our classification algorithm, as it comes from a dataset not included in the training or testing sets, thus providing an independent testing case. Using the BIGMaC results as a guide, we apply brGDGT-derived calibrations specific to lakes or soils and peats as needed downcore and obtain a relatively stable temperature estimate for this area that is in general agreement with the pollen record.

While our Giraffe pipe results showcase the usefulness of our approach when applied to clear changes in depositional environments; the application of BIGMaC in the Cobham site shows that this approach may not be suitable in cases where the depositional environment is changing rapidly and thereby results in mixed sources of GDGTs. It is possible that the future integration of a mixing model in the BIGMaC workflow could improve its performance in this type of scenario.

Ultimately, we show that the combined set of branched and isoprenoid GDGTs is an effective tool for identifying depositional environments that can be used in combination with more established proxies to gain a better understanding of past environments.

Open Research

The GDGT fractional abundance data used for training the BIGMaC algorithm in the study are directly available at Pangaea via Naafs (2017), Guo et al. (2020), Dearing Crampton-Flood et al. (2019), Guo et al. (2021), and Inglis et al. (2019); as well as on Zenodo via Martínez-Sosa et al. (2023b), Martínez-Sosa et al. (2023a) and Pérez-Angel et al. (2020). V1.0 of the BIGMaC algorithm used for the classification of samples based on GDGT fractional abundances is preserved at Martínez-Sosa et al. (2023), available via MIT license and developed openly in the `tidymodels` environment in R.

Acknowledgments

We would like to thank Patrick Murphy for his assistance with the lipid analysis, Dr. Jeffrey Donnelly and the Woods Hole Oceanographic Institution Seafloor Samples Laboratory for access to marine sediment samples, and Dr. Cody Routson for contributing Alaskan lake samples. The hyperparameter tuning of the models was performed using the Ocelote cluster from the University of Arizona. This research was funded by the American Chemical Society Petroleum Research Fund, grant 60772-ND2, and by CONACYT through the student scholarship 440897. Ioana Stefanescu and Bryan Shuman ac-

knowledge support from the Microbial Ecology Collaborative Project through the National Science Foundation grant EPS-1655726. Francien Peterse acknowledges funding from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) through Veni grant no. 863.13.016 and Vidi grant no. 192.074. Lina Pérez-Ángel and Julio Sepúlveda acknowledge support from NSF Sedimentary Geology and Paleobiology grant 1929199. We also thank Serhiy Buryak for assisting with the sampling of the Giraffe pipe sediments.

References

Baxter, A., van Bree, L., Peterse, F., Hopmans, E., Villanueva, L., Verschuren, D., & Sinninghe Damsté, J. S. (2021). Seasonal and multi-annual variation in the abundance of isoprenoid GDGT membrane lipids and their producers in the water column of a meromictic equatorial crater lake (Lake Chala, East Africa). *Quaternary Science Reviews*, *273*, 107263.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Chen, Y., Zheng, F., Yang, H., Yang, W., Wu, R., Liu, X., ... others (2022). The production of diverse brGDGTs by an Acidobacterium providing a physiological basis for paleoclimate proxies. *Geochimica et Cosmochimica Acta*, *337*, 155–165.

Collinson, M. E., Steart, D. C., Harrington, G. J., Hooker, J. J., Scott, A. C., Allen, L. O., ... Gibbons, S. J. (2009). Palynological evidence of vegetation dynamics in response to palaeoenvironmental change across the onset of the Paleocene-Eocene Thermal Maximum at Cobham, Southern England. *Grana*, *48*(1), 38–66.

Cramwinckel, M. J., Huber, M., Kocken, I. J., Agnini, C., Bijl, P. K., Bohaty, S. M., ... others (2018). Synchronous tropical and polar temperature evolution in the Eocene. *Nature*, *559*(7714), 382–386.

Dang, X., Ding, W., Yang, H., Pancost, R. D., Naafs, B. D. A., Xue, J., ... Xie, S. (2018, May). Different temperature dependence of the bacterial brGDGT isomers in 35 Chinese lake sediments compared to that in soils. *Org. Geochem.*, *119*, 72–79.

