**Manuscript outline**

*Topic 6) Model-driven identification of OM molecular signatures controlling biogeochemical transformation in river corridors*

*#---------------------------------------------Document explanation---------------------------------------------#*

Please make edits in **“Suggestion Mode”**. This draft (v1.0) reflects the current summary paragraph in the [editorial document here](https://docs.google.com/document/d/1k0G3YBb3P1gGnnzCOOUqthC0N7X-BXrxzts2HRTQRe4/edit?usp=drive_web&ouid=117093633409831549208), which is based on the 7/19 and 7/28 meeting discussions [here](https://docs.google.com/document/d/1AejjFVq9BkmrvORGzlmVDl0bqpO6G6MWs4PDUkfdtrA/edit). Refer to the same documents for other notes and comments that are being  added through the development until 10/30/2021.

#-----------------------------------------------------------------------------------------------------------------------#

**Summary paragraph (please revise this paragraph** [**here**](https://docs.google.com/document/d/1k0G3YBb3P1gGnnzCOOUqthC0N7X-BXrxzts2HRTQRe4/edit?usp=drive_web&ouid=117093633409831549208)**; upon the revision, this paragraph will be updated)**

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

Model-driven identification of OM molecular signatures controlling biogeochemical transformation in river corridors. 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.
* 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: Experimental datasets**

* Elaborate 2019 WHONDRS data (elemental composition, NPOC, respiration rate) and other ancillary meta-data.

**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 - use the identified OM molecular signatures to predict respiration rates and compare correlation with lab-incubated respiration rate data.
* Establish that OM molecular signatures exist which can be used to infer ecosystem respirations.

Figure ideas

* Distribution of model-predicted thermodynamic properties.

Chart

Description automatically generated

* Scatter plots of predicted rates and experimental data.

Chart, scatter chart

Description automatically generated

* Highlight identified OM molecular signatures in a distribution of thermodynamic properties (lambda, delG) of the entire OM pool.

Chart

Description automatically generated with medium confidence

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

Questions / Hypotheses

* Are there any differences/similarity in OM molecular signatures identified for specific conditions (i.e., spatial variation -  surface vs. sediment, stream order; climatic variation - intermittent vs. perennial streams; ecological variation - vegetation vs. bare)?

Paragraphs

Analysis ideas:

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.
* Use boxplots to assess the difference in distribution of various thermodynamic properties of OM molecular signatures across datasets. Use p-values and Pearson correlation to assess statistical significance.

Chart

Description automatically generated

* PCA 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: Distinction between core and satellite species could be inferred based on thermodynamic properties and OM molecular signatures.**

Questions / Hypotheses

* Is the thermodynamic properties and OM molecular signatures consistent between core and satellite species?

Paragraphs

Analysis ideas:

* Compare thermodynamic properties of core and satellite species.
* Analyse the distribution of OM molecular signatures of specific environmental conditions across core and satellite groups.

Figure ideas

* Compare difference in distribution of various thermodynamic properties of OM molecular signatures between core and satellite species. Use p-values and Pearson correlation to assess statistical significance.

Chart, box and whisker chart

Description automatically generated A picture containing application

Description automatically generated

* Compare correlation between species occupancy % and thermodynamic properties for all species and for OM molecular signatures.

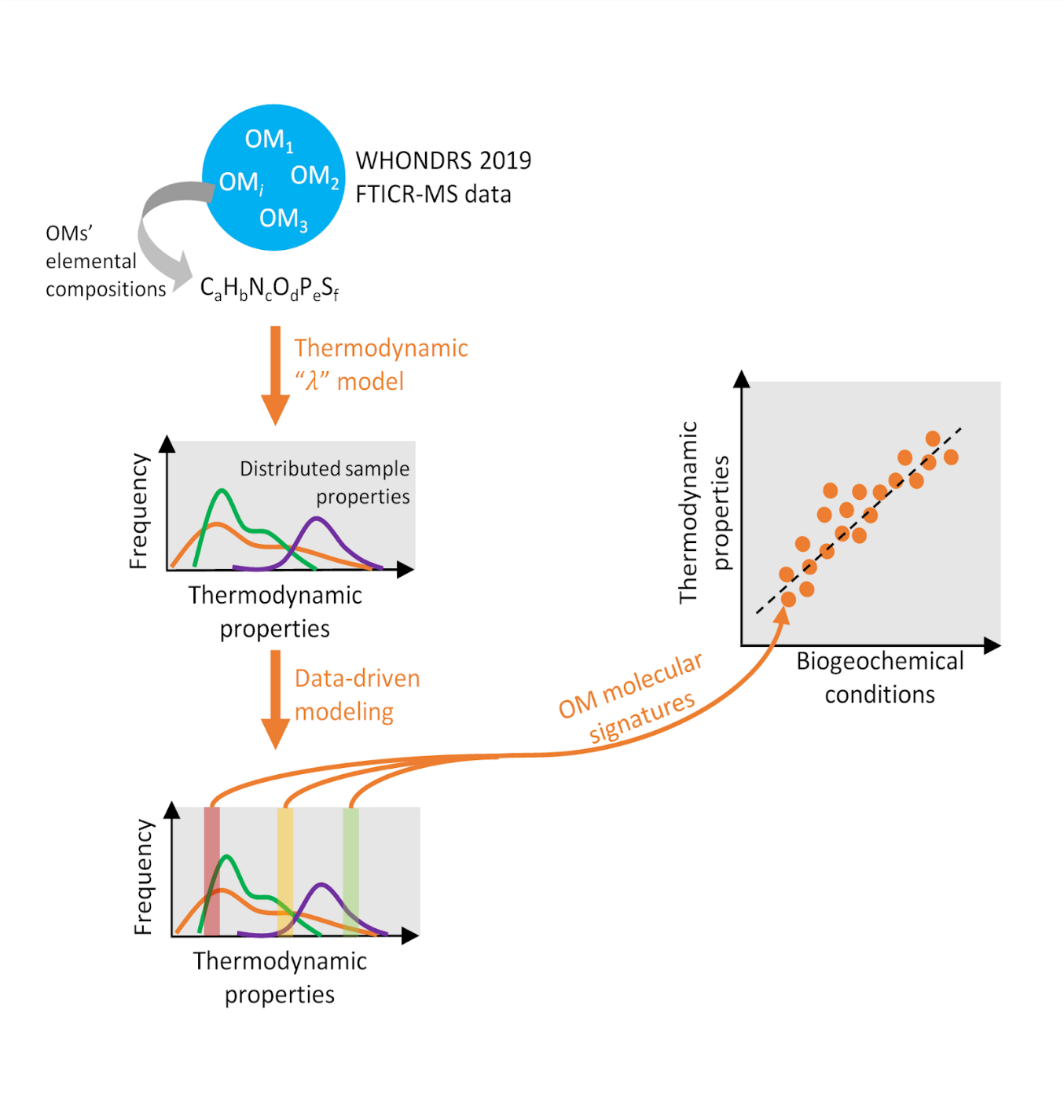
Chart

Description automatically generated with low confidence

**DISCUSSION**

* Concluding remarks
  + OM molecular signature can be widely used 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

**HYPOTHESIS FIGURE**

****

**Figure legend:**

An illustration of our model- and data-driven approach to extract obscure inferences from high-throughput OM data. OM metabolomics data typically contain molecular-level signatures of condition-specific biogeochemistry that are identifiable through advanced data-driven modeling techniques.