**Topic 4 Outline**

**Deadline: October 4, 2021**

Title: Examining the ecological processes influencing the assembly of molecules into OM assemblages

Summary Paragraph: Dissolved organic matter (DOM) assemblages in freshwaters are formed from complex mixtures of compounds that are highly variable across time and space due to the environmental heterogeneity of river networks. We propose that ecologically significant DOM assemblages can be extracted from a continental-scale dataset of fluvial FTICR-MS organic matter molecular characterization, nutrient levels, stable water isotopes (2H and 18O), and other physicochemical parameters. To identify assemblages, we will apply artificial neural network and machine learning on DOM samples collected across continental US. Once DOM assemblages are identified we will use variation across space and time to investigate metabolomic processes in metacommunity ecology and community metabolomes. We will address questions about the origins of DOM  assemblages in streams and rivers. Specifically, we will examine the relative importance of biotic versus abiotic processes, watershed characteristics, coupled nutrient cycles, and sediment metabolism on DOM assemblage formation. We argue that this data driven approach will reveal common continental-scale DOM assemblages and the metabolomic processes that generate these groups of compounds.

**INTRODUCTION**

* The importance of OM in aquatic ecosystems
  + OM is both a resource for heterotrophic biota and a consequence of ecological processes that correspondingly, constrains and reflects metabolisms and OM sources in a watershed
  + In lotic ecosystems, the dissolved fraction of OM has the potential to be highly mobile and can be transported downstream or captured and metabolized by microbes. (Abiotic factors such as UV degradation and absorbance (?) can also control the fate of DOM.)
  + While all DOM is likely to be ecologically significant, we here focus on identifying and interpreting DOM assemblages that can represent a response to ecological processes and/or are likely to affect future ecological processes.
    - This is similar to “response and effect” traits in Martiny et al. 2015
  + We are learning that the distribution of different DOM molecules is not universal across stream ecosystems despite that it is a consequence of degradation of fundamental biomolecules.
* Challenges to understanding the relevance of DOM species/formations
  + Literature overview use of artificial neural network in biogeochemistry and microbial ecology: papers [Danczak (2020)](https://www.nature.com/articles/s41467-020-19989-y) and [Larsen (2012)](https://www.nature.com/articles/nmeth.1975).
* What is currently known about how DOM/DOC varies spatially (i.e., patterns) and what attributes cause this variation (controls).
  + Regional/global variation in OC processing rates (see Tiegs et al. 2019)
  + Hyporheic vs surface water (PNNL pubs)
  + Microbial (biotic) vs Abiotic
    - Nutrients, watershed…
    - What controls DOM persistence and reactivity in freshwaters? Molecular Composition and Time → Check Kothawala et al 2021 (Trends in Ecology & Evolution)
* Another paragraph…
* Specific research questions to be addressed using the WHONDRS data:
  + 1) How do DOM class formations and inherent molecular attributes vary for sediment and surface water compartments? Do these assemblages correspond to a specific class of molecules? Like a cluster of lignin-derived DOM, proteic, lipidic, etc?
  + 2) What is the relative influence of biotic versus abiotic features (i.e., watershed characteristics, elemental macronutrient ratios, oxidation state, sediment metabolism) on DOM assemblage formation in sediment and surface water compartments and how do these relationships vary across sites?
  + 3) How does the source (autochthonous, allochthonous) and pathway (primary vs secondary production) of DOM help identify its molecular characteristics and can these relationships be predicted by watershed conditions (landscape and stream order)?

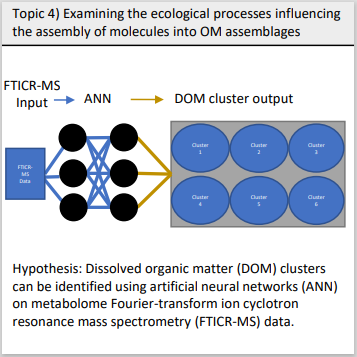


Figure 1. We hypothesize that DOM clusters can be identified using machine learning techniques on FTICR-MR data. MORE

**Methods**

1. **Data source:**
   1. Original dataset from WHONDRS (Processed\_S19S\_Sediments\_Water\_2-2)