Dearing Crampton-Flood, E., Tierney, J. E., Peterse, F., Kirkels, F. M. S. A., & Sinninghe Damsté, J. S. (2019). *Global soil and peat branched GDGT compilation dataset* [Dataset]. PANGAEA. doi: 10.1594/PANGAEA.907818

Dearing Crampton-Flood, E., Tierney, J. E., Peterse, F., Kirkels, F. M., & Sinninghe Damsté, J. S. (2020). BayMBT: A Bayesian calibration model for branched glycerol dialkyl glycerol tetraethers in soils and peats. *Geochimica et Cosmochimica Acta*, *268*, 142–159.

De Jonge, C., Hopmans, E. C., Zell, C. I., Kim, J.-H., Schouten, S., & Damsté, J. S. S. (2014). Occurrence and abundance of 6-methyl branched glycerol dialkyl glycerol tetraethers in soils: Implications for palaeoclimate reconstruction. *Geochimica et Cosmochimica Acta*, *141*, 97–112.

De Jonge, C., Radujković, D., Sigurdsson, B. D., Weedon, J. T., Janssens, I., & Peterse, F. (2019). Lipid biomarker temperature proxy responds to abrupt shift in the bacterial community composition in geothermally heated soils. *Organic Geochemistry*, *137*, 103897.

De Jonge, C., Stadnitskaia, A., Hopmans, E. C., Cherkashov, G., Fedotov, A., & Sinninghe Damsté, J. S. (2014). In situ produced branched glycerol dialkyl glycerol tetraethers in suspended particulate matter from the Yenisei River, Eastern Siberia. *Geochim. Cosmochim. Acta*, *125*, 476–491.

de Rosa, M., de Rosa, S., Gambacorta, A., Minale, L., & Bu'lock, J. D. (1977). Chemical structure of the ether lipids of thermophilic acidophilic bacteria of the Caldariella group. *Phytochemistry*, *16*(12), 1961–1965.

De Rosa, M., Gambacorta, A., Nicolaus, B., Chappe, B., & Albrecht, P. (1983). Isoprenoid ethers; backbone of complex lipids of the archaebacterium Sulfolobus solfataricus. *Biochimica et Biophysica Acta (BBA)-Lipids and Lipid Metabolism*, *753*(2), 249–256.

Dunkley Jones, T., Eley, Y. L., Thomson, W., Greene, S. E., Mandel, I., Edgar, K., & Bendle, J. A. (2020). OPTiMAL: a new machine learning approach for GDGT-based palaeothermometry. *Climate of the Past*, *16*(6), 2599–2617. Retrieved from `https://cp.copernicus.org/articles/16/2599/2020/` doi:

10.5194/cp-16-2599-2020

El Bouchefry, K., & de Souza, R. S. (2020). Learning in big data: Introduction to machine learning. In *Knowledge discovery in big data from astronomy and earth observation* (pp. 225–249). Elsevier.

Engle, M. A., & Brunner, B. (2019). Considerations in the application of machine learning to aqueous geochemistry: Origin of produced waters in the northern US Gulf Coast Basin. *Applied Computing and Geosciences*, *3*, 100012.

Fleming, L. E., & Tierney, J. E. (2016). An automated method for the determination of the TEX_{86} and paleotemperature indices. *Org. Geochem.*, *92*, 84–91.

Greenwell, B., Boehmke, B., & Gray, B. (2020). Package 'vip'. *Variable Importance Plots*, *12*(1), 343–66.

Guo, J., Glendell, M., Meersmans, J., Kirkels, F., Middelburg, J. J., & Peterse, F. (2020). Assessing branched tetraether lipids as tracers of soil organic carbon transport through the Carminowe Creek catchment (southwest England). *Biogeosciences*, *17*(12), 3183–3201.

Guo, J., Glendell, M., Meersmans, J., Kirkels, F. M. S. A., Middelburg, J. J., & Peterse, F. (2020). *Branched tetraether lipids in Carminowe Creek catchment (southwest England)* [Dataset]. PANGAEA. doi: 10.1594/PANGAEA.918523

Guo, J., Ma, T., Liu, N., Zhang, X., Hu, H., Ma, W., ... Peterse, F. (2021). *Branched tetraether lipids and bacterial communities along an aridity soil transect in Inner Mongolia, northern China* [Dataset]. PANGAEA. doi: 10.1594/PANGAEA.938067

Guo, J., Ma, T., Liu, N., Zhang, X., Hu, H., Ma, W., ... Peterse, F. (2022). Soil pH and aridity influence distributions of branched tetraether lipids in grassland soils along an aridity transect. *Organic Geochemistry*, 104347.

