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**[2], **Francien Peterse**[4], **Bryan N. Shuman**[3], **Alberto V. Reyes**[5]

[1]Department of Geosciences, The University of Arizona, 1040 E 4th St, Tucson, Arizona 85721, USA
[2]Department of Geological Sciences and Institute of Arctic and Alpine Research (INSTAAR), University of Colorado, Colorado, USA
[3]Department of Geology and Geophysics, University of Wyoming, Wyoming, USA
[4]Department of Earth Sciences, Utrecht University, Utrecht, Netherlands
[5]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 they also have unique responses to environmental factors.
- The BIGMaC algorithm captures the correlation 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), including both the archaeal isoprenoid GDGTs (isoGDGTs) and the bacterial branched GDGTs (brGDGTs), have been used in paleoclimate studies to reconstruct temperature in marine and terrestrial archives. However, GDGTs are present in many different types of environments, with relative abundances that strongly depend on the depositional setting. This suggests that GDGT distributions can be used more broadly to infer paleoenvironments in the geological past. In this study, we analyzed 1153 samples from a variety of modern sedimentary settings for both isoGDGT and brGDGTs. We used machine learning on the GDGT relative abundances from this dataset to relate the lipid distributions to the physical and chemical characteristics of the depositional settings. We observe a robust relationship between the depositional environment and the lipid distribution profiles of our samples. This dataset was used to train and test the **B**ranched and **I**soprenoid **G**DGT **Ma**chine learning **C**lassification algorithm (BIGMaC), which identifies the environment a sample comes from based on the distribution of GDGTs with high accuracy. 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 information, provides new information about the paleoenvironment of this site, and helps improve paleotemperature reconstruction. In cases where paleoenvironments are unknown or are changing, BIGMaC can be applied in concert with other proxies to generate more refined paleoclimatic 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 such as the alkyl chains, the stereochemistry of the glycerol group (Weijers et al., 2006), and most importantly, culture studies (Chen et al., 2022; Halamka et al., 2022, 2021; Sinninghe Damsté et al., 2011), they are considered to have a bacterial source.

In contrast, isoprenoid glycerol dibiphytanyl glycerol tetraether GDGTs (isoGDGTs) are produced by Archaea (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 an identified stereoisomer (Sinninghe Damsté et al., 2002, 2018).

Both isoprenoid and branched GDGTs are used in paleoclimate studies as their distribution follows 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, Schouten, et al., 2007). The Methylation index of Branched Tetraethers (MBT'_{5Me}) proxy isolates the relationship between the methylation of brGDGTs and temperature (De Jonge, Hopmans, et al., 2014) and has been widely used for ter-

restrial paleoclimate reconstructions (Pancost et al., 2013; Peterse et al., 2012; Weijers, Schefuß, et al., 2007).

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 organisms to the isoGDGT pool compared with members of Thaumarchaeota (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 from (Rattanasriampaipong et al., 2022; Taylor et al., 2013). 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, has been proposed to broadly discriminate between marine and terrestrial environments (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 can in turn be used to discriminate between samples formed in terrestrial or marine environments, as well as whether terrestrial samples were formed in 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 improve our application of GDGT-based paleotemperature proxies by better constraining which environmental setting the lipids are coming from. This requires characterizing multidimensional, nonlinear relationships between the occurrence and distribution of GDGT lipids and their source environment, as well as a framework that allows researchers to easily apply these relationships to new unclassified samples.

To address and incorporate all of these factors, we make use of machine learning, which provides a way to model highly dimensional and nonlinear data with complex interactions and missing values (El Bouchefry & de Souza, 2020). Machine learning has previously been used in the Geosciences to discriminate between magma (Ueki et al., 2018) as well as water (Engle & Brunner, 2019) sources. Similarly, these tools have also been specifically applied to biomarkers and GDGTs (Véquaud et al., 2022; Peaple et al., 2021; Zheng et al., 2019). Here, we use a compilation of 1153 globally dispersed samples 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 GDGTs. We further 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, as well as the limitations of this approach in an application to a peatland dataset that spans the Paleocene-Eocene Thermal Maximum (PETM).

2 Materials and Methods

2.1 Global Dataset

We compiled a total of 1153 globally distributed (Fig. 1) samples from different depositional environments: coastal, 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 paleoenvironmen-

tal 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, 2017), 215 are marine and coastal 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 sediment (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 included in the original dataset.

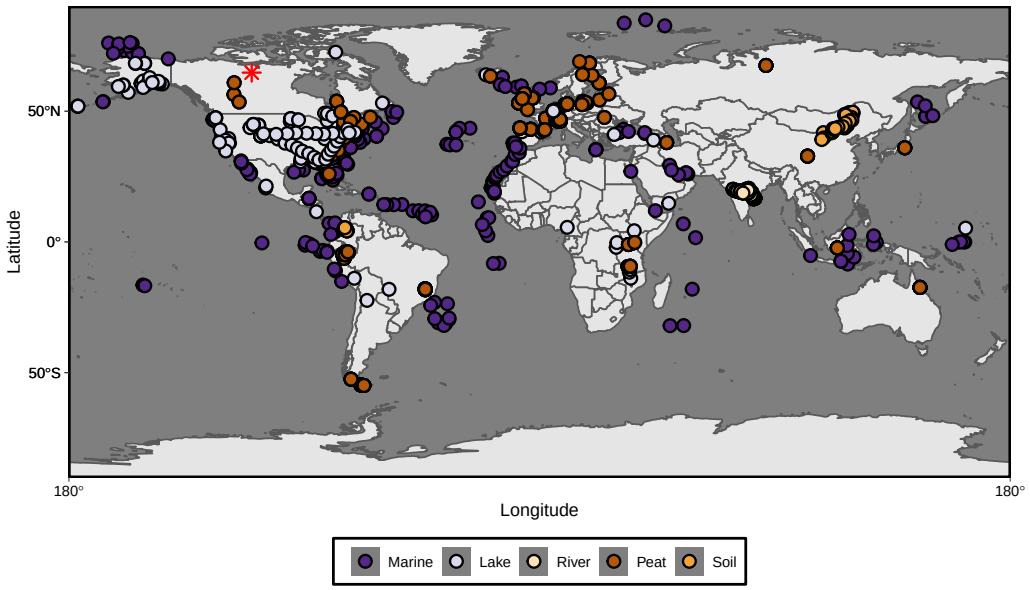

Figure 1. World map showing the distribution of the samples included in this work. Color code reflects the depositional environment which these samples were collected from. Red asterisk shows the modern location of the Giraffe pipe.

All marine sediment samples were processed at the University of Arizona following the method used in Martínez-Sosa et al. (2021). Briefly, samples 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), and 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 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). 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 2.2.1).

We collected the environmental parameters associated with the samples using the data available in the source datasets. 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 Machine Learning

For our machine learning analyses we use two different but complementary approaches. We first performed unsupervised machine learning on the dataset (with the samples' depositional environment unlabeled), which allows for the exploration of complex patterns presented by the predictor variables (GDGT abundance). The end product of this section is the identification of the major GDGT-derived clusters. Next, we applied supervised machine learning, where the dataset is split into a training set and a test set, and the environment of each sample is assigned to one of the major clusters identified in the unsupervised step. The training set is used to map the relationship between the predictor variables to the response variable (the environment). The test set is then used to evaluate the performance of the mapped relationship.

For this work, all analyses were performed in `R` (R Core Team, 2022).