1. **Data processing**
   1. JianJun summarised whole set in R
      1. **Processed set DOM\_Comp**: **DOM composition data** from sediments and water samples (91085 observations and 558 samples in total); DOM peaks and composition were generated by following Hu et al [3]. We included all assigned and unassigned peaks in this data.
      2. **MetaData**: **Metadata** for 558 samples from sediments and water (558 observations and 49 variables in total).
   2. Topic 1 data set made available by Michaela

1. **Analyses:**

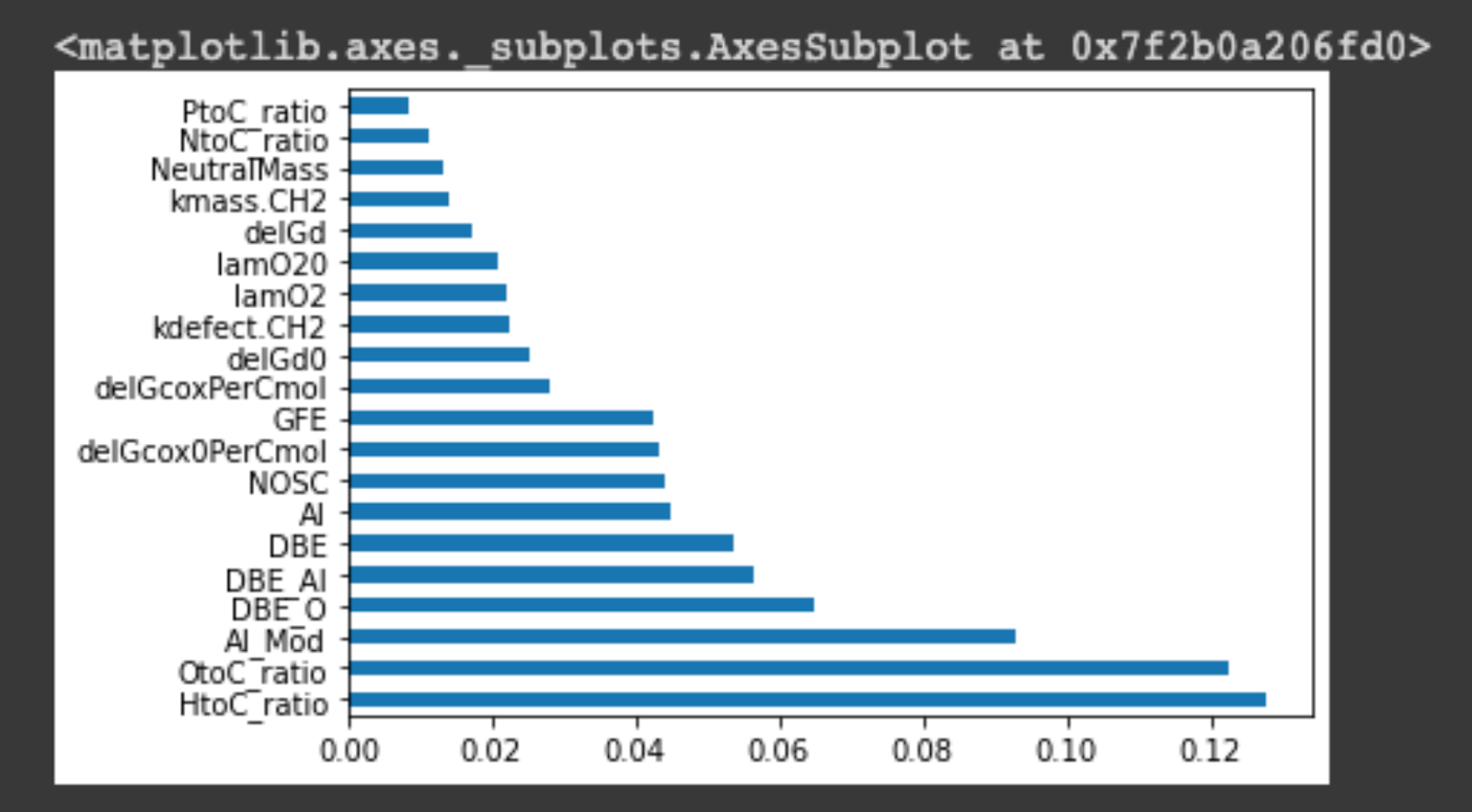
We plan to build and run 3 sets of analyses: sediment only, surface water only, and both combined; on their own to look at DOM clusters and with metadata (and pictures?!) to look at relationships to environmental parameters.

