ESS 469/569: Final Project

Project leads:

Iris Kern, [iriskern@uw.edu](mailto:iriskern@uw.edu), Github: @irissylvera

Josh Sacks, [jssacks@uw.edu](mailto:jssacks@uw.edu), Github: @jssacks

Project helper: Ayesha Mohammed, [ayeshamo@uw.edu](mailto:ayeshamo@uw.edu), Github: @ayeshazmo

Project GitHub:

<https://github.com/UW-ESS-DS/MLGEO2023_ESS569_MICROBIAL_METABOLITES>

**Predicting Marine Inorganic Nitrogen Concentrations from Microbial Metabolite Composition Using Machine Learning**

# **Scientific Motivation**

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 (Cloern et al., 2014). 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) (Martin and Fitzwater, 1988). 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 (Passow and Carlson, 2012). Global climate change is expected to drive increased ocean stratification and lower nutrient inputs into the surface ocean (Li et al., 2020) 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 (Moran et al., 2022). Metabolomics is the measurement of all small molecules in a system, often using liquid chromatography-mass spectrometry (Boysen et al. 2018). The marine microbial metabolome (all metabolites in a system) is controlled by a combination of taxonomy and cellular physiology (Heal et al. 2021, Boysen et al. 2021). 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.

Research Question:

**Can we predict inorganic nitrogen (N) concentrations from intracellular microbial metabolite concentrations?**

1. **Data**

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). We will also switch the comparison to test the ability of machine learning models trained on the environmental samples to predict the amended nutrient concentrations of the mesocosm experiment samples

PERI-DICE:

The PERI-DICE metabolite dataset is a set of samples collected from the PERI-DICE mesocosm 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 at the University of Washington, School of Oceanography. This dataset is currently unpublished and so was not evaluated for its FAIR principles.

The inorganic nitrogen concentrations (uM) that are being predicted from the various models implemented refers to the various amounts of added nitrate (NO3-) and ammonium (NH4+) to each mesocosm, which are two of the most predominant inorganic forms of nitrogen present in surface seawater. This experiment took place in a laboratory setting, implying significant control over and knowledge of the added inorganic nitrogen to each treatment group.

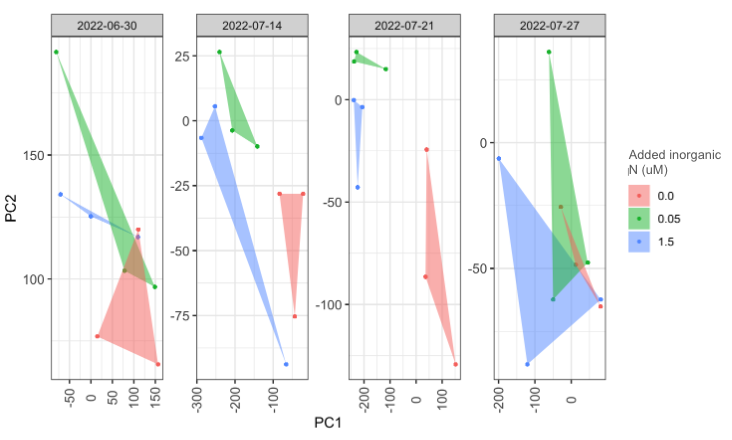


Figure 0.1 (Iris): Principal Component Analysis (PCA) of the metabolites present in the PERI-DICE dataset. The colors correspond with the amount of added inorganic nitrogen, showing groupings of metabolites based on that nutrient addition.

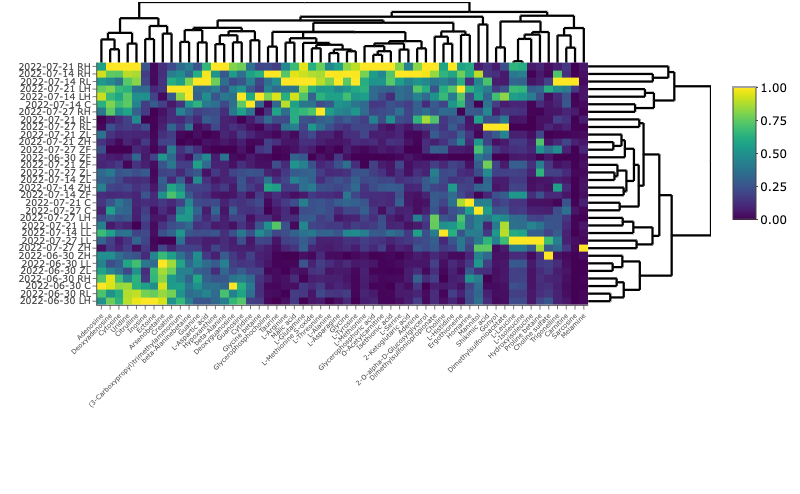


Figure 0.2 (Iris): Heatmap of metabolites present in PERI-DICE dataset. Shows clustering of treatments and metabolite concentration in minimum/maximum scaling space across the 4 week sampling period.

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. 2021. This dataset met most fair principles as it was findable and accessible (accessed through the supplemental information contained in the paper). It was not entirely interoperable and machine readable as it required some manual compilation from an excel sheet into a useful dataframe in a csv. The dataset was reusable as it included extensive metadata documentation that made it easy and straightforward to understand and incorporate into this project.

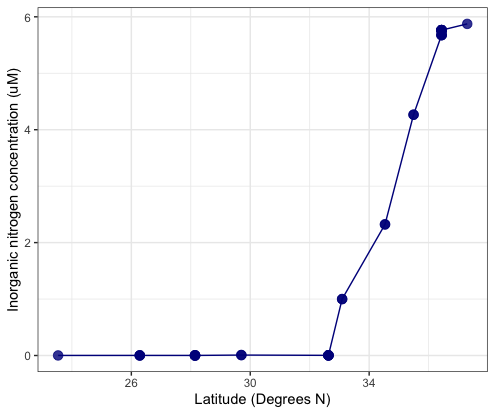


Figure 0.3 (Josh): Inorganic nitrogen concentrations along the Gradients 1 transect along 158 degrees W. The nitrogen concentrations show a relatively bimodal distribution with a sharp transition between very low concentrations in the south and high concentrations in the north.

