**Title:** Predicting Nutrient Ratios from Marine Microbial Metabolite Composition Using Machine Learning

**Project Leads:** Iris Kern and Joshua Sacks

**Project Helper:** Ayesha Mohammed

**Introduction:**

Marine microbial communities drive biogeochemical cycles in the ocean and are the base of marine ecosystems. Light driven primary production by phytoplankton controls the amount of fixed carbon entering a system which controls microbial biomass. Phytoplankton growth in the ocean is limited by different nutrients in different regions of the ocean, primarily nitrogen in the vast, oligotrophic oceanic gyres or iron in the polar and equatorial regions (termed high nutrient low chlorophyll (HNLC) regions). Differences in nutrient input rates and ratios determine the limiting nutrients in different areas of the ocean and different limiting nutrients lead to different microbial community compositions, microbial physiologies, organic matter compositions, and biogeochemical cycling rates. Global climate change is expected to drive increased ocean stratification and lower nutrient inputs into the surface ocean with potentially substantial impacts on marine microbial communities and carbon cycling.

Metabolites are small, biologically produced organic molecules that are the building blocks of macromolecules, the intermediates of metabolic processes, and serve important cellular functions such as resource storage, signaling molecules, and stress response. Metabolomics is the measurement of all small molecules in a system, often using liquid chromatography-mass spectrometry. The marine microbial metabolome (all metabolites in a system) is controlled by a combination of taxonomy and cellular physiology. Therefore, the overall meaning of a metabolite measurement is the combined signal of these two factors. Here, we propose to use machine learning to identify relationships between marine microbial metabolites and nutrient concentrations to learn how microbial communities shift in response to changes in nutrients.

**Project Overview:**

Research Question:

Can we predict nutrient concentrations of a marine environment from its microbial metabolome?

Question type:

Regression

Datasets:

This project will use a controlled mesocosm experiment (PERI-DICE) metabolite dataset where nutrients were amended into natural communities as a test dataset to develop the machine learning model to predict nutrient concentrations. We will then test the robustness and generalizability of this model on an environmental dataset with samples containing a wide range of nutrient concentrations (Gradients 1).

Train: PERI-DICE:

The PERI-DICE metabolite dataset is a set of samples collected from the PERI-DICE experiment. This experiment collected metabolites weekly on a 4 week long nutrient enrichment experiment, with treatments designed to test how nutrient ratio and input rate influenced the microbial communities present in the nitrogen limited North Pacific Subtropical Gyre. The seawater was collected off the coast of Hawaii and incubated with daily nutrient amendments for a 4 week period. This dataset was generated in the Ingalls Lab.

Test: Gradients 1:

The Gradients 1 metabolite dataset is a set of samples collected on the Gradients 1 research cruise (KOK1606) in April, 2016. This cruise transited latitudinally along 158 degrees W from the nitrogen limited North Pacific Subtropical Gyre to the iron limited North Pacific Subpolar Gyre. Metabolites were sampled from the surface ocean (15m depth) along the transect capturing environmental communities experiencing a range of nutrient conditions. This dataset was generated in the Ingalls Lab and is published in Heal et al. 2019.

Machine Learning Approaches:

Multiple approaches can be taken to answer the research question. The possible approaches are the following and are listed in order of most applicable:

1. Linear Regression - LOOCV
   1. Linear regression is used when the dependent variable/s continues and follows a linear relationship with the independent variable.
   2. Multiple linear regressions would be necessary as there are multiple predictors influencing the target variable.
   3. Weakness: This approach assumes a linear relationship exists between the dependent and independent variables. If no such relationship exists, the results may turn out distorted and inaccurate.
2. Random Forest
   1. Random Forest method involves using multiple decision trees and averaging their predictions. This approach addresses the weaknesses of decision trees and improves generalization.
   2. This approach should be used when dealing with more complex relationships.
   3. Weakness: Random Forests are much more complicated and are harder to interpret compared to a single decision tree.
3. Decision Trees
   1. This approach splits up the data into subsets based on the most significant attribute of each node. This approach is best used when simple relationships exist between variables that are easily interpretable and when you want to understand individual decisions.
   2. Decision trees are best used when a nonlinear relationship exists between the variables
   3. Weakness: decision trees are very sensitive to small variations in data and will fail when dealing with complex relationships
4. Neural Networks
   1. Neural networks can capture complex relationships in data. This model consists of multiple layers of interconnected nodes. It is capable of learning hierarchical representations from input data. The output layer typically has one neuron, and the model learns to predict one continuous value
   2. This model can model non-linear and highly complex relationships in data. The model can capture intricate patterns and features.
   3. Weakness: The model requires huge amounts of data, is computationally intensive, and lacks simple interpretability.
5. Support Vector Machines (SVM)
   1. This model is used for regression and classification tasks. SVM works by finding the hyperplane that best predicts the target variable values.
   2. SVM is suitable for datasets with many features. The model is versatile due to the use of kernel functions, which allows it to capture non-linear relationships
   3. Weakness: Sensitivity to choice of kernel parameters and can be computationally expensive

To choose among these models, it is recommended to start with simple models and build up to more complex models.