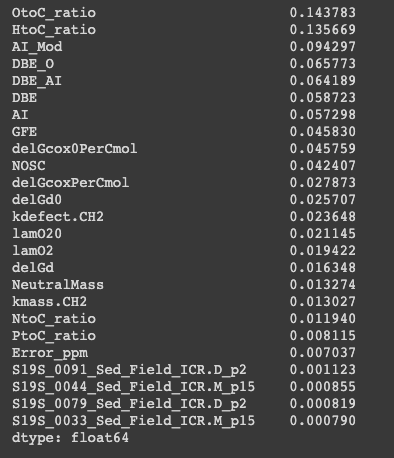
1. Random Forest models RF (and other ML models later on) for all 3 sets
   1. We plan to use Random Forest models to identify predictors of DOM class formation in sediment and surface water samples.
   2. We constructed three random forest models (currently for DOM data only and numerical data only), one each for sediment and surface water compartments, and one combined, to distinguish which predictors variables are “important” in explaining DOM class formation in sediment and surface waters.
      1. Add table of variables used in the random forest models which will describe data type (i.e., numerical vs. categorical)
      2. Describe how we rank predictors as important and any cutoffs for selecting the top predictors.
2. Calculate DOM diversity index to be included in analyses
   1. molecular richness DR, Gini-Simpson index DA, and the functional molecular diversity index DF (similar to Mentges et al. 2017).
3. Include Metadata in the analyses
4. Use pictures from the sites as input and see if CNN or similar can predict what kind of OM we have in the water
5. There is the potential to use isotope data for water retention calculations