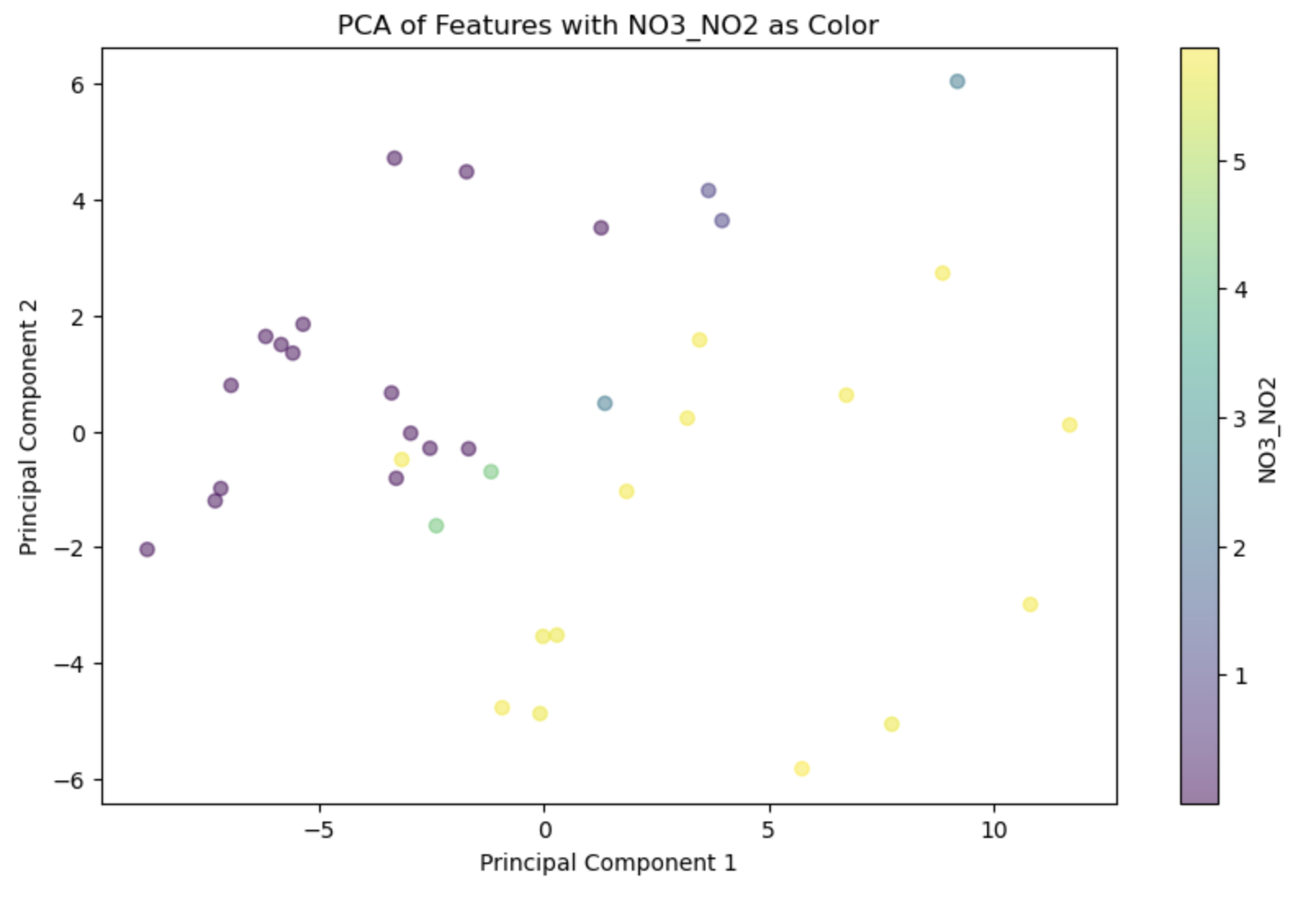


Figure 0.4 (Ayesha): The PCA of Gradients 1 is done using two Principal components. The distinct clusters can be identified from the colors.

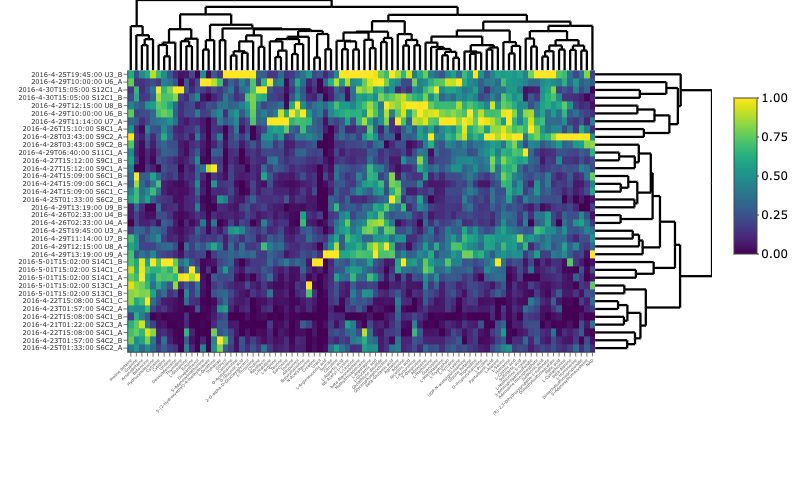


Figure 0.5 (Iris): Heatmap of metabolites present in Gradients 1 dataset. Shows concentration of metabolites in minimum/maximum scaling space at locations specified over transect.

Data Normalization:

Both the PERI-DICE and Gradients 1 datasets consist of raw metabolite concentrations (uM), inorganic nitrogen concentrations (uM), and particulate carbon concentrations (uM). The metabolite concentrations were normalized to biomass by dividing the metabolite concentrations by the particulate carbon concentrations for a fourth data column of normalized metabolite concentration (uM per uM). This is necessary as metabolite trends have a strong correlation with biomass, which is not a significant component of our analysis in predicting inorganic nitrogen concentrations. Normalizing to biomass minimizes variation across these diverse samples.

The other normalization conducted to filter for shared metabolites across datasets.. Unfiltered, the PERI-DICE dataset consists of 172 unique metabolites and the Gradients 1 dataset consists of 84 unique metabolites. There were 68 metabolites shared between the two datasets which

# **ML Methods, Training, and Performance**

Our problem is a supervised regression problem as we are trying to predict a continuous, known value (inorganic nitrogen concentration) from metabolite concentration. We therefore tried three regression type machine learning approaches: multiple linear regression (MLR), random forest regression (RFR), and principal component regression (PCR). This allows us to compare our approaches to identify the best one for our problem as well as testing the robustness of our results (do different methods give us the same answers?).

1. **Multiple Linear Regression:**

**Package:**

Scikit-learn package in Python ([Scikit-learn: Machine Learning in Python](https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.)

**Hyperparameters:**

No hyperparameters were set for the Linear Regression. The fit\_intercept parameters is set to “true”, however, this is done by default.

**Model Performance:**

PERI-DICE:

We began by training and testing the Peri-Dice data set. The model was evaluated using the MSE (mean squared error), and the R^2 value. To create the model, the PERI-DICE dataset was randomly split into two equal data sets. One was used to train the model, and the other was used to test. The model was fit for the data and then predicted values for the NO3\_NO2 concentration. A scatter plot was generated, the points consist of the actual vs predicted NO3\_NO2 concentrations.

Figure 1.1 displays the plot of PERI-DICE NO3\_NO2 actual values vs prediction values. The Test MSE was 0.597and the Train MSE was 0.686. The MSE measures the average squared difference between the actual and predicted values. In this case, the test MSE is slightly lower than the training MSE, suggesting that the model generalizes well to unseen data. The R-squared result for the test was 0.728 and for the train was 0.712. The R-squared value represents the proportion of variance in the dependent variable (NO3\_NO2) concentration) that is predictable from the independent variables (metabolites concentration). In this case, both the training and test R-squared values are relatively high, indicating that the model, explains a significant portion of the variance in the NO3\_NO2 concentration.