2.2.1 Unsupervised Machine Learning

For the unsupervised machine learning analysis we centered and scaled the fractional abundances of GDGTs across the whole dataset. We tested the optimal number of clusters for this dataset using the `fviz_nbclust()` function of the *factoextra* package (Kassambara & Mundt, 2020) and by performing a silhouette analysis using the `pam()` (Partitioning Around Medoids) method from the *cluster* package (Maechler et al., 2019). Samples were separated into clusters by applying the fuzzy version of the k-means clustering algorithm using the `cmeans()` function from the *e1071* package (Meyer et al., 2020). The best performing number of clusters from the silhouette analysis was used and the analysis was iterated a maximum of 100 times.

Following the cluster analysis and prior to the supervised machine learning, we curated the identified groups by hand, reassigning any samples that were incorrectly classified to their correct (real-world) environment. This preserves the natural variability in the samples that ultimately contributes to some amount of error in the classification model.

2.2.2 Supervised Machine Learning

For the supervised machine learning we worked in the *tidymodels* and *tidyverse* environments (Kuhn & Wickham, 2020; Wickham et al., 2019), where we used the fractional abundances of GDGTs as predictor variables and the curated classification from the previous unsupervised step as the response variables. The dataset was split in a 3:1 ratio, preserving the distribution of sample types, for the training and test sets using the function `initial_split()` from the *rsample* package (Kuhn et al., 2019). We further generated a validation set from the training set with 10 partitions for tuning the hyperparameters—parameters whose values control the learning process—using the `vfold_cv()` function from the *rsample* package.

We tested the performance of four different classification models (Random Forest, XGBoost, K Nearest Neighbour and Naive Bayes) plus a control non-informative (null) model. Hyperparameters for each model, except XGBoost, were tested using a regular grid through the `grid_regular()` function from the *dials* package (Kuhn, 2020a). The

hyperparameters for the XGBoost model were selected using a latin hypercube design with 30 parameter value combinations using the `grid_latin_hypercube()` function from the *dials* package. The hyperparameter tuning was run at the University of Arizona High-Performance Computing facility. Finally, the best hyperparameter values were selected by comparing their ROC-AUC score on the validation set (Table S1).

We tested the performance of each model with the best hyperparameter combination on the validation set and selected the model that produced the best F1 and ROC-AUC score. This model was then trained and tested using the `last_fit()` function from the *tune* package (Kuhn, 2020b).

2.3 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.

2.4 Cobham Lignite Bed

The Cobham lignite bed, Kent, UK ($\sim 48°$N palaeolatitude) is composed by 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, which was then used to perform a fuzzy k-means classification. This analysis separated the dataset into four groups consisting between 219 and 465 samples each. When we compare the composition of each cluster using Principal Component Analysis (PCA), we observe clear differences between depositional environments (Fig. 2a and b, and Table 1). 87% of the peat samples fall within Group 1, while 85% of the lacustrine samples are assigned to Group 2. In turn, 92% of the river samples are assigned to Group 3, and 92% 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 (44%).

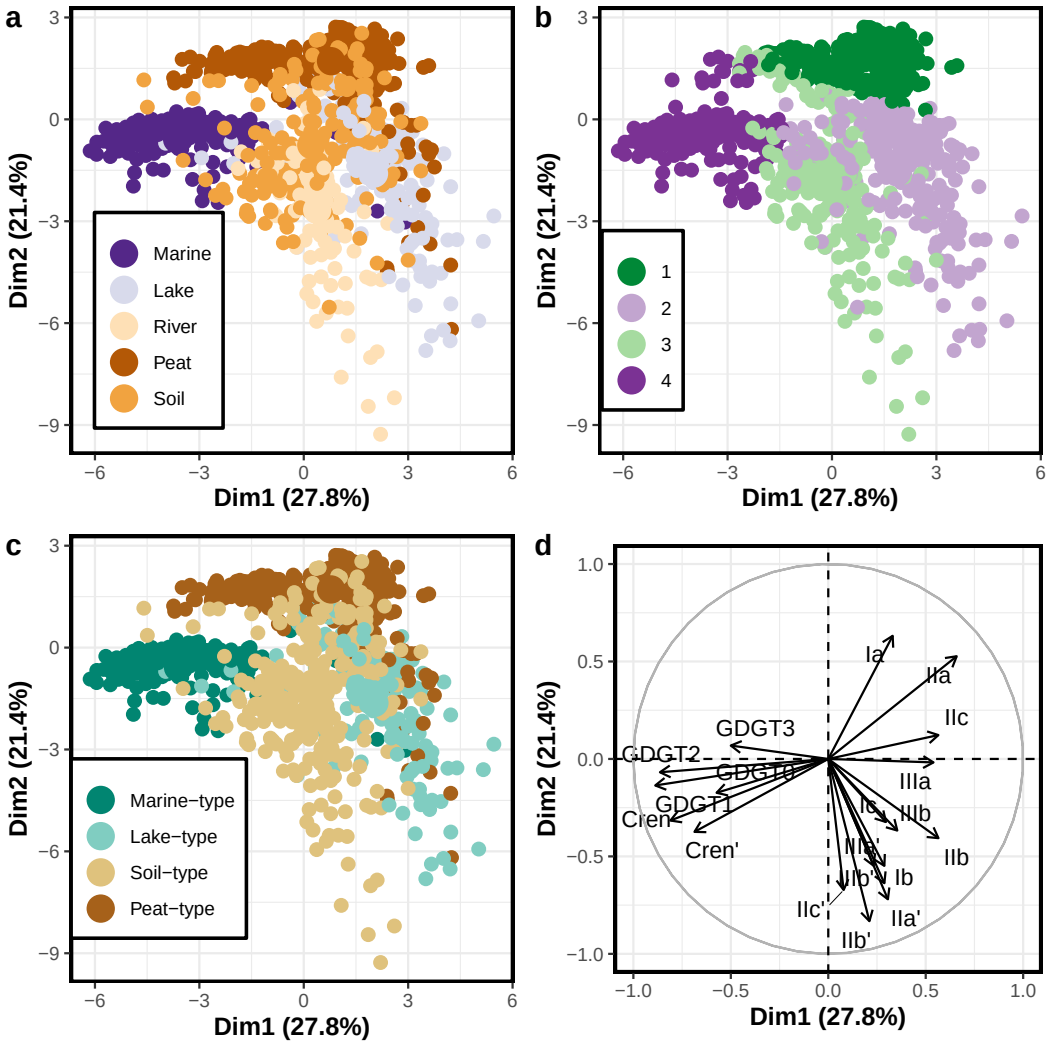

Figure 2. Samples from the dataset plotted in reduced dimensional space based on the fractional abundance of GDGTs. Plots show the same analysis with samples colored based on the depositional environment (a), their assigned group based on the fuzzy k-means analysis (b), and the hand-curated clusters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

Given the distinctive clustering, we renamed them based on the dominant depositional environment (Fig. 2b and c). Group 1 was renamed as *Peat-type*, Group 2 as *Lake-type*, Group 3 as *Soil-type*, and finally Group 4 as *Marine-type*. Samples for which the cluster assignment did not match their depositional environment were manually reassigned to the appropriate group (Table 1). For example the original dataset from Naafs (2017) includes only peats and so all samples from this dataset were reassigned as *Peat-type*, regardless of whether they fell in Group 1 or not. The k-means derived and manually curated clusters maintain their core distributions (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 peat samples and 20% of the soil samples; *Peat-type*, with 476 samples, has all of the peat samples and only one lake sample. Similarly, Group 4, with 225 samples, had 92% of the marine samples, while *Marine-type* includes all of

Table 1. Percentage of each type of sample assigned to each of the four clusters determined by fuzzy k-means analysis (top) as well as the four manually curated clusters (bottom). At the bottom is the total number of samples from each type, and the last column shows the total number of samples in each cluster (fuzzy k-means and curated). The highest percentage for each type of sample in the clusters is indicated in bold.