Guo, J., Yuan, H., Song, J., Li, X., Duan, L., Li, N., & Wang, Y. (2022). Influence of bottom seawater oxygen on archaeal tetraether lipids in sediments: Implications for archaeal lipid-based proxies. *Marine Chemistry*, 104138.

Halamka, T. A., McFarlin, J. M., Younkin, A. D., Depoy, J., Dildar, J., & Kopf, S. H. (2021). Oxygen limitation can trigger the production of branched GDGTs in culture. *Geochemical Perspectives Letters*, *19*, 36 – 39.

Halamka, T. A., Raberg, J. H., McFarlin, J. M., Younkin, A. D., Mulligan, C., Liu, X.-L., & Kopf, S. H. (2022). Production of diverse brGDGTs by Acidobacterium Solibacter usitatus in response to temperature, pH, and O_2 provides a culturing perspective on br GDGT proxies and biosynthesis. *Geobiology*.

Hamblin, A., Stasiuk, L., Sweet, A., Lockhart, G., Dyck, D., Jagger, K., & Snowdon, L. (2003). Post-kimberlite Eocene strata within a crater basin, Lac de Gras, Northwest Territories, Canada. In *International kimberlite conference: Extended abstracts* (Vol. 8).

Hopmans, E. C., Schouten, S., & Damsté, J. S. S. (2016). The effect of improved chromatography on GDGT-based palaeoproxies. *Organic Geochemistry*, *93*, 1–6.

Hopmans, E. C., Weijers, J. W., Schefuß, E., Herfort, L., Damsté, J. S. S., & Schouten, S. (2004). A novel proxy for terrestrial organic matter in sediments based on branched and isoprenoid tetraether lipids. *Earth and Planetary Science Letters*, *224*(1-2), 107–116.

Huguet, C., Hopmans, E. C., Febo-Ayala, W., Thompson, D. H., Sinninghe Damsté, J. S., & Schouten, S. (2006). An improved method to determine the absolute abundance of glycerol dibiphytanyl glycerol tetraether lipids. *Org. Geochem.*, *37*(9), 1036–1041.

Inglis, G. N., Farnsworth, A., Collinson, M. E., Carmichael, M. J., Naafs, B. D. A., Lunt, D. J., ... Pancost, R. D. (2019). Terrestrial environmental change across the onset of the PETM and the associated impact on biomarker proxies: A cautionary tale. *Global and Planetary Change*, *181*, 102991.

Inglis, G. N., Farnsworth, A., Collinson, M. E., Carmichael, M. J., Naafs, B. D. A.,

Lunt, D. J., ... Pancost, R. D. (2019). *Terrestrial environmental change across the onset of the PETM and the associated impact on biomarker proxies: a cautionary tale* [Dataset]. PANGAEA. Retrieved from `https://doi.org/10.1594/PANGAEA.901285` doi: 10.1594/PANGAEA.901285

Jahren, A. H., & Sternberg, L. S. L. (2003). Humidity estimate for the middle Eocene Arctic rain forest. *Geology*, *31*(5), 463–466.

Kim, J.-H., Van der Meer, J., Schouten, S., Helmke, P., Willmott, V., Sangiorgi, F., ... Sinninghe Damsté, J. S. J. (2010). New indices and calibrations derived from the distribution of crenarchaeal isoprenoid tetraether lipids: Implications for past sea surface temperature reconstructions. *Geochimica et Cosmochimica Acta*, *74*(16), 4639–4654.

Kirkels, F. M., Ponton, C., Galy, V., West, A. J., Feakins, S. J., & Peterse, F. (2020). From Andes to Amazon: Assessing branched tetraether lipids as tracers for soil organic carbon in the Madre de Dios River system. *Journal of Geophysical Research: Biogeosciences*, *125*(1), e2019JG005270.

Kirkels, F. M., Usman, M. O., & Peterse, F. (2022). Distinct sources of bacterial branched GMGTs in the Godavari River basin (India) and Bay of Bengal sediments. *Organic Geochemistry*, *167*, 104405.

Kirkels, F. M., Zwart, H. M., Usman, M. O., Hou, S., Ponton, C., Giosan, L., ... others (2022). From soil to sea: sources and transport of organic carbon traced by tetraether lipids in the monsoonal godavari river, india. *Biogeosciences*, *19*(17), 3979–4010.

