Review of Cromartie *et al.* for *Biogeosciences*

Joseph B. Novak

Recommendation

Major revision.

Summary

Cromartie et al. present a new machine learning approach to probabilistically assess the provenance of brGDGTs in terrestrial sedimentary archives. This work improves upon previous work by Martinez-Sosa et al. (2023), the BIGMAC algorithm, by generating probability estimates that permit analysis of the likely relative contributions of brGDGTs from various sources to a sediment sample rather than discrete sample classifications. The improvement upon the BIGMAC algorithm is a contribution towards ongoing efforts to utilize brGDGTs as proxies of past climate change in the geologic record.

The writing is mostly clear, although there are some places where I was confused by the word choice or sentence structure. Wherever possible, I provided suggestions to revise the wording for clarity. I urge the editors to find a machine learning expert to evaluate the methodology of this work, as this technique does not fall within my expertise. My recommendation for a major revision is based upon my concerns regarding section 4.2.4 where a new brGDGT wetlands index is proposed (see major comments).

I look forward to the publication of this work after my comments are addressed.

Major Comments

Introduction

The introduction would benefit from some clarification as to why it was necessary to use five machine learning techniques to generate the model described here. Did you try five machine learning methods and then settle on one as the best? Are you somehow combining the output of all five models? Machine learning is generally a confusing (and intimidating!) methodology for many people, including some who would want to use your algorithm. Clarity on why you took this approach will make people more likely to understand what you did and therefore more likely to use your algorithm (and cite your work! 😊). I think an additional 1–3 sentences in the paragraph at lines 81–92 would be very helpful for clarifying this point.

Materials and Methods

L220–221: Do you mean that you are using the probability estimates as a means of understanding changes in brGDGT provenance through time? Because that is a different thing than using them as an environmental proxy. Please clarify.

Discussion

Figure 6a and 7a: why is the pollen water depth reconstruction plotted on a log scale? This seems a bit odd, should this not be plotted on a linear scale? Please explain.

Section 4.2.4: I question whether including this section distracts from the larger point of the paper.

Is this index not redundant since you are tracking basically the same thing with the % peat probability? Perhaps more importantly, there should be some sort of validation regarding whether this index is useful for identifying wetlands in a modern dataset. For example, is this index value higher in modern samples from wetlands than in those from dry soils or lakes? I think this section generally should be expanded upon significantly if the authors want to claim that this index can be used this way.

Minor Comments

Abstract

L21: I think you mean "Branched glycerol dialkyl glycerol tetraethers (brGDGTs) **are** critical molecular biomarkers" rather than "…serve as critical molecular biomarkers."

L22–23: Is the sentence starting with "Despite their success…" necessary?

L25: here and throughout, make sure to use the proper prime symbol ′

L25–26: "…where ecosystems are sensitive to diverse environmental factors." Do you mean that depositional environments in arid and semi-arid regions are prone to change in response to water stress?

L31: "…obtained from the identical records." Do you mean from the same samples? Or from the same cores?

L34: Typo. "brGDGT provenance" not "brGDGTs provenance."

L36: I think a word is missing here. Do you mean potential biases in brGDGT paleotemperature reconstructions?

Introduction

L39–43: These two sentences are largely redundant. Could they be combined?

L47: "**their** potential" not "its potential." I would consider removing this initial clause in the sentence and starting with "A key challenge…" as this is a more focused start to the paragraph.

L47–48: I think you need a citation here since this thought is informed by previous work.

L56–58: I think you mean "The MBT′$_{5Me}$ index is correlated to temperature in lake sediments, peats, and soils [CITATIONS]."

L59–65: Somewhere in here it would be good to mention that MBT′$_{5Me}$ is systematically higher in soils than in lakes. This is usually the major source of concern when it comes to dealing with brGDGTs from potentially mixed sources, at least in lake sediments.

L78: what do you mean by ecological changes? As in bacterial ecology? This may be a word choice issue, I was really surprised to see the word "ecology" here.

L99 vs L100: do you mean depositional or provenance? Because those are different things. The words cannot be used interchangeably.

Materials and Methods

L123: I think you mean limited data, not limited publication.

L135: “A C_{46} **internal** standard.”

L139: “…the C_{46} **internal** standard.”

L176: Check the journal’s referencing policies. I am not sure that “ibid” is permitted within this citation style.

L189–191: This sentence is missing a verb. Please clarify.

L192: same comment regarding ibid.

L196: same comment regarding ibid.

Figure 2: consider making the background color for the soil dataset symbol a lighter shade of brown. I found it hard to read the dark font against the dark brown background.

L222: do you mean “periods of mixed brGDGT provenance?”

L244–245: why did you retain the original sample to sample curve?

L252–254: I am not sure this is necessary to explain.

Results

L279–280: these sentences should be combined.

Discussion

L362: I think you mean that the two timeseries are qualitatively similar.

L402: italicize *in situ*

L429–431: I am not sure what you mean. The De Jonge et al. 2024 study found that MBT'_{5Me} are generally reproducible between laboratories.

L450: the ′ symbol should not be subscripted

L468: you can simply report this p-value as “$p < 0.001$”

L475–482: why are some of the brGDGTs written in square brackets sometimes but not other times?