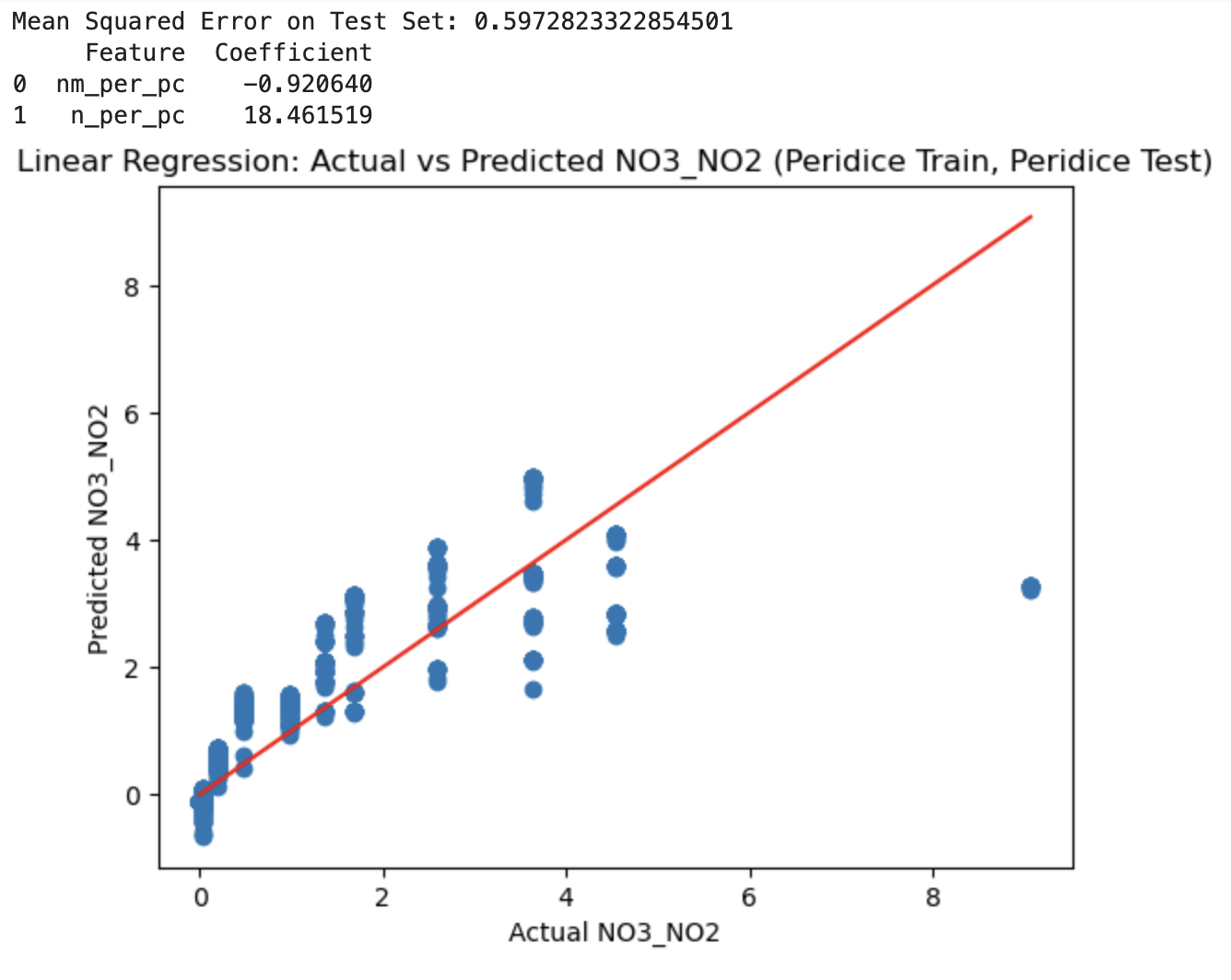
****

Figure 1.1 (Ayesha): The PERI-DICE dataset was used to train and test a Linear Regression model and the scatter plot of actual vs predicted results is depicted.

The plot of actual vs predicted NO3\_NO2 concentrations, however, depicts a different picture than the R-squared and MSE. The plot shows that the model is not capturing variability. The pattern indicates that the model may be oversimplified and not adequately capture the relationships in the data. These results could be because the model is limited and not right to capture the complexities of the relationship between data. The model may also be underfitting the training data and is not a right fit.

Gradients 1:

The Linear Regression model was next trained and tested using the Gradients 1 data set. This model was evaluated using the MSE and the R^2 values. The dataset, similar to PERI-DICE, was randomly split in half with one half to be used for training and the other half to be used for testing.

Figure 1.2 displays the plot of the Gradients 1 NO3\_NO2 actual values vs predicted values. The test MSE was 1.9 and the train MSE was 2.71e-28. The test R squared error was 0.7 and the train value was 1.0. Because the train MSE value was so close to 0, and the train R-squared is exactly 1, this indicates that the train data set is performing exceptionally well. Such a low training MSE and a perfect R-squared might suggest that the model has essentially memorized the training data fitting it perfectly. However, the test MSE is 1.9, which is significantly higher. This suggests that the model is not generalizing well to new, unseen data. The high test MSE indicates a substantial discrepancy between the model’s predictions and the actual values in the test set.

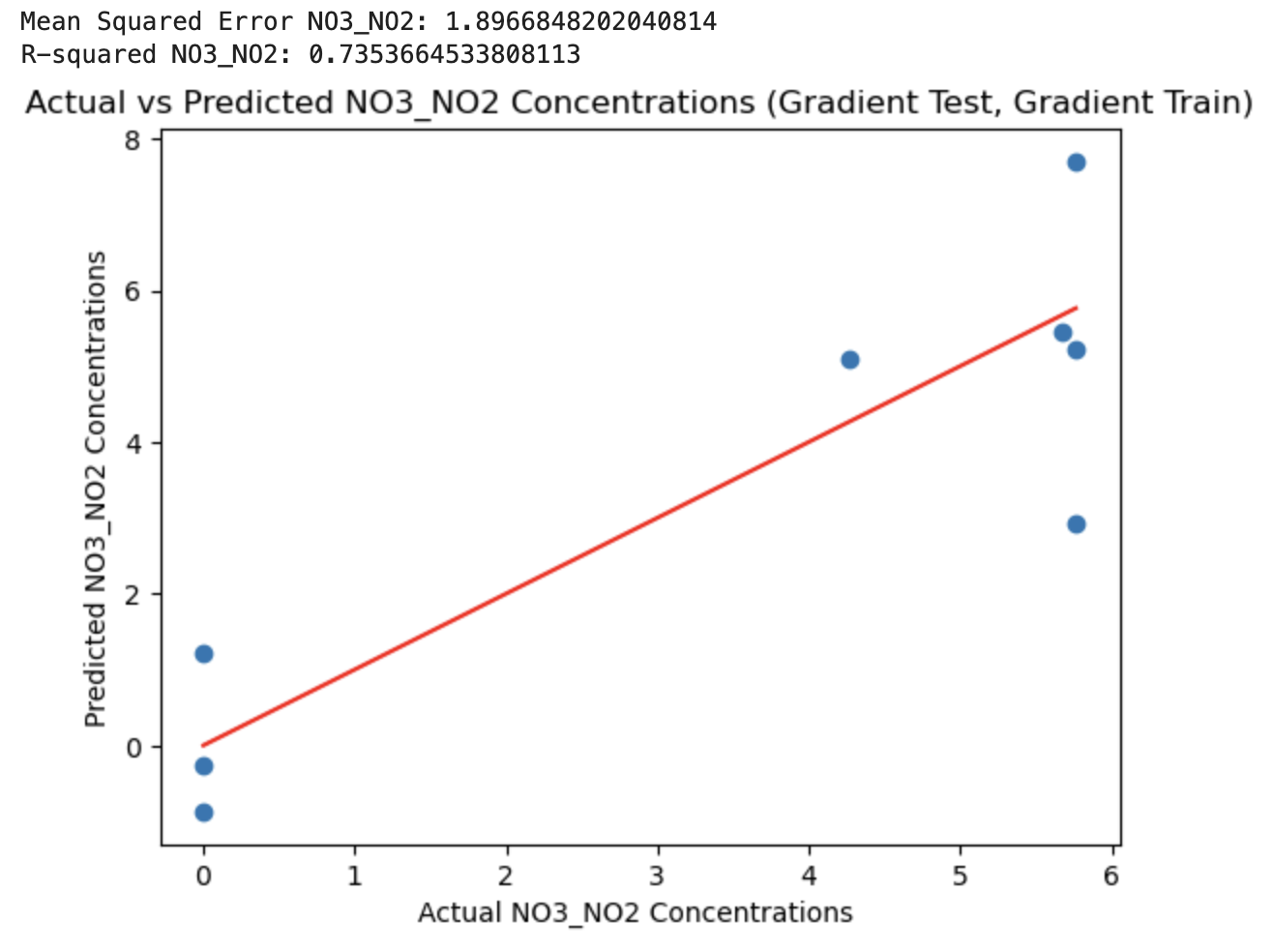
****

Figure 1.2 (Ayesha): Gradients 1 dataset was used to train and test a Linear Regression model and the scatter plot of actual vs predicted results is depicted.

The discrepancy between the test and train MSE and R-squared values is likely due to overfitting. The model appears to be overfitting the training data, capturing noise and specific patterns that don’t generalize well. It is possible that the model is not properly capturing the complex patterns of data.

**Performance Across Datasets:**

Next, we trained and tested the model by applying both the Gradients 1 and PERI-DICE datasets to each other. In Figure 1.3, the actual vs predicted plot of the model trained on the PERI-DICE data set and tested with the Gradients 1 dataset is shown. The test MSE is 14.6 and the train MSE is 0.668. The test R-squared is -1.05 and the train R-squared is 0.715. The new results indicate a significant increase in the MSE for the test set along with a negative R-squared value. These metrics suggest that the model’s performance has deteriorated, and it might be struggling to capture the underlying patterns in the data.

The model may be overfitting the training data. The large difference between the training and test MSE, along with the negative R-squared for the test set, suggests that the model may be overfitting by learning the training data too well and capturing noise. The model complexity may also not align with the data set.