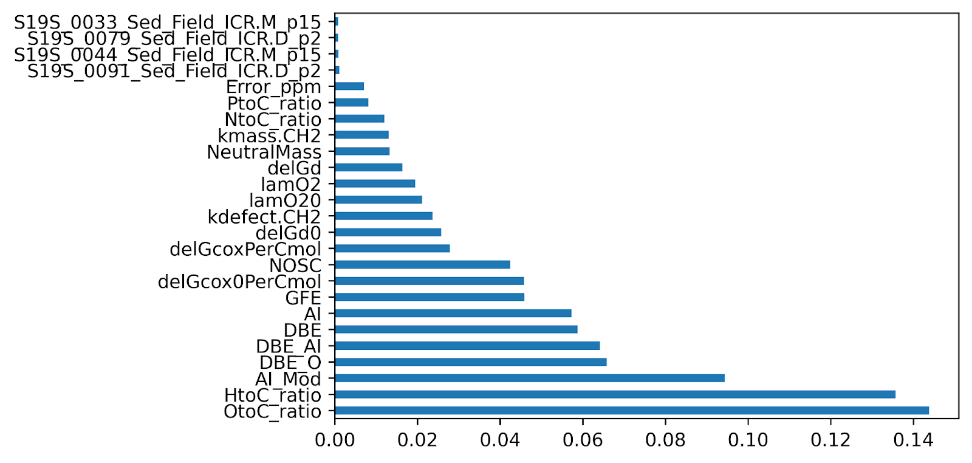
**RESULTS**

1. Identifying clusters in
   1. **waterXsediment**: class (lignin, proteic, etc); Nosc, DBE

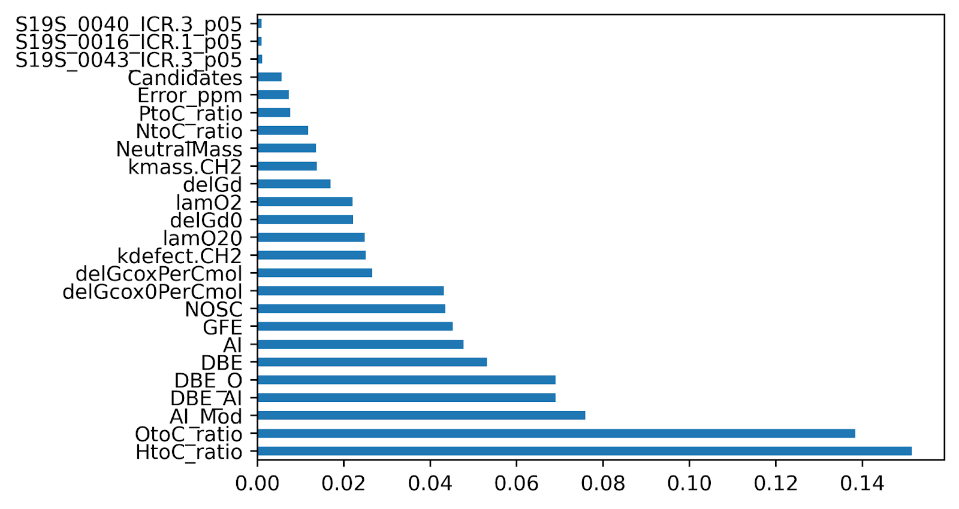
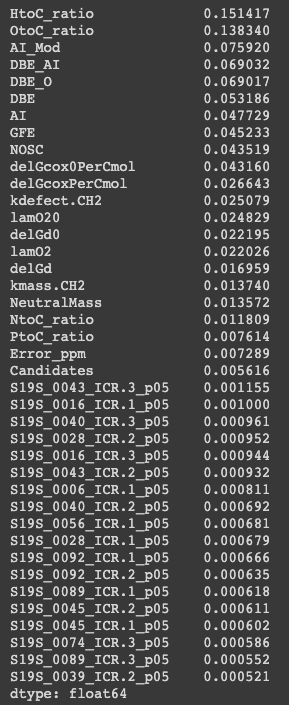
1. Elemental ratios and molecular characteristics (i.e., modified aromaticity index [AI\_mod], DBE, others??) were ranked as “important” predictors of DOM class formation.
2. Figure. Random Forest output for DOM class formation in whole data set (**water and sediment combined**)



1. **Sediment only**



1. **Surface Water only**

Formation of classes (lignin, tannin, lipid etc) is controlled by similar parameters in water, sediment, and overall (combined water+sediment).

Next immediate steps are to include diversity calculations and include metadata. Consider separating core and satellite DOM for future analyses.

Sensitivity analyses (based on Franceschini et al. 2019), PcOA, and ANN (to see if sampling images can be used to predict DOM types /patterns /diversity) are next in line.

**DISCUSSION**

* 1) DOM class formations and inherent molecular attributes vary for sediment and surface water compartments in these ways...
  + Assemblages correspond to a certain classes of molecules (consider lignin-derived DOM, proteic, lipidic, etc)
* 2) Discuss how DOM assemblages related to certain inherent features (watershed characteristics, elemental macronutrient ratios, oxidation state, sediment metabolism, etc.) and to what extent they were distinct for sediment and surface water compartments.
  + Discuss how and why these relationships may vary across sites
* 3) Evaluate the OM source (autochthonous, allochthonous) and pathway (primary vs secondary production) of DOM to help identify how well these relationships can be predicted by watershed conditions (landscape and stream order)

**REFERENCES**

Danczak et al. 2020

Garayburu-Caruso et al. 2020. https://www.mdpi.com/2218-1989/10/12/518

Franceschini S, Tancioni L, Lorenzoni M, Mattei F, Scardi M (2019) An ecologically constrained procedure for sensitivity analysis of Artificial Neural Networks and other empirical models. PLoS ONE 14(1): e0211445.

Fudyma et al. 2021. Coupled Biotic-Abiotic Processes Control Biogeochemical Cycling of Dissolved Organic Matter in the Columbia River Hyporheic Zone. https://www.osti.gov/servlets/purl/1777984

Hu et al

Larsen et al. 2012

Liang C, Schimel J, Jastrow J. The importance of anabolism in microbial control over soil carbon stor- age. Nature Microbiology, 2:17105, 07 2017.

Mentges et al. 2017. <https://www.frontiersin.org/articles/10.3389/fmars.2017.00194/full#h3>

Tiegs et al. 2019. **Global patterns and drivers of ecosystem functioning in rivers and riparian zones.** https://www.science.org/doi/10.1126/sciadv.aav0486