	Lake	Marine	Peat	River	Soil	Total
Group 1	7.4%	0%	**87%**	0%	20.4%	465
Group 2	**85%**	6%	6%	8%	31%	244
Group 3	6%	3%	4.4%	**92.4%**	**44%**	219
Group 4	3%	**92%**	3%	0%	5.1%	225
Peat-type	0.6%	0%	**100%**	0%	0%	476
Lake-type	**97.5%**	0%	0%	0%	0%	158
Soil-type	1.2%	0%	0%	**100%**	**100%**	303
Marine-type	0.6%	**100%**	0%	0%	0%	216
Total	*162*	*215*	*475*	*105*	*196*	

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, peat and soil samples. 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 soil samples from Group 2 to *Soil-type*.

3.2 Within-Group Analyses

Once the unsupervised machine learning demonstrated that the dataset can be differentiated into *Marine-type*, *Lake-Type*, *Soil-type*, and *Peat-type* groups, we analyzed the GDGT distribution of each group to assess their influence on the clustering results as well as how well they correlated with environmental parameters.

3.2.1 GDGT Distribution

Across the entire dataset, we observe that 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 four groups. *Marine-type* samples have a higher proportion of crenarchaeol and GDGT-0 compared with the other groups (Fig. 3a). As previously reported (Martínez-Sosa et al., 2021), *Lake-type* samples show a higher proportion of IIIa and lower Ia than both soils and peats (Fig. 3b and c). While our data also shows that from the terrestrial groups, *Soil-type* has a preference for 6-methyl isomers, in contrast to *Lake-type* and *Peat-type*; an analysis of the brGDGT distribution of just the *Soil-type* samples shows that it is the river samples that contain a higher proportion of 6-methyl brGDGTs, while soils have a higher proportion of 5-methyl isomers (Fig. S1). Additionally, while the proportion of isoGDGTs is generally low in the terrestrial groups, *Soil-type* samples show a higher proportion of crenarchaeol than *Lake-type* and *Peat-type* samples, but lower than *Marine-type* (Fig. 3a).

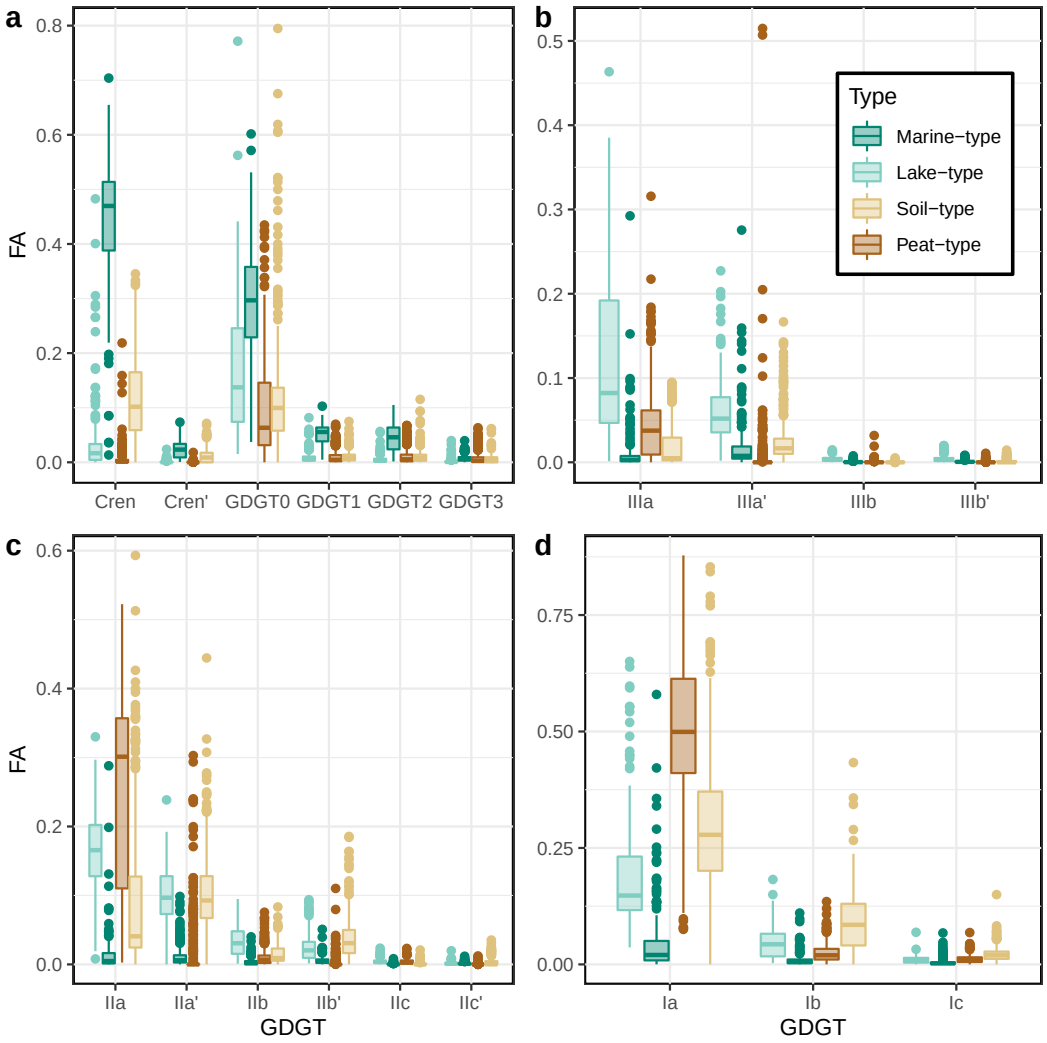

Figure 3. Box plots showing the distribution of the fractional abundance (FA) of all GDGTs in each of the curated clusters, following the color code of Figure 2. GDGTs separated by isoGDGTs (a), hexamethylated brGDGTs (b), pentamethylated brGDGTs (c), and tetramethylated brGDGTs (d).

3.2.2 GDGT Influence

To better understand the effect that each compound has on each group, we performed a Non-Metric Multidimensional Scaling (NMDS) on the fractional abundance of GDGTs (Fig. 4). For this analysis, we excluded four outlier samples from the *Marine-type* group: AII72-BC21 (North Atlantic), U (Port Wells, Alaska), CHN752-PC7 (North Atlantic), and FISH-1 (Long Island Sound) as they strongly skewed the data. These samples had no relation to each other, spatial or otherwise. All NDMS analysis reach convergence for two dimensions with stress < 0.2.

The NMDS results show that for the *Marine-type* set (Fig. 4a and d) the first dimension is driven by a positive relation with isoGDGTs and a negative relation with brGDGTs. The second dimension, in turn, is mostly dominated by a negative relation with GDGT-0. We also observe a strong relationship ($\rho = 0.82$, Spearman's correlation) between