The plot in Figure 1.3 shows that all predicted values are consistently below zero and are the same. This may have happened for a couple of reasons. The most likely reason is that the relationship between the data is non-linear. The model assumes a linear relationship exists between the dependent and independent variables. If no such relationship exists, the Linear Regression model will not be suitable. Another reason why such a distribution is shown is that the data sets are not similar enough. The PERI-DICE and Gradients 1 datasets may not be similar enough to work with each other.

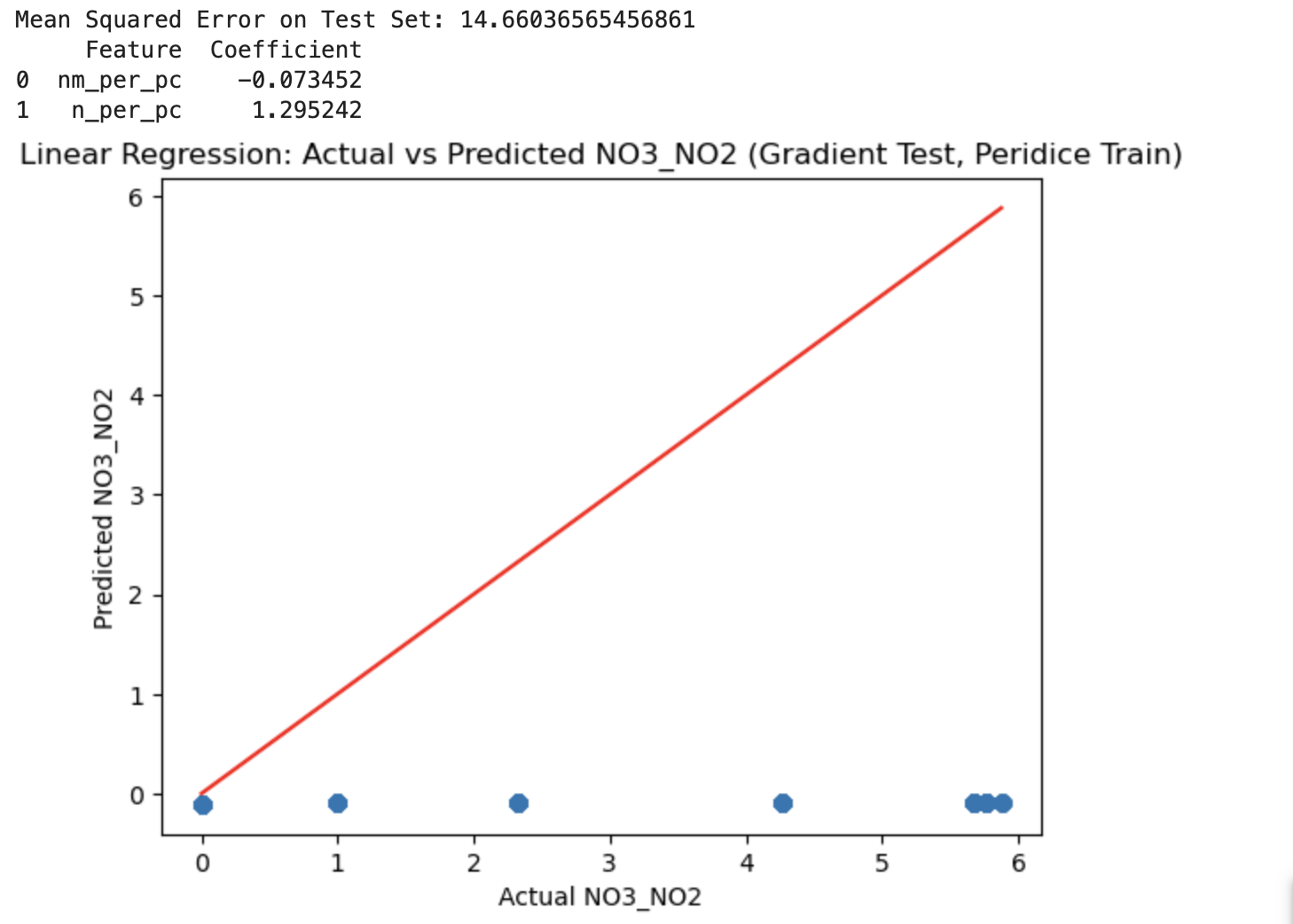
****

Figure 1.3 (Ayesha): The PERI-DICE dataset was used to train and the Gradients 1 dataset was used to test a Linear Regression model and the scatter plot of actual vs predicted results is depicted.

Finally, we trained a model with the Gradients 1 dataset and tested it with the PERI-DICE dataset. Figure 1.4 displays the actual vs predicted scatter plot for the model. This model performed the worst of all of the models. The test MSE was 244190 and the train MSE was 0.017. The test R-squared was -104059 and the train R-squared was 0.998. The results indicate a large discrepancy between the training and test set performance.e The training R-squared is very high (0.998), indicating an excellent fit to the training data. However, the test R-squared is negative, which is unusual and suggests a poor fit to the test data. Additionally, the test MSE is very high, indicating significant errors in the predictions on the test.

The model is likely overfitting the training data, as evidenced by the large difference between the training and test set performance. The relationship between independent and dependent variables may also not be linear. That would also explain the results.

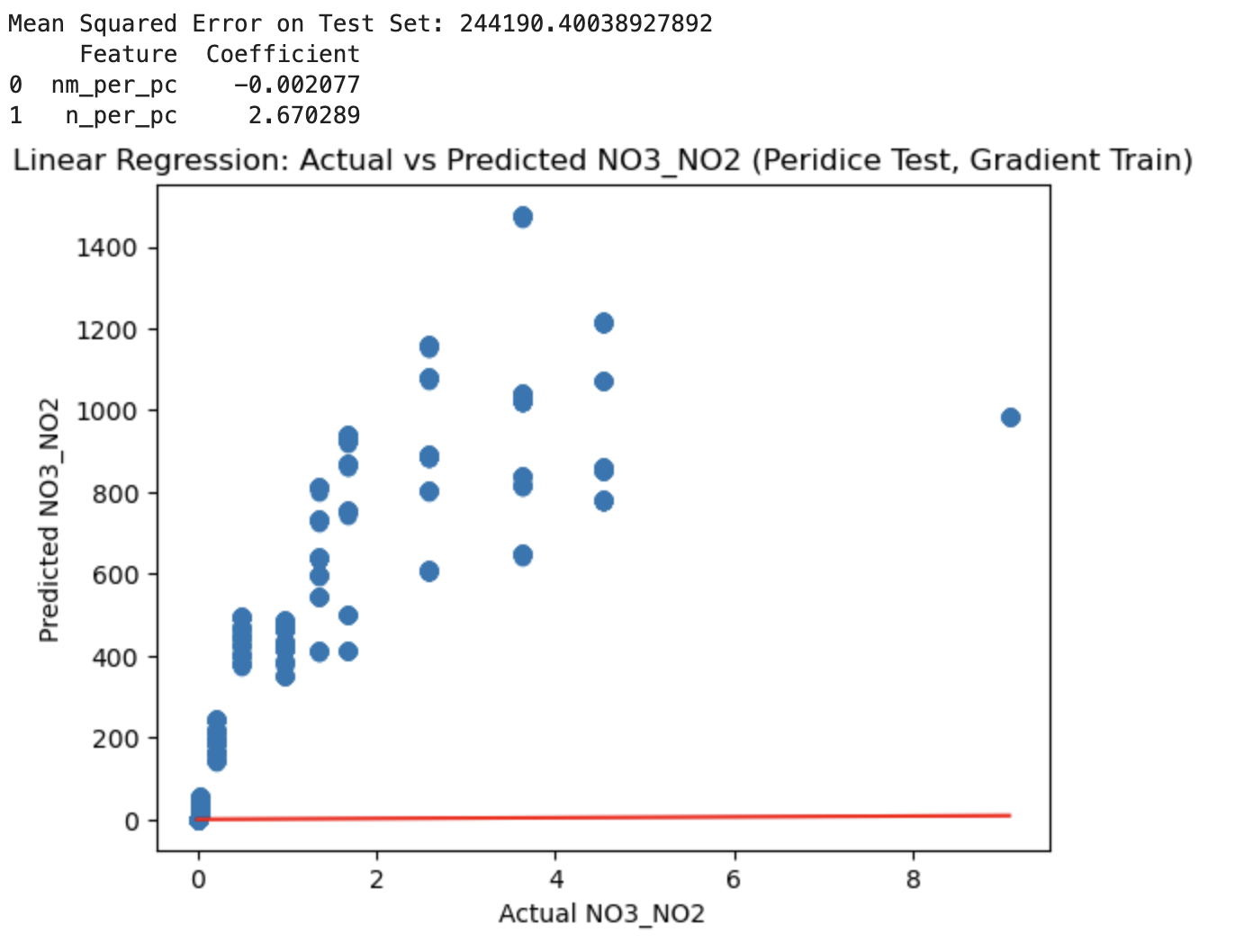


Figure 1.4 (Ayesha): The Gradients 1 dataset was used to train and the PERI-DICE dataset was used to test a Linear Regression model and the scatter plot of actual vs predicted results is depicted.