Kuhn, M., & Wickham, H. (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. [Computer software manual]. Retrieved from `https://www.tidymodels.org`

Kumar, D. M., Woltering, M., Hopmans, E. C., Damste, J. S. S., Schouten, S., & Werne, J. P. (2019). The vertical distribution of Thaumarchaeota in the water column of Lake Malawi inferred from core and intact polar tetraether lipids. *Organic Geochemistry*, *132*, 37–49.

Langworthy, T. A. (1977). Long-chain diglycerol tetraethers from Thermoplasma acidophilum. *Biochimica et Biophysica Acta (BBA)-Lipids and Lipid Metabolism*, *487*(1), 37–50.

Lauretano, V., Kennedy-Asser, A. T., Korasidis, V. A., Wallace, M. W., Valdes, P. J., Lunt, D. J., ... Naafs, B. D. A. (2021). Eocene to oligocene terrestrial southern hemisphere cooling caused by declining p co2. *Nature Geoscience*, *14*(9), 659–664.

Li, J., Pancost, R. D., Naafs, B. D. A., Yang, H., Zhao, C., & Xie, S. (2016). Distribution of glycerol dialkyl glycerol tetraether (gdgt) lipids in a hypersaline lake system. *Organic Geochemistry*, *99*, 113–124.

Liu, X., Lipp, J. S., & Hinrichs, K.-U. (2011). Distribution of intact and core GDGTs in marine sediments. *Organic Geochemistry*, *42*(4), 368–375.

Locarnini, M., Mishonov, A., Baranova, O., Boyer, T., Zweng, M., Garcia, H., ... others (2018). World ocean atlas 2018, volume 1: Temperature.

Maechler, M., et al. (2019). Finding groups in data": Cluster analysis extended Rousseeuw et al. *R package version*, *2*(0).

Martínez-Sosa, P., Tierney, J. E., & Meredith, L. K. (2020). Controlled lacustrine microcosms show a brGDGT response to environmental perturbations. *Org. Geochem.*, 104041.

Martínez-Sosa, P., Tierney, J. E., Stefanescu, I. C., Crampton-Flood, E. D., Shuman, B. N., & Routson, C. (2021). A global Bayesian temperature calibration for lacustrine brGDGTs. *Geochimica et Cosmochimica Acta*, *305*, 87–105.

Martínez-Sosa, P., Tierney, J., Pérez-Angel, L., Stefanescu, I. C., Guo, J., Kierkels, F., ... Reyes, A. V. (2023). *BIGMaC GDGT algorithm* [Software]. Zenodo. (V. 1.0) doi: 10.5281/zenodo.7513557

Martínez-Sosa, P., Tierney, J., Pérez-Angel, L., Stefanescu, I. C., Guo, J., Kirkels,

F., ... Reyes, A. V. (2023b). *Giraffe kimberlite pipe core GDGT fractional abundance* [Dataset]. Zenodo. doi: 10.5281/zenodo.7540094

Martínez-Sosa, P., Tierney, J., Pérez-Angel, L. C., Stefanescu, I. C., Guo, J., Kirkels, F., ... Reyes, A. V. (2023a). *Environmental data and fractional abundance of iso and branched GDGT data used to train the BIGMaC algorithm* [Dataset]. Zenodo. doi: 10.5281/zenodo.7522415

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, *10*, 1–16.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2020). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, 2018, R package version 1.7-0.*

Naafs, B., Rohrssen, M., Inglis, G. N., Lähteenoja, O., Feakins, S. J., Collinson, M. E., ... others (2018). High temperatures in the terrestrial mid-latitudes during the early Palaeogene. *Nature Geoscience*, *11*(10), 766–771.

Naafs, B. D. A. (2017). *Global biomarker (GDGT) database for peatlands* [Dataset]. PANGAEA. Retrieved from `https://doi.org/10.1594/PANGAEA.883765` doi: 10.1594/PANGAEA.883765

Naafs, B. D. A., Inglis, G. N., Zheng, Y., Amesbury, M., Biester, H., Bindler, R., ... others (2017). Introducing global peat-specific temperature and pH calibrations based on brGDGT bacterial lipids. *Geochimica et Cosmochimica Acta*, *208*, 285–301.

Naeher, S., Peterse, F., Smittenberg, R. H., Niemann, H., Zigah, P. K., & Schubert, C. J. (2014). Sources of glycerol dialkyl glycerol tetraethers (GDGTs) in catchment soils, water column and sediments of Lake Rotsee (Switzerland)– Implications for the application of GDGT-based proxies for lakes. *Organic Geochemistry*, *66*, 164–173.