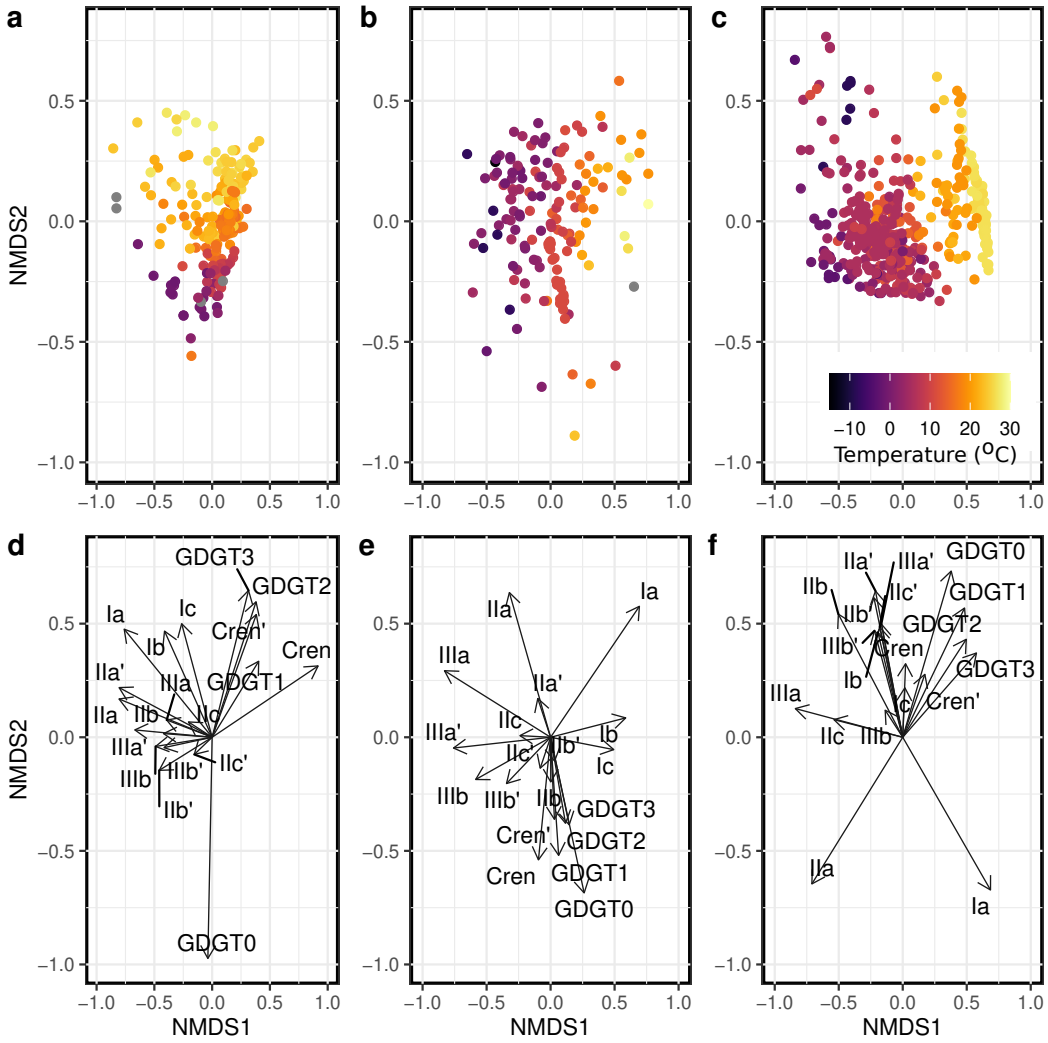

Figure 4. NMDS plots for *Marine-type* (a and d), *Lake-type* (b and e), and *Peat-type* (c and f). Panels a to c show the samples from each group colored based on mixed layer temperature (a), or MAAT (b and c), while panels d to f show the contribution of each GDGT to each group.

the second dimension and mixed layer temperature (Fig. 4a). For the *Lake-type* samples (Fig. 4b and e) the first dimension is dominated by a positive relation with the tetramethylated brGDGTs (Ia, Ib, and Ic) and a negative relation with the rest of the brGDGTs. The second dimension is driven by a negative relation with isoGDGTs and cyclic brGDGTs, and a positive relation with non-cyclic brGDGTs. The first dimension has a high correlation ($\rho = 0.83$) with mean annual air temperature (MAAT) (Fig. 4b), but we find no strong correlation ($\rho < |0.4|$) between the second dimension and any of the environmental parameters analyzed. Finally, the *Peat-type* set (Fig. 4c and f) shows a strong positive relation between Ia and the first dimension, and a negative relation with most of the other brGDGTs, closely following MAAT with a correlation of $\rho = 0.80$ (Fig. 4c). The second dimension has primarily a positive relation with Ia and IIa, while most of the other compounds show a negative relation, once again we were unable to find a strong correlation between this dimension and any environmental parameter. We do not discuss the NMDS results for the *Soil-type* samples because their spatial distribution is extremely limited (Fig. 1) and thus their location dominates the GDGT distributions. We

also do not observe any strong relationships between the NMDS dimensions and other additional environmental parameters, such as pH, elevation, and depth.

3.3 Supervised Machine Learning

The manually-curated labels generated after 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 we observe that 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. Finally, we observe the same result when measuring accuracy, where Random Forest performed the best (0.96), followed by XGBoost (0.94), K-Nearest Neighbour (0.92), Naive Bayes (0.88), and the null model (0.41). Based on these results we chose the Random Forest algorithm. We observe that the performance of this algorithm in the test set is similar to the one observed for the training set (0.94 and 0.99 for F1 and ROC-AUC respectively, Fig. 5). This result suggests that the algorithm is not overfitting the data.

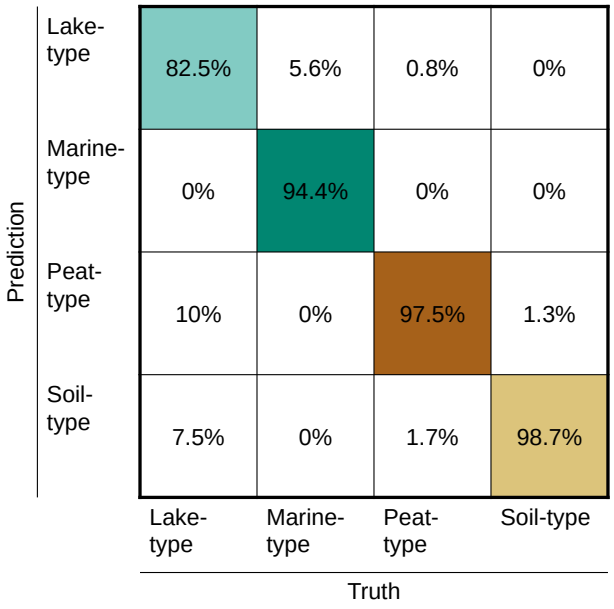

Figure 5. Confusion matrix showing the performance of the BIGMaC Random Forest algorithm in the test dataset. Columns show the true label of the samples and rows the predicted label. Diagonal cells are color-coded based on Fig. 2.

Finally, we diagnose the importance that each predictor variable has on the trained classification algorithm. We observe from this analysis 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. These values were calculated using the default values in the *ranger* package (Wright et al., 2019).

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

3.4 Applications

To demonstrate that the model can be successfully used to analyze changes in depositional environments through time, we test 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.

3.4.1 Giraffe Kimberlite Pipe

The lithology of the Giraffe kimberlite pipe core has previously been described, thus making it a good test case for the application of our classification algorithm. When we apply the BIGMaC algorithm to this core, we observe that the predicted cluster for each sample strongly aligns with the corresponding lithological section (Fig. 6). 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. 6).

To further investigate the results of our classification, the fractional abundance of brGDGTs was used to calculate CBT', 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) (Fig. 6b). We observe that in general the peat section has much lower CBT' values (associated with lower pH), than those observed in the lacustrine section. While this trend is maintained for most of the core, we observe a marked decrease in CBT' values in the section within the lacustrine facies that is classified as *Peat-type*.

Based on the BIGMaC classification, we applied either the global soil/peat calibration (Dearing Crampton-Flood et al., 2020) for samples classified as *Peat-type* and *Soil-type*, or the global lake calibration (Martínez-Sosa et al., 2021) for samples classified as *Lake-type*. Our compounded temperature reconstruction has a mean temperature of 19.1°C and a standard deviation of 3.2°C. Overall we observe a stable period with no clear trends in temperature. The mean difference in the predicted temperature for the entire core between the soil and lake calibrations is 6.7°C, with the lake calibration consistently generating higher temperatures. During the *Peat-type* excursion section the mean difference between both calibrations is 5.7°C.