**Summary:**

Overall, the multiple linear regression models as currently shown may not be suitable for capturing the complexity present in the data. The common theme of overfitting and poor generalization across different datasets suggests that the linear regression models, as currently configured, struggle to capture the underlying patterns in data. Models trained on one dataset struggled to generalize well to a different dataset which led to poor performance on the test set. The extreme overfitting indicates a need for model simplification or regularization to prevent fitting noise in the training data. It is also important to note that it is likely that the data is non-linear. The constant poor performance hints at the non-linearity of the data. Non-linear relationships in the data may not be adequately captured by the simple linear models. There is also the possibility of the data sets being too different. Models trained on one dataset and tested on another show poor generalization, indicating that the datasets may have significant differences. In summary, the multiple linear regression model may not be suitable for this situation.

1. **Random Forest Regression:**

**Packages:**

RFR: Random Forest Package in R (https://www.stat.berkeley.edu/~breiman/RandomForests/cc\_home.htm)

Hyperparameter Tuning: Caret package in R (<https://topepo.github.io/caret/>)

**Hyperparameters:**

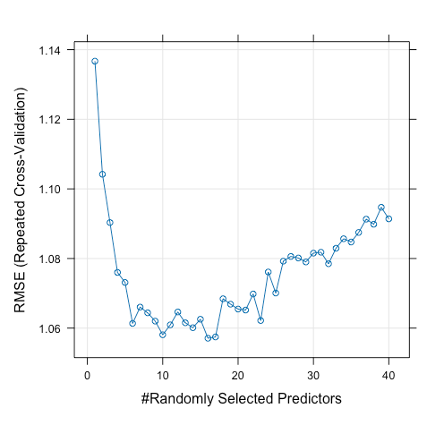
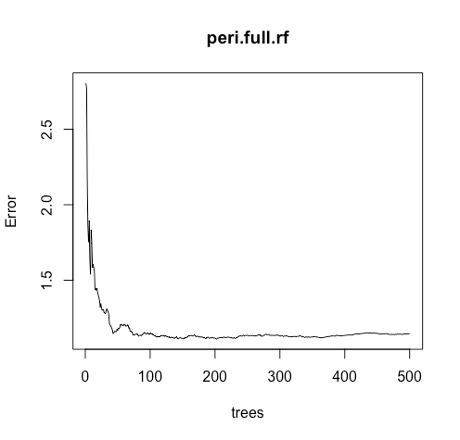
The two hyperparameters that can be tuned in RFR are the number of trees (ntree) and the number of variables considered at each decision point (mtry). The number of trees used is usually selected be a large enough number (in this case, 500) such that the mean squared error of the regression converges. Therefore, this value is typically not tuned. The mtry parameter is therefore the main hyperparameter tuned for RFR. We tuned mtry for each RFR model (G1 and PERI-DICE) independently using a grid search approach provided by the caret package in R. We searched values for mtry between 1 and 40 and found that 16 for PERI-DICE and 32 for G1 were the optimal mtry values based on having the lower root mean squared error values. Example plots showing the OOB MSE for different numbers of trees and the MSE for different mtry values for the PERI-DICE RFR model are shown in Figure 2.1 

Figure 2.1 (Josh): Hyperparameter selection for the PERI-DICE RFR model. Left panel shows the MSE (error) with increase numbers of decisions trees included in the ensemble. Right panel shows the MSE for different numbers of variable predictors included at each step in the RFR. The optimal number was shown to be 16 as it had the lowest MSE value.

**Model Performance:**

PERI-DICE:

We started by training and testing on the entire PERI-DICE dataset. We evaluated this using the out of bag (OOB) mean squared error rate (MSE) and the pseudo-R2 value (here after R2) to evaluate how well the RFR performed. Ensemble models such as RFRs randomly sample from the existing sample set in a bootstrapped manner when creating each decision tree. Each decision tree can then be evaluated on the samples not included in the tree (the “out of bag” samples) to validate the method performance. The R2 value provides an estimate in the regression of how well the predicted data matches the actual data. The OOB MSE was 1.14 and the R2 was 40.74% for the PERI-DICE dataset suggesting that the model does a decent but imperfect job of predicting the N concentration in the mesocosms (Figure 2.2). To assess over and underfitting of the model, we randomly split the dataset into a “test” and “train” dataset and evaluated the MSER of the test dataset when predicted by the “train” dataset trained model. The result was an MSE of 1.13 which is close to the train model (MSE: 1.28) suggesting that there was not overfitting. The model may be underfit as it explains less than 50% of the variation.

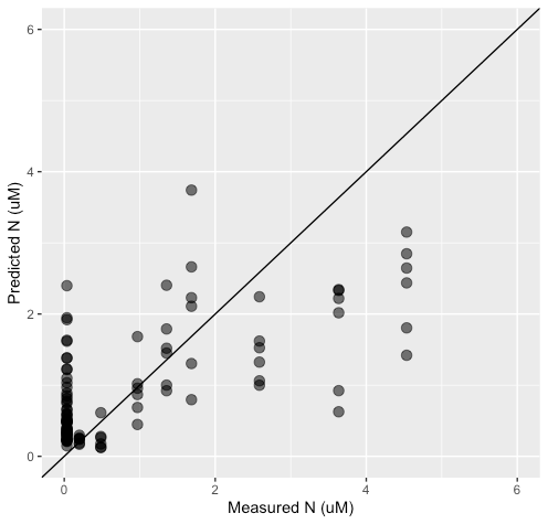


Figure 2.2 (Josh): OOB Predicted vs measured N for PERI-DICE RFR model. The black plotted line is the 1:1 line.

Gradients 1:

The G1 RFR model trained and tested on the entire dataset had an OOB MSE of 1.04 and an R2 of 85.43% (Figure 2.3). This suggests that the model performs well at predicting N from metabolite concentrations. The assessment of overfitting by splitting the data into a “test” and “train” dataset showed potential for some overfitting. The MSE of the test set predicted by the model (MSE: 1.76) was higher than the MSE of the trained model (MSE: 1.26). The model is likely not underfit given its high predicted capacity with an R2 of greater than 80% of the split model and an R2 of greater than 85% for the full model.

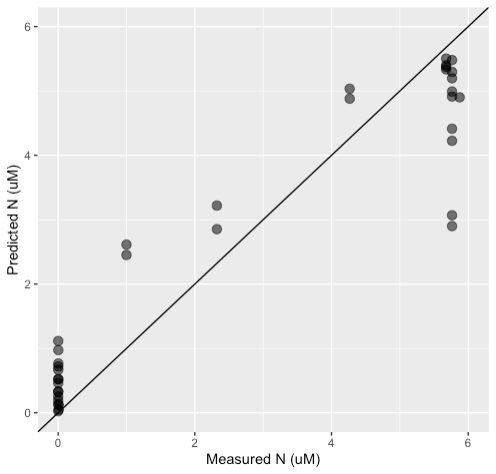


Figure 2.3 (Josh): OOB Predicted vs measured N for PERI-DICE RFR model. The black plotted line is the 1:1 line.