Parmar, A., Katariya, R., & Patel, V. (2019). A review on random forest: An ensemble classifier. In *International conference on intelligent data communication technologies and internet of things (icici) 2018* (pp. 758–763).

Peaple, M. D., Tierney, J. E., McGee, D., Lowenstein, T. K., Bhattacharya, T., & Feakins, S. J. (2021). Identifying plant wax inputs in lake sediments using machine learning. *Organic Geochemistry*, *156*, 104222.

Pérez-Angel, L. C., Sepúlveda, J., Molnar, P., Montes, C., Rajagopalan, B., Snell, K., ... Dildar, N. (2020). Soil and air temperature calibrations using branched GDGTs for the Tropical Andes of Colombia: Toward a pan-tropical calibration. *Geochemistry, Geophysics, Geosystems*, *21*(8), e2020GC008941.

Pérez-Angel, L. C., Sepúlveda, J., Molnar, P., Montes, C., Rajagopalan, B., Snell, K., ... Dildar, N. (2020). *In situ temperature and brGDGTs measurements in soils from the Tropical Andes of Colombia and a tropical soil brGDGT compilation dataset* [Dataset]. Zenodo. doi: 10.5281/zenodo.3939270

R Core Team. (2022). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Raberg, J. H., Miller, G. H., Geirsdóttir, Á., & Sepúlveda, J. (2022). Near-universal trends in brGDGT lipid distributions in nature. *Science Advances*, *8*(20), eabm7625.

Rattanasriampaipong, R., Zhang, Y. G., Pearson, A., Hedlund, B. P., & Zhang, S. (2022). Archaeal lipids trace ecology and evolution of marine ammonia-oxidizing archaea. *Proceedings of the National Academy of Sciences*, *119*(31), e2123193119.

Revelle, W., & Revelle, M. W. (2015). Package 'psych'. *The comprehensive R archive network*, *337*, 338.

Schouten, S., Hopmans, E. C., & Damsté, J. S. S. (2013). The organic geochemistry

of glycerol dialkyl glycerol tetraether lipids: A review. *Organic geochemistry*, *54*, 19–61.

Schouten, S., Hopmans, E. C., Schefuß, E., & Damste, J. S. S. (2002). Distributional variations in marine crenarchaeotal membrane lipids: a new tool for reconstructing ancient sea water temperatures? *Earth and Planetary Science Letters*, *204*(1-2), 265–274.

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In *Emerging technology in modelling and graphics: Proceedings of iem graph 2018* (pp. 99–111).

Sinninghe Damsté, J. S., Rijpstra, W. I. C., Hopmans, E. C., den Uijl, M. J., Weijers, J. W., & Schouten, S. (2018). The enigmatic structure of the crenarchaeol isomer. *Organic Geochemistry*, *124*, 22–28.

Sinninghe Damsté, J. S., Rijpstra, W. I. C., Hopmans, E. C., Weijers, J. W., Foesel, B. U., Overmann, J., & Dedysh, S. N. (2011). 13, 16-Dimethyl octacosanedioic acid (iso-diabolic acid), a common membrane-spanning lipid of Acidobacteria subdivisions 1 and 3. *Applied and Environmental Microbiology*, *77*(12), 4147–4154.

Sinninghe Damsté, J. S., Schouten, S., Hopmans, E. C., Van Duin, A. C., & Geenevasen, J. A. (2002). Crenarchaeol: the characteristic core glycerol dibiphitanyl glycerol tetraether membrane lipid of cosmopolitan pelagic crenarchaeota. *Journal of lipid research*, *43*(10), 1641–1651.

Sinninghe Damsté, J. S., Weber, Y., Zopfi, J., Lehmann, M. F., & Niemann, H. (2022). Distributions and sources of isoprenoidal GDGTs in Lake Lugano and other central European (peri-) alpine lakes: Lessons for their use as paleotemperature proxies. *Quaternary Science Reviews*, *277*, 107352.

Siver, P. A., & Lott, A. M. (2023). History of the Giraffe Pipe locality inferred from microfossil remains: a thriving freshwater ecosystem near the Arctic Circle during the warm Eocene. *Journal of Paleontology*, *97*(2), 271–291.

Taylor, K. W., Huber, M., Hollis, C. J., Hernandez-Sanchez, M. T., & Pancost, R. D. (2013). Re-evaluating modern and Palaeogene GDGT distributions: Implications for SST reconstructions. *Global and Planetary Change*, *108*, 158–174.