3.4.2 Cobham Lignite Bed

Our application of the BIGMaC algorithm to the Cobham lignite bed shows a marked difference in the depositional environment prediction for the pre-PETM and PETM sections (Fig. 7). Almost all samples up to 54.15 cm 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, we observe 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 MBT'_{5Me} values, nor are they organized in any evident pattern within the unit.

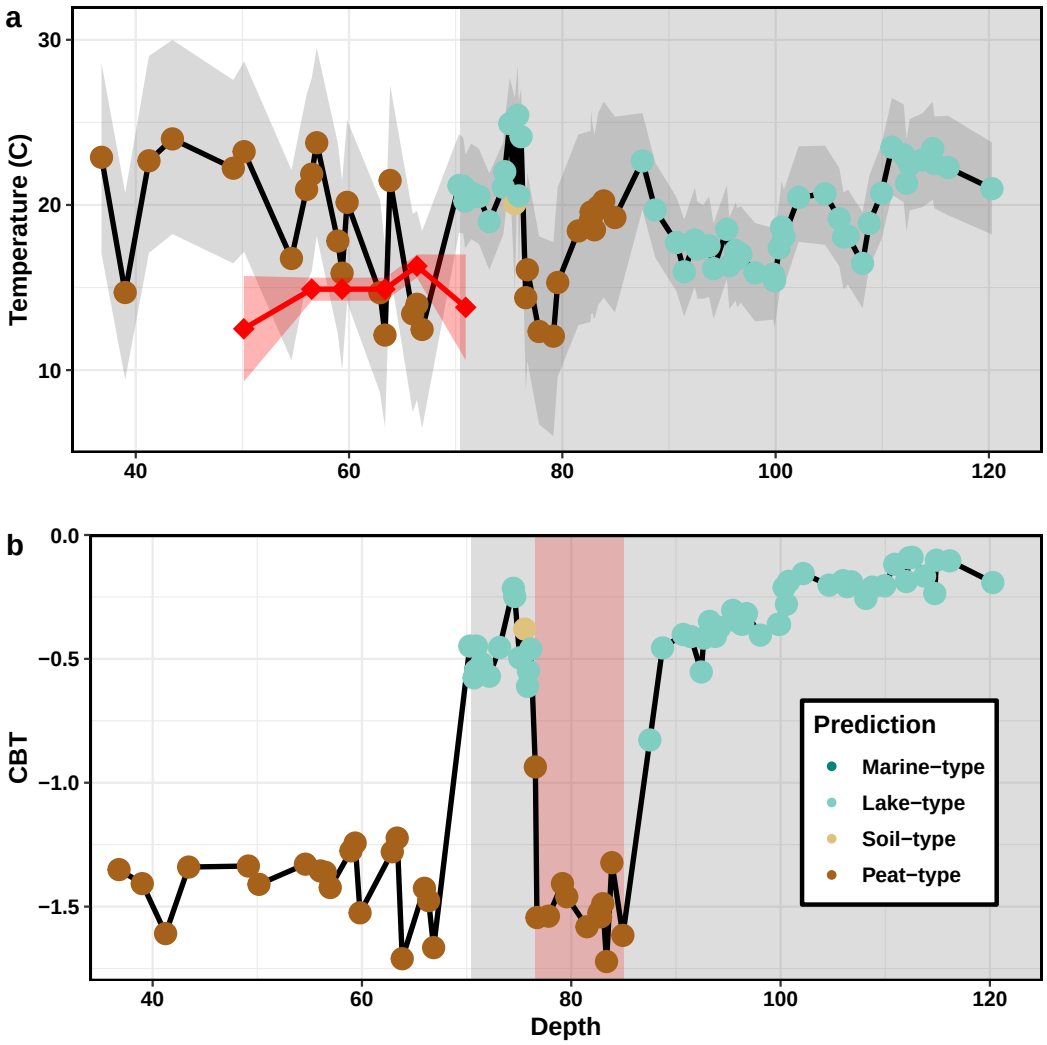

Figure 6. Inferred temperature (a) and CBT' (b) calculated from Giraffe core brGDGTs plotted against vertical-equivalent depth. The temperature reconstruction was generated by applying the Dearing Crampton-Flood et al. (2020) Bayesian calibration for *Peat* and *Soil-type* samples, and Martínez-Sosa et al. (2021) calibration for *Lake-type* samples. Palynological estimates of MAT with their associated error from Wolfe et al. (2017) are shown in red diamonds in (a). Samples are color-coded based on the predicted groups. White and gray shading indicates peat and lacustrine sediments in the core, respectively. The acid excursion is shaded in red (b).

4 Discussion

4.1 Unsupervised Machine Learning

The fuzzy k-means analysis shows that the compiled global dataset is best described by four clusters that are strongly defined by depositional environment (Table 2; Fig. 2). The marine samples form the most distinct cluster, which is probably driven by the higher abundance of isoGDGTs compared with other environments. The terrestrial environments (lakes, rivers, peats and soils) have GDGT distributions more closely related to each other but still form distinct clusters (except for rivers which cluster with soils) in agreement