**Performance across datasets:**

We then tested the robustness and generalizability of the PERI-DICE and G1 trained RFR models by applying them to each other. The application of the PERI-DICE model to the G1 data (Figure 2.4) resulted in a MSE of 9.92 while the G1 model applied to the PERI-DICE data (Figure 2.5) resulted in an MSE of 11.71. These extremely high MSE values relative to the original models (MSEs of 1.04 to 1.14) paired with inspection of the model predictions suggested an almost complete lack of generalizability of these models. This lack of generalizability could be due to small training set size, batch specific variations in the metabolite datasets, fundamental differences between mesocosm experiments and environmental samples, or a combination of all three factors.

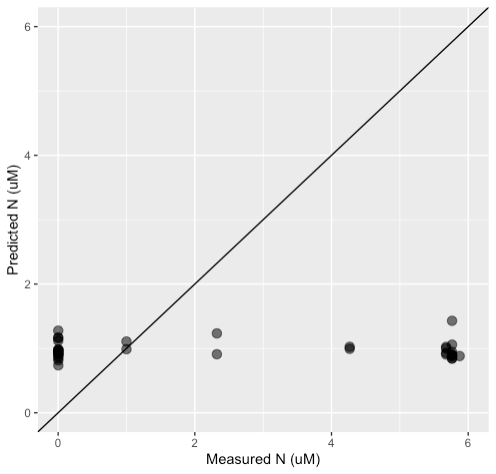


Figure 2.4 (Josh): Predicted N concentrations of G1 dataset by PERI-DICE RFR model. Black solid line is the 1:1 line.

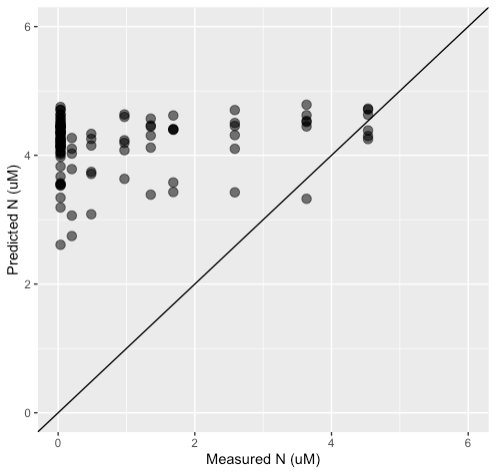


Figure 2.5 (Josh): Predicted N concentrations of PERI-DICE dataset by G1 RFR model. Black solid line is the 1:1 line.

**Summary:**

Both RFR models showed the ability to predict N concentration from metabolite concentration within each dataset. The G1 model performed exceptionally well, with the OOB predicted values explaining over 85% of the variance in the actual values. However, both models performed exceptionally poorly when applied to the other dataset suggest a severe lack of generalizability to these models.

**Insights from variable importance metrics:**

While both random forest models performed poorly when tested on each other the comparison of the 10 most important variables in each model contained 4 shared metabolites (Figure 2.6). Guanine, arsenobetaine, carnitine, and L-proline were important for both G1 and PERI-DICE suggesting that the concentrations of these metabolites are strongly related to inorganic nitrogen concentrations and therefore warrant future research. Guanine is particularly interesting as a predictor of inorganic nitrogen concentration as it is accumulated by dinoflagellates and other eukaryotic phytoplankton groups as a nitrogen storage compound under high nitrogen conditions (Mojzes et al. 2020). Carnitine is also intriguing as it is exclusively produced by eukaryotic species and therefore may indicate microbial community shifts from cyanobacterial dominance to eukaryotic dominance under changing nutrient conditions (Durham et al. 2022).

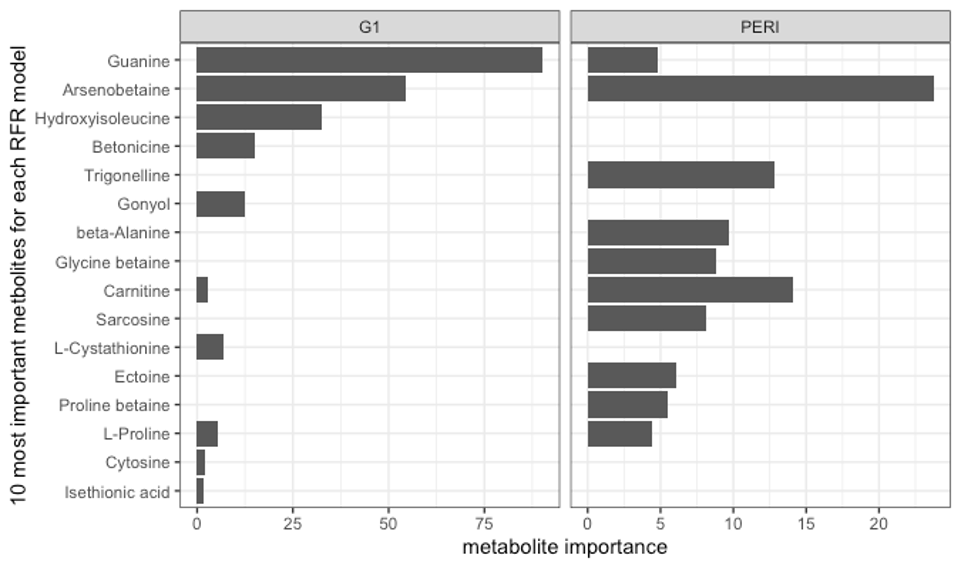


Figure 2.6 (Josh): Top 10 most important variables contributing to each RFR model.

1. **Principal Component Regression:**

**Package:**

Liland K, Mevik B, Wehrens R (2023). \_pls: Partial Least Squares and Principal Component Regression\_. R package version 2.8-3,

<https://CRAN.R-project.org/package=pls>

**Hyperparameters:**

The only hyperparameter to be considered in a Principal Component Regression is the number of principal components, which act as the regressor when implementing the multiple linear regression. For both datasets, the 1st and 2nd principal components were metabolic signature and metabolite concentration, which were implemented in both models to predict the inorganic nitrogen concentration (uM). Both the PERI-DICE and Gradients 1 datasets were filtered for metabolites determined from the metrics determined by the Random Forest Model (Figure 2.6). Based on the Principal Component Regression, the number of principal components that fits the most variation within the dataset was calculated to be 2 for both PERI-DICE and Gradients 1. The method of best fit was evaluated by using the Principal Component Regression Model and was established as Single Value Decomposition Principal Component (SVDPC). The root-mean squared error (RMSE) was calculated from the k-fold cross-validation method, with the Leave-One-Out (LOO) cross validation method. Any additional principal components results in an increase in the RMSE value.

**Model Performance:**

PERI-DICE:

The model training began by training and testing on the PERI-DICE dataset, using LOO Cross Validation. This model was evaluated by the hyperparameters discussed above, with 2 principal components accounting for 79% of the variance within the dataset. The RMSE calculated from the k-fold cross validation was 1.392 using 2 principal components.

