**Model-driven identification of OM molecular signatures controlling biogeochemical transformation in river corridors**

**SUMMARY PARAGRAPH**

Organic matter (OM) assemblages in surface water and sediment are determined by complex biogeochemical processes, including those governed by microbes, and thus highly variable in spatiotemporal spaces. This continental-scale, ultrahigh-resolution river metabolomic dataset provides an opportunity to investigate global trends of riverine OM characteristics (e.g., source, mobility, bioavailability, reactivity, parent vs. product compounds) through both data-driven and physics-based modeling. In particular, models applying thermodynamic theory will allow better synthesis and identification of key OM signatures (e.g., thermodynamic favorability, reaction energy efficiency) that control their biogeochemical transformation (e.g., oxidative respiration rate), and can help improve the representation of biogeochemical mechanisms in more integrated models such as reactive transport models. Further, when coupled with existing meta-data of climate, hydrology, geology, and ecology, OM thermodynamics will allow better predictions of biogeochemical and ecosystem dynamics under changing climate. Finally, through data-driven approaches like Sparse Identification of Nonlinear Dynamics (SINDy), we will identify key compound sets that control OM thermodynamics and aerobic respiration rates. The results could be analyzed along with microbiome data to better understand microbe-metabolite interactions provided that microbiome data will be collected from the same locations in future.

**INTRODUCTION**

* State the challenge of extracting meaningful inferences from high-throughput OM data (number of observations, OMs >> number of sample points). This challenge, commonly seen in any high-throughput data, requires better synthesis of data and advanced analytical approaches.
* Explain that OM chemistry can translate to variable thermodynamic properties and be linked to microbial respiration kinetics.
  + Characterizing the OM pools via their thermodynamic properties can facilitate mechanistic understanding of biogeochemical transformations.
  + Introduce thermodynamic “lambda” model that can be used to infer thermodynamic properties of OM.
* Introduce the derivation of OM molecular signatures from thermodynamic properties.
  + OMs pools grouped based on their molecular signatures can offer clear and direct connections with biogeochemical transformations.
  + Briefly introduce the concept of data-driven approach, i.e., SINDy, which is ideal to identify OM signatures from highly distributed data.
  + Introduce the availability of ancillary meta-data that can be used to make insightful comparisons of biogeochemical transformation across varying conditions against the identified OM molecular signatures.
* This study: Briefly go through objectives below. Summarize our findings.
* Make a final statement that the approach of identifying OM molecular signature can be widely used to infer implications of changing environmental conditions.

**OBJECTIVES**

1. Synthesize the distribution of OM thermodynamic properties across samples that represent different climatic and ecological conditions.
2. Identification of OM molecular signatures that are derived from condition-specific distribution of thermodynamic properties using SINDy.
3. Compare variation and similarity of OM molecular signatures across spatial (surface vs. sediment, along stream orders), climatic (intermittent vs. perennial streams) and ecological (vegetation vs. bare) domains to understand the relationship between OM molecular signatures and environmental conditions.
4. Identification of congruence between core-satellite OM species grouping, model-predicted OM thermodynamic properties, and inferred OM molecular signatures.

**METHODS**

**Section 1: Thermodynamic “lambda” model**

**Section 2: Data-driven SINDy approach**

**Section 3: Multivariate and correlation analyses**

* Elaborate 2019 WHONDRS data (elemental composition, NPOC, respiration rate).
* Ancillary meta-data
  + Original meta-data: <https://drive.google.com/file/d/1qox0YK82gVVzzShrrqHqN4l7rpd3xlco/view?usp=sharing>
  + Additional meta-data summarized here: <https://docs.google.com/document/d/1mmV9LN23HZIETTHc7uTEHCQUzFx1Nfnvb5KlmEFqw6w/edit>
  + A new meta-data file including intermittency here: <https://drive.google.com/file/d/17arlZPHjzyGbyypYbPeTE3AStWRZzimR/view?usp=sharing>

**RESULTS**

**Section 1: Reduced-order prediction of ecosystem respiration using identified OM molecular signatures.**

Questions / Hypotheses

* Do OM molecular signatures reasonably predict ecosystem respirations?

Paragraphs

Analysis ideas:

* Use model-predicted OM thermodynamic properties to identify OM molecular signatures.
* Validate the results - compare predicted respiration rates with lab-measured respiration rates.

Figure ideas

* Distribution of model-predicted OM thermodynamic properties (lambda, delG) from individual sites.

Chart

Description automatically generated

* OM molecular signatures identified from all distributions of OM thermodynamic properties across sites.

Chart

Description automatically generated with medium confidence

* Scatter plots of correlation between predicted rates and experimental data.

Chart, scatter chart

Description automatically generated

**Section 2: OM molecular signatures translate to condition-specific biogeochemical transformations.**

Questions / Hypotheses

* Are patterns in OM molecular signatures reflect environmental variations (i.e., spatial variation -  surface vs. sediment, stream order; climatic variation - intermittent vs. perennial streams; ecological variation - vegetation vs. bare)?

Paragraphs

Analysis ideas:

* Multivariate analysis
* Correlation analysis

Figure ideas

* Highlight identified OM molecular signatures in the Van Krevelen diagram.

Scatter chart, qr code

Description automatically generated

* Highlight categorizable patterns of OM molecular signatures under specific conditions in a bipartite network of the OM pool subsets with relevant network centrality measures.
* Use boxplots to assess the difference in distribution of various thermodynamic properties of OM molecular signatures across datasets, with significance test and correlation analysis.

Chart

Description automatically generated

* Unconstrained ordination plots of FTICR data vs PCA plots of OM molecular signatures grouped based on meta-data subsets.

Chart, scatter chart

Description automatically generated

**Section 3: Congruence between core-satellite OM species and thermodynamic properties and OM molecular signatures.**

Questions / Hypotheses

* Can core and satellite species be inferred based on OM thermodynamic properties and molecular signatures?
* Do core-satellite OM molecular signatures vary upon environmental conditions?

Paragraphs

Analysis ideas:

* Compare thermodynamic properties of occupancy-based core vs. satellite species.
* Analyse the distribution of OM molecular signatures of core and satellite groups upon environmental conditions.

Figure ideas

* Compare distribution of OM thermodynamic properties between core and satellite species, with significance test and correlation analysis.

Chart, box and whisker chart

Description automatically generated

* Compare core vs. satellite OM molecular signatures under varying environmental conditions.

A picture containing application

Description automatically generated

* Compare correlation between occupancy % and thermodynamic properties for all OM species vs. OM molecular signatures. Color clusters based on environmental conditions.

Chart

Description automatically generated with low confidence

**DISCUSSION**

* Concluding remarks
  + Key findings from this paper
  + OM molecular signature can be applied to infer implications for changing environmental conditions.
* Comparing core-satellite OM species grouping based on other approaches (e.g., molecular weight, aromaticity in Topic 1) vs. molecular signatures identified here.
* Future works