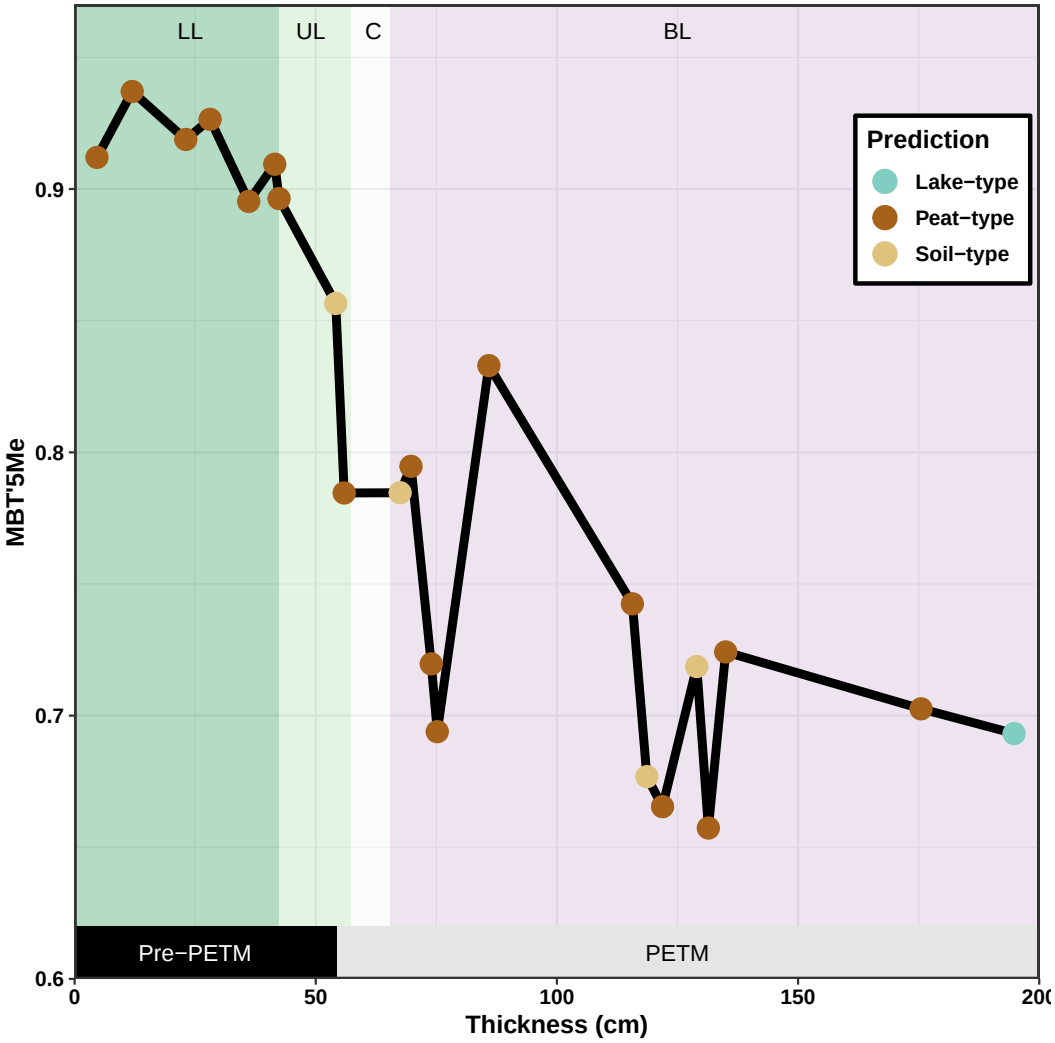

Figure 7. Calculated MBT'_{5Me} values of the Cobham lignite bed across the site thickness (cm). Samples are color coded based on the BIGMaC predicted groups. Different units are colored and labeled on the top as: lower laminated lignite (LL, dark green), upper laminated lignite (UL, light green), clay (C, white), and blocky lignite (BL, purple).

with previous work that has argued for clear differences between GDGTs in soils and lakes (Russell et al., 2018; Tierney et al., 2010; Tierney & Russell, 2009).

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 samples more closely resemble soils rather than peats or lakes. While this could be interpreted as soil-derived GDGTs dominating river inputs, our river data come from only two locations and primarily from only one system (the Godavari river) so this could be particular to that 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. 8). Membership to the soil-dominated Group 3 is higher in the lower Godavari basin, as well as from the wet (post-monsoon) season (Fig. 8 c and d). In contrast, membership to the lake-dominated Group 2 is overall higher in

the wet season, and in the upper basin year-round (Fig. 8 a and b). These results are in line with those presented in Kirkels, Zwart, et al. (2022), where it was noted that GDGTs from soils have a stronger influence on the river 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 samples from the dry season as well as those from the upper basin.

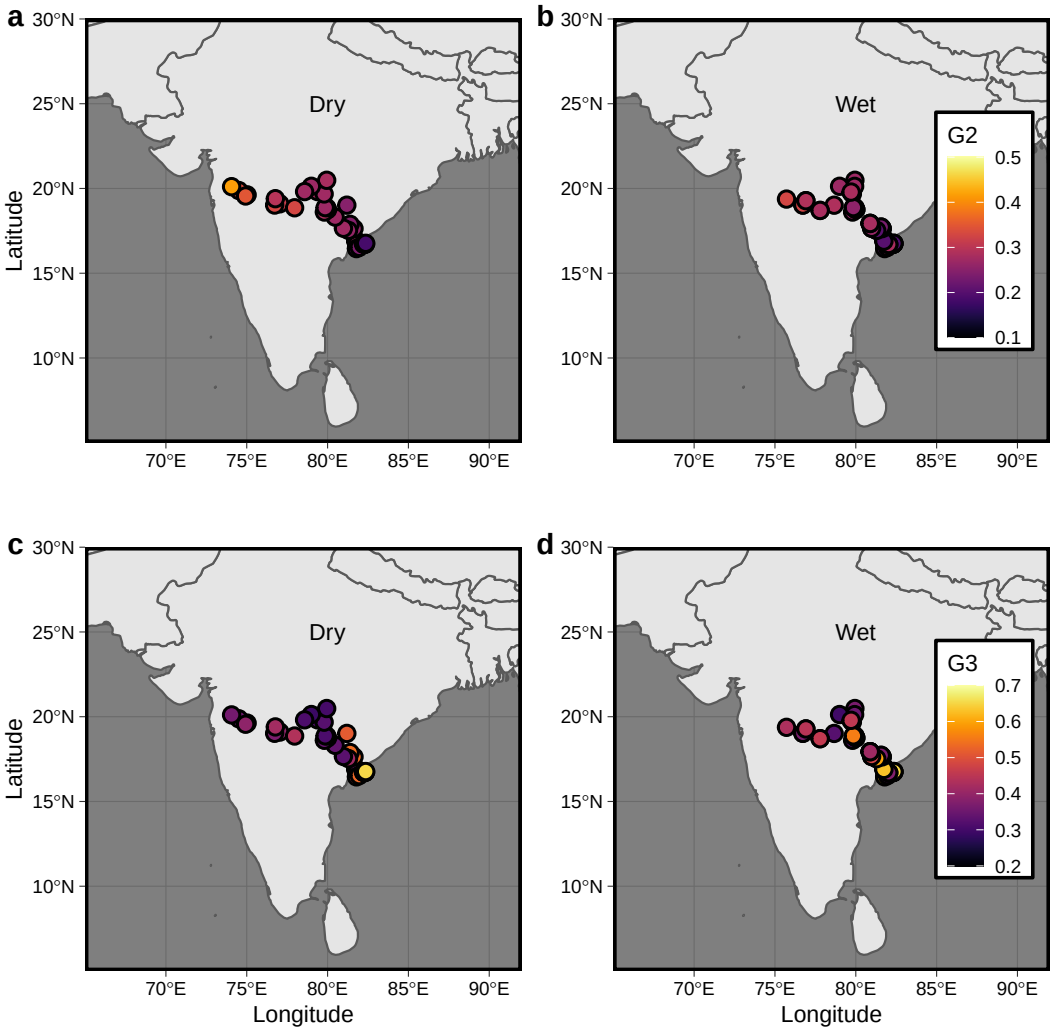

Figure 8. Maps for the Godavari River sample locations for the dry (left column) and wet (right column) seasons. Maps show the sample memberships, calculated through fuzzy k-means analysis, to the lake-dominated Group 2 (a and b), and to the soil-dominated Group 3 (c and d).

4.2 Manually Curated Clusters

While our fuzzy k-means clusters show strong patterns that reflect environmentally relevant relationships (Fig. 2a), some samples whose depositional environment had been unequivocally documented cluster in unrelated groups (*i.e.* soil samples plotting as peats). Since our intention with the supervised machine learning was to test whether GDGT distributions can be used to classify the true depositional environment, we manually re-assigned any samples that fell within the incorrect group. The manually curated clusters are very

similar to the statistical groupings (Fig. 2b) but preserve the "errors" (i.e., soils that look like peats) in the dataset, thus guarding against overfitting.

Soils are highly diverse environments with diffuse boundaries; they are often in contact with other depositional environments. 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). This may explain why soil samples are spread across most of the fuzzy k-means clusters (Fig. 2). Even given the limited number of locations from which the soil samples derive, the diverse nature of soils is potentially influencing our results, particularly in transitory environments, such as the transition from soil to lacustrine sediments in a lake shore. It is possible that these transitory locations require a more in-depth analysis, with the use of more extensive datasets.

4.3 GDGT Distribution

The GDGT profiles of the curated clusters show characteristic patterns that reflect known qualities of GDGTs in their respective environments. For example, as expected, the *Marine-type* samples have a much higher proportion of isoGDGTs, while the terrestrial clusters have a higher proportion of brGDGTs (Fig. 3). As previously described by Martínez-Sosa et al. (2021), *Lake-type* samples have a preference for 5-methyl isomers, although some work has suggested that 6-methyl brGDGTs can dominate in lacustrine environments with lower oxygen conditions (van Bree et al., 2020). Both *Peat-type* samples and soil samples from the *Soil-type* cluster also have a higher proportion of 5-methyl isomers, but river samples within the the *Soil-type* cluster show a clear preference for 6-methyl brGDGTs (Fig. 3b,c and Fig. 9). In addition, *Lake-type* samples have a higher proportion of IIIa, and a lower proportion of Ia, compared with the other terrestrial environments (Fig. 3b,d). Overall, the particular GDGT profiles from these depositional environments suggest that each may have a unique microbial community that responds to the environment in distinct ways (Raberg et al., 2022; De Jonge et al., 2019; Tierney & Russell, 2009).