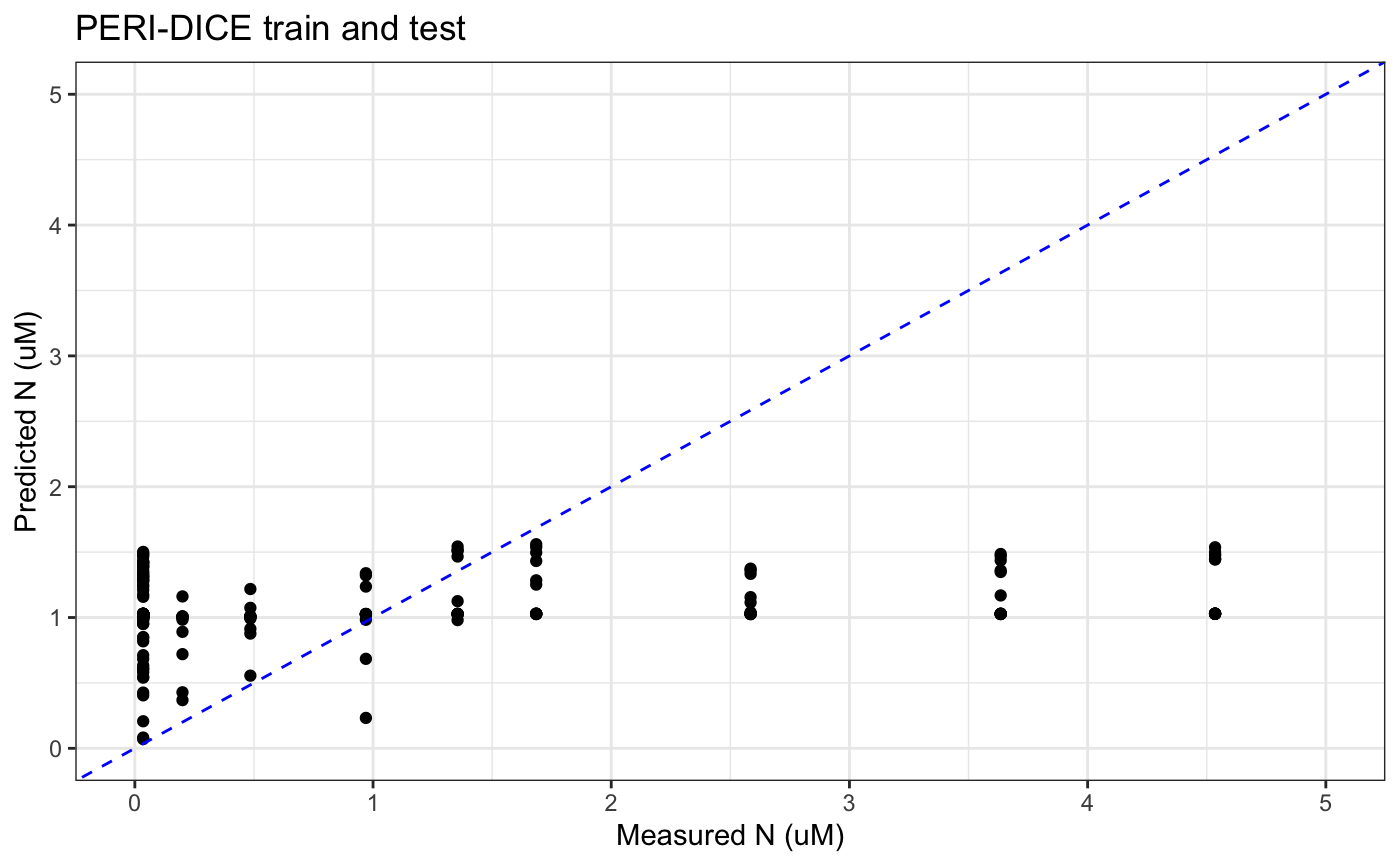


Figure 3.1 (Iris): Predicted vs. measured N (uM) from metabolite concentrations for the PERI-DICE PCR model. The blue dashed line is the 1:1 line.

Gradients 1:

The model training began by training and testing on the Gradients 1 dataset to develop an estimation of measured N (uM) from metabolite concentrations using LOO cross-validation. The metabolite concentrations were normalized to particulate carbon concentrations in the sample to reduce variation introduced from biomass. This model was evaluated using the hyperparameters discussed above, with 2 principal components accounting for 85% of the variance within the dataset. The RMSE calculated from the k-fold cross-validation was 2.204 using 2 principal components.

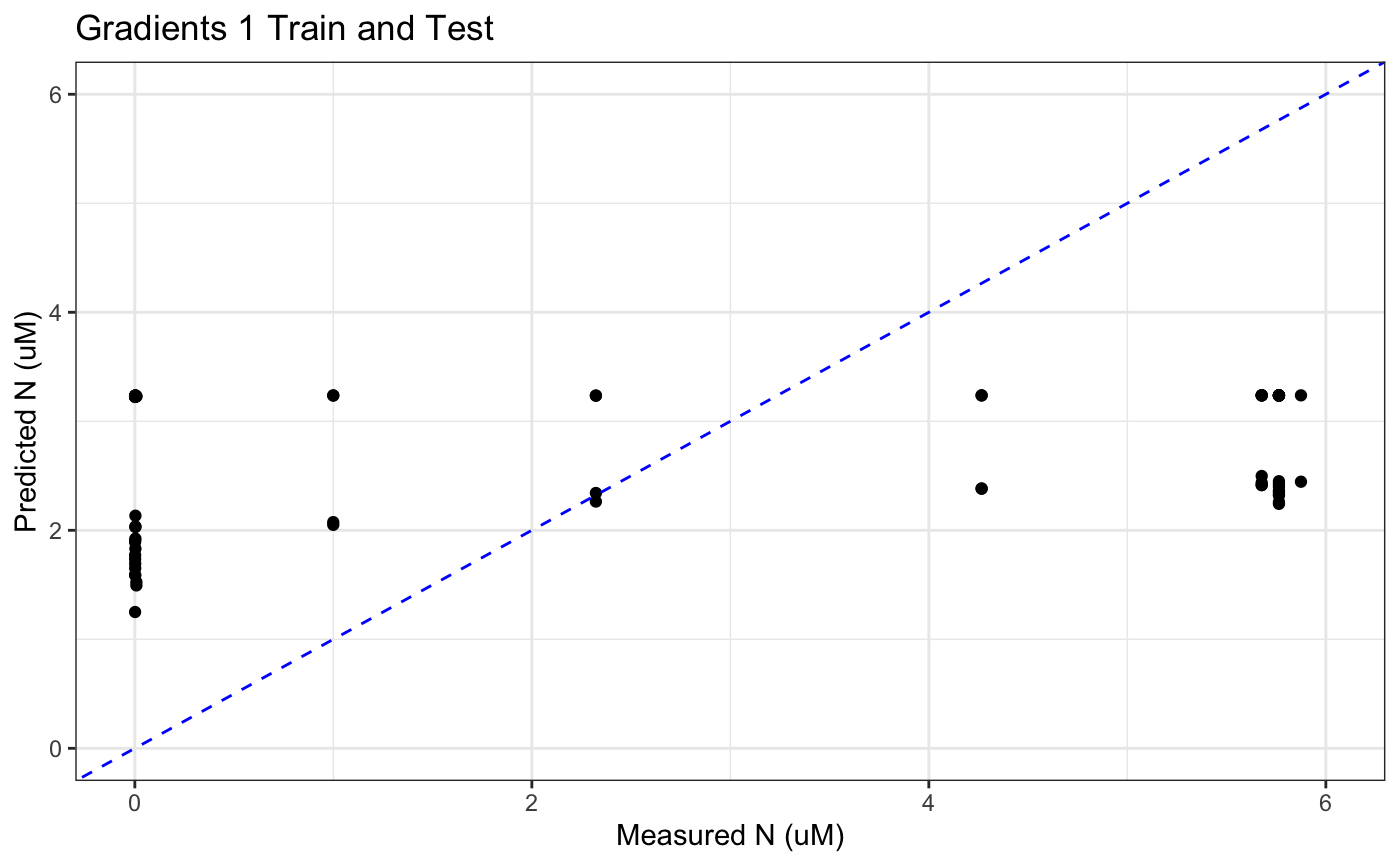


Figure 3.2 (Iris): Predicted vs. measured inorganic nitrogen (uM) from metabolite concentrations for the Gradients 1 PCR model. The blue dashed line is the 1:1 line.

**Performance Across Datasets:**

The next step in analyzing these models was to compare their applicability to other datasets. First the PCR model was trained on the PERI-DICE dataset and tested on the Gradients 1 dataset and then the PCR model was trained on the Gradients 1 dataset and tested on the PERI-DICE dataset.