Tierney, J. E., Russell, J. M., Eggermont, H., Hopmans, E., Verschuren, D., & Sinninghe Damsté, J. S. (2010). Environmental controls on branched tetraether lipid distributions in tropical East African lake sediments. *Geochim. Cosmochim. Acta*, *74*(17), 4902–4918.

Ueki, K., Hino, H., & Kuwatani, T. (2018). Geochemical discrimination and characteristics of magmatic tectonic settings: A machine-learning-based approach. *Geochemistry, Geophysics, Geosystems*, *19*(4), 1327–1347.

Véquaud, P., Thibault, A., Derenne, S., Anquetil, C., Collin, S., Contreras, S., ... Huguet, A. (2022). FROG: A global machine-learning temperature calibration for branched GDGTs in soils and peats. *Geochimica et Cosmochimica Acta*, *318*, 468–494.

Weijers, J. W., Schouten, S., Hopmans, E. C., Geenevasen, J. A., David, O. R., Coleman, J. M., ... Sinninghe Damsté, J. S. (2006). Membrane lipids of mesophilic anaerobic bacteria thriving in peats have typical archaeal traits. *Environmental Microbiology*, *8*(4), 648–657.

Weijers, J. W., Schouten, S., van den Donker, J. C., Hopmans, E. C., & Damsté, J. S. S. (2007). Environmental controls on bacterial tetraether membrane lipid distribution in soils. *Geochimica et Cosmochimica Acta*, *71*(3), 703–713.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686

Windler, G., Tierney, J. E., DiNezio, P. N., Gibson, K., & Thunell, R. (2019). Shelf exposure influence on Indo-Pacific Warm Pool climate for the last 450,000

years. *Earth and Planetary Science Letters*, *516*, 66–76.

Wolfe, A. P., Reyes, A. V., Royer, D. L., Greenwood, D. R., Doria, G., Gagen, M. H., ... Westgate, J. A. (2017). Middle Eocene CO_2 and climate reconstructed from the sediment fill of a subarctic kimberlite maar. *Geology*, *45*(7), 619–622.

Wright, M. N., Wager, S., Probst, P., & Wright, M. M. N. (2019). Package 'ranger'. *Version 0.11*, *2*.

Yavitt, J., Harms, K., Garcia, M., Wright, S., He, F., & Mirabello, M. (2009). Spatial heterogeneity of soil chemical properties in a lowland tropical moist forest, Panama. *Soil Research*, *47*(7), 674–687.

Zell, C., Kim, J.-H., Moreira-Turcq, P., Abril, G., Hopmans, E. C., Bonnet, M.-P., ... Damsté, J. S. S. (2013). Disentangling the origins of branched tetraether lipids and crenarchaeol in the lower Amazon River: Implications for GDGT-based proxies. *Limnology and Oceanography*, *58*(1), 343–353.

Zhang, Y. G., Zhang, C. L., Liu, X.-L., Li, L., Hinrichs, K.-U., & Noakes, J. E. (2011). Methane Index: A tetraether archaeal lipid biomarker indicator for detecting the instability of marine gas hydrates. *Earth and Planetary Science Letters*, *307*(3-4), 525–534.

Zhao, B., Castañeda, I. S., Salacup, J. M., Thomas, E. K., Daniels, W. C., Schneider, T., ... Bradley, R. S. (2022). Prolonged drying trend coincident with the demise of norse settlement in southern greenland. *Science advances*, *8*(12), eabm4346.

Zheng, Y., Heng, P., Conte, M. H., Vachula, R. S., & Huang, Y. (2019). Systematic chemotaxonomic profiling and novel paleotemperature indices based on alkenones and alkenoates: Potential for disentangling mixed species input. *Organic Geochemistry*, *128*, 26–41.

Zheng, Y., Liu, H., Yang, H., Wang, H., Zhao, W., Zhang, Z., ... Liu, W. (2022). Decoupled Asian monsoon intensity and precipitation during glacial-interglacial transitions on the Chinese Loess Plateau. *Nature Communications*, *13*(1), 5397.

Zheng, Y., Pancost, R. D., Liu, X., Wang, Z., Naafs, B., Xie, X., ... Yang, H. (2017). Atmospheric connections with the north atlantic enhanced the deglacial warming in northeast china. *Geology*, *45*(11), 1031–1034.