Each cluster also has a characteristic pattern of GDGT influence, which affects their relationship with environmental parameters (Fig. 4). Notably, for *Marine-type* samples the first dimension is dominated by a negative relation with brGDGTs and a positive one with isoGDGTs (Fig. 4d) and it is not associated with temperature (Fig. 4a), unlike the other groups. While we speculate that this dimension is related to terrestrial influence, we did not find a relationship with the distance from the core sites to land or water depth, suggesting that it possibly represents a complex response to several environmental influences. The second dimension, which inversely follows GDGT-0, more closely follows the mixed layer temperature (Fig. 4a). 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 NMDS analysis shows no relation between GDGT-0 and brGDGTs, which suggests that GDGT-0 is not influenced by terrestrial sources (Fig. 3 b-d). Our results suggest that temperature strongly influences the abundance of this lipid 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}.

The first dimension of the *Lake-type* cluster follows MAAT (Fig. 4b) and the GDGT distribution along this dimension reflects the pattern associated with the MBT'_{5Me} index, with a positive relationship for Ia, Ib, and Ic, and a negative relationship with the remaining brGDGTs. In this first dimension, isoGDGTs do not seem to exert much influence. The second dimension seems to capture relative amounts of isoGDGTs vs. brGDGTs,

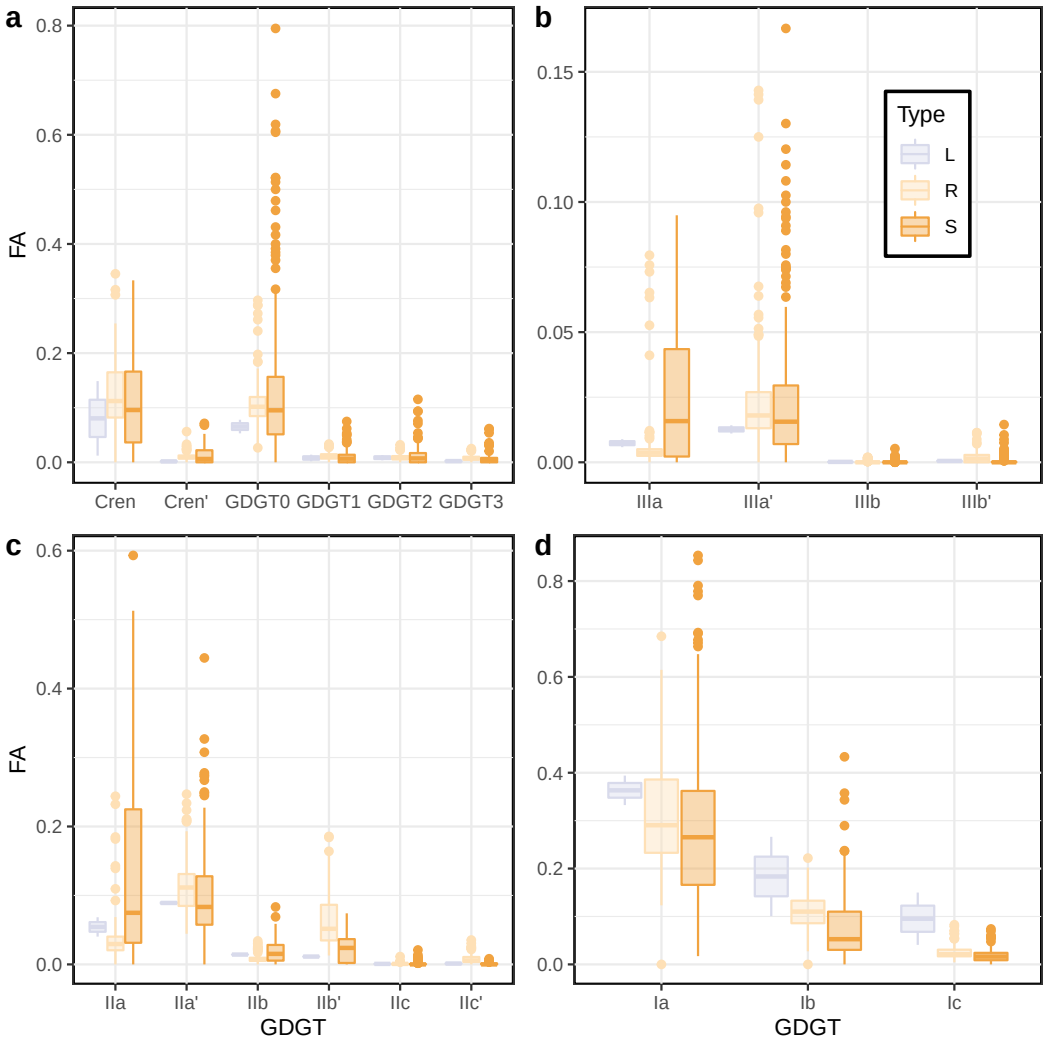

Figure 9. Box plots showing the distribution of the fractional abundance (FA) of all GDGTs in samples from the *Soil-type* cluster, following the color code of Figure 1. GDGTs separated by isoGDGTs (a), hexamethylated brGDGTs (b), pentamethylated brGDGTs (c), and tetramethylated brGDGTs (d).

but again, we were unable to find an environmental parameter that shows a relationship with this dimension; for example, lake depth is not associated with this axis of variability ($\rho = 0.13$). We speculate that this dimension reflects changes in microbial communities. These changes 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).

The *Peat-type* samples show a pattern similar to the lake cluster, with the first dimension following temperature, as shown by temperature increasing along the first axis on the NMDS analysis (Fig. 4c). The GDGT distribution in turn, follows to some extent the pattern of the MBT'_{5Me} index, with Ia, Ib and Ic plotting opposite to the rest of the brGDGTs. However, a unique feature of this cluster is that Ib and Ic appear to be less important, and less abundant than Ia. This is in line with previous work that has noted that there are relatively fewer brGDGTs with cyclopentane rings in peatlands, likely

because they are acidic (Naafs et al., 2017; Weijers, Schouten, et al., 2007). The GDGT distribution for the second dimension somewhat resembles the pattern for the CBT' index, with Ia and IIa negatively relating to this dimension. However, we found no relationship between this dimension and pH. Previous work has suggested that the abundance of isoGDGTs, particularly 1 – 4, could be related to factors such as water content or redox state (Yang et al., 2019); we observe that these GDGTs indeed have a positive relationship with the second dimension, suggesting that this could be the environmental driver.

4.4 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), suggesting that the observed performance is not due to overfitting the training set.

When we apply the BIGMaC algorithm to the complete dataset, we can investigate the importance of each GDGT in the model. The importance metric is calculated based on how much each GDGT contributes to decreasing the probability of incorrectly classifying a sample (Gini impurity) (Greenwell et al., 2020). This analysis shows that the two compounds that contribute the most to the classification are IIa' and crenarchaeol. While these compounds have not been substantially linked to any particular environmental response in previous work, PCA (Fig. 2d) suggests that they are strongly associated with *Soil-type* and *Lake-type* (IIa'), as well as *Marine-type* (crenarchaeol) samples. It is possible that the importance of IIa' is due to its association with *Lake-type* and *Soil-type* samples but not *Peat-type* samples, thus helping the classification algorithm split the terrestrial environments. Similarly, the association between crenarchaeol and *Marine-type* helps distinguish this group from the terrestrial environments.

4.5 Applications

Our GDGT analysis of the Giraffe core shows a good agreement with its previously described stratigraphy (Wolfe et al., 2017; Hamblin et al., 2003), with the sections of the core described as peat and lake, respectively, being correctly identified as such by BIGMaC (Fig. 6b). However, BIGMaC also reveals additional information about changes in the depositional environment in the lacustrine facies that was not evident in the stratigraphic description, which interpreted the environment to be a shallow lacustrine setting with intermittent wet and dry periods (Hamblin et al., 2003). Between 76.5 and 85 meters, within the lacustrine section, BIGMaC indicates a transition to a peatland environment, followed by a brief transitional period between *Soil-type* and *Lake-type* (Fig. 6b). This predicted feature is corroborated by the CBT' index, which also suggests a period of acidification in the lake section that matches the *Peat-type* section (Fig. 6b). Previous work reported the presence of acidophilic freshwater diatoms in this section of the core, consistent with our interpretation of an acidic depositional environment (Siver et al., 2010). While we cannot completely discard the possibility that the lake became acidic (rather than transitioning to a peatland), lakes show a muted response of CBT' to pH between a range of 4.3 to 10 (Martínez-Sosa et al., 2021). Given this, the observed change in CBT' in this section ($\sim$1 unit) would require the pH of the lake to be below 4.3, i.e., well beyond the range of the global calibration. Conversely, if we assume the CBT' values were recorded in a peat environment, they are consistent with a pH between 4 and 5, which is more in line with the conditions expected based on the observed diatoms (Siver et al., 2010). It is important to note that the species of diatom in this section, *Actinella*

giraffensis, does not match any extant species, although its closest relative *A. parva* is only known to inhabit freshwater bodies.

Our temperature reconstruction for the Giraffe pipe with the environmental correction for the different sections of the core suggests a relatively stable climate with no clear trend (Fig. 6a). The mean temperature of our reconstruction (19°C) agrees with independent studies. A pollen reconstruction on this site (red diamonds in Fig. 6a), 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. 6a). 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.

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. Prior to the PETM, the algorithm consistently suggests that the site is a peatland environment (Fig. 7). In contrast, during the PETM the algorithm struggles to assign a consistent depositional environment to the blocky lignite unit. Moreover, the PETM samples are primarily classified as *Peat-type* and *Soil-type*, suggesting that the same temperature calibration should be used as during the pre-PETM, thus undercutting any potential correction to the temperature reconstruction from Inglis et al. (2019). 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). From this perspective, the oscillating results from BIGMaC likely point to an unstable, dynamically changing depositional environment with mixed sources of brGDGTs. Since BIGMaC is 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.

5 Conclusions

Our analyses of 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 given environment. There is a strong association between temperature and the *Lake-type* and *Peat-type* groups, with a possible smaller effect of pH or conductivity on the latter group. *Marine-type* samples are also clearly influenced by temperature, but also seem to be affected by another environmental factor that drives changes in the relative proportion of isoGDGTs and brGDGTs, an observation that deserves further study. 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 have the strongest influence on separating the different groups identified here, possibly due to their association with *Marine-type* samples. 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 observed 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 Section

The GDGT fractional abundance data used for training the BIGMaC algorithm in the study are directly available at Pangea via https://doi.org/10.1594/PANGAEA.883765, https://doi.org/10.1594/PANGAEA.938067, https://doi.org/10.1594/PANGAEA.907818, https://doi.org/10.1594/PANGAEA.918523, and https://doi.org/10.1594/PANGAEA.901285; as well as on Zenodo via https://doi.org/10.5281/zenodo.7540094, https://doi.org/10.5281/zenodo.7522415 and https://doi.org/10.5281/zenodo.3939270. V1.0 of the BIGMaC algorithm used for the classification of samples based on GDGT fractional abundances is preserved at https://doi.org/10.5281/zenodo 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. 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 acknowledge 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, 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., & 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., 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.

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., 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* [data set]. 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.

Kassambara, A., & Mundt, F. (2020). Extrac and Visualize the Results of Multivariate Data Analyses. R Package Version 1.0. 3. *R package version*.

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. (2020a). *dials: Tools for Creating Tuning Parameter Values.* R package version 0.0.

Kuhn, M. (2020b). Tune: Tidy Tuning Tools. *R package version 0.0*, *1*.

Kuhn, M., Chow, F., Wickham, H., et al. (2019). Rsample: General resampling infrastructure. *R package version 0.0*, *5*.

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.

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, January). *BIGMaC GDGT algorithm.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.7513557` doi: 10.5281/zenodo.7513557

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. D. A. (2017). *Global biomarker (GDGT) database for peatlands* [data set]. 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.

Pancost, R. D., Taylor, K. W., Inglis, G. N., Kennedy, E. M., Handley, L., Hollis, C. J., ... others (2013). Early Paleogene evolution of terrestrial climate in the SW Pacific, Southern New Zealand. *Geochemistry, Geophysics, Geosystems*, *14*(12), 5413–5429.

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.

Peterse, F., van der Meer, J., Schouten, S., Weijers, J. W., Fierer, N., Jackson, R. B., ... Sinninghe Damsté, J. S. (2012). Revised calibration of the MBT–

CBT paleotemperature proxy based on branched tetraether membrane lipids in surface soils. *Geochimica et Cosmochimica Acta*, *96*, 215–229.

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.

Russell, J. M., Hopmans, E. C., Loomis, S. E., Liang, J., & Damsté, J. S. S. (2018). Distributions of 5-and 6-methyl branched glycerol dialkyl glycerol tetraethers (brGDGTs) in East African lake sediment: Effects of temperature, pH, and new lacustrine paleotemperature calibrations. *Organic Geochemistry*, *117*, 56–69.

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.

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. *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., Wolfe, A. P., & Edlund, M. B. (2010). Taxonomic descriptions and evolutionary implications of Middle Eocene pennate diatoms representing the extant genera Oxyneis, Actinella and Nupela (Bacillariophyceae). *Plant Ecology and Evolution*, *143*(3), 340–351.

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. (2009). Distributions of branched GDGTs in a tropical lake system: implications for lacustrine application of the MBT/CBT paleoproxy. *Organic Geochemistry*, *40*(9), 1032–1036.

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 char-

acteristics of magmatic tectonic settings: A machine-learning-based approach. *Geochemistry, Geophysics, Geosystems*, *19*(4), 1327–1347.

van Bree, L. G., Peterse, F., Baxter, A. J., De Crop, W., Van Grinsven, S., Villanueva, L., ... Sinninghe Damsté, J. S. (2020). Seasonal variability and sources of in situ brGDGT production in a permanently stratified African crater lake. *Biogeosciences*, *17*(21), 5443–5463.

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., Schefuß, E., Schouten, S., & Sinninghe Damsté, J. S. (2007). Coupled thermal and hydrological evolution of tropical Africa over the last deglaciation. *Science*, *315*(5819), 1701–1704.

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., & Sinninghe Damsté, J. S. (2007). Environmental controls on bacterial tetraether membrane lipid distribution in soils. *Geochim. Cosmochim. 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. (2019). A fast implementation of random forests. *R package version 0.11*, *2*, 123–136.

Yang, H., Xiao, W., Słowakiewicz, M., Ding, W., Ayari, A., Dang, X., & Pei, H. (2019). Depth-dependent variation of archaeal ether lipids along soil and peat profiles from southern China: Implications for the use of isoprenoidal GDGTs as environmental tracers. *Organic Geochemistry*, *128*, 42–56.

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.

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.