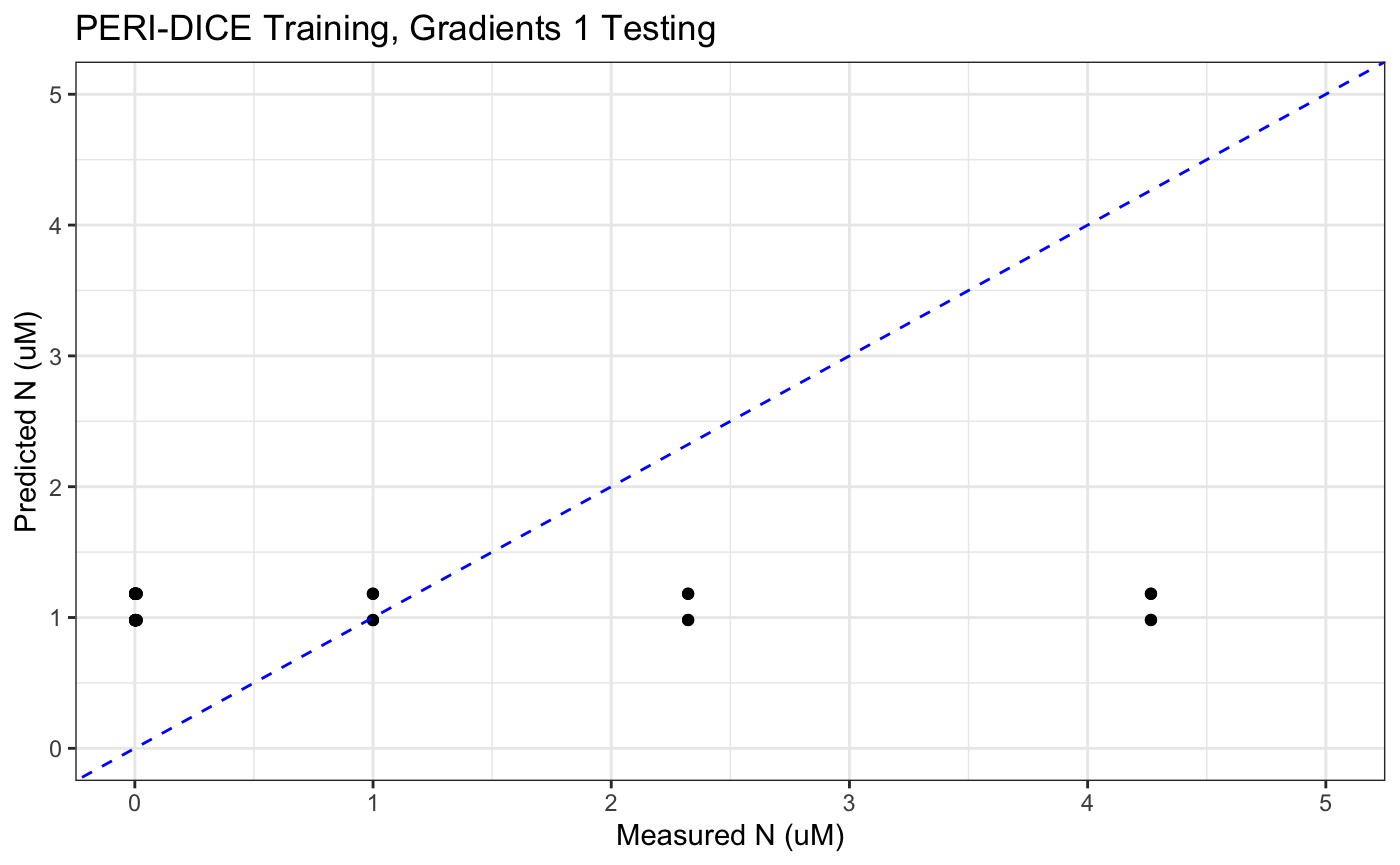


Figure 3.3 (Iris): Predicted vs. measured inorganic nitrogen (uM) for the PERI-DICE PCR model tested on the Gradients 1 data. The blue dashed line is the 1:1 line.

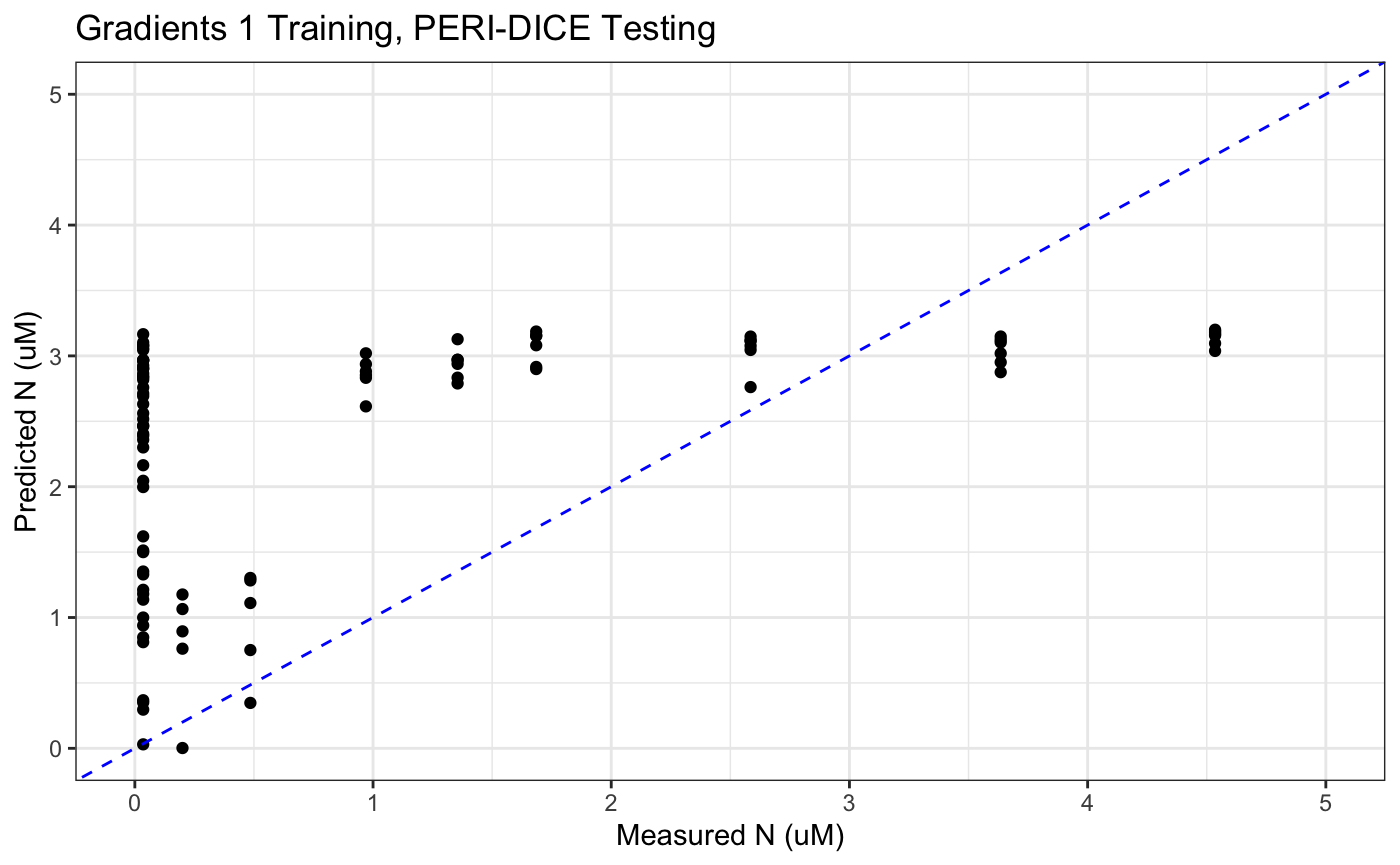


Figure 3.4 (Iris): Predicted vs. measured inorganic nitrogen (uM) for the Gradients 1 PCR model tested on the PERI-DICE data. The blue dashed line is the 1:1 line.

When the model was trained on PERI-DICE and tested on Gradients 1 data with 2 principal components, the resulting predictions had a root mean squared error (RMSE) of 9.322. When the model was trained on Gradients 1 and tested on PERI-DICE, the resulting predictions had an RMSE of 3.317.

**Summary:**

The Gradients 1 trained model performs significantly better on PERI-DICE than the PERI-DICE trained model does on the Gradients 1 model, however the RMSE values are significantly higher than when the models tested on themselves (1.392 and 2.204 respectively for PERI-DICE and Gradients 1). This high RMSE reflects a need to improve accuracy of the Principal Component Regression in order to be confident in its findings. These high RMSE values also reflect a potential overfitting of the model to their respective training datasets, as the natural variability of inorganic nitrogen as well as metabolite concentrations makes it difficult for a model based on metabolites in one portion of the ocean to accurately represent the metabolites in a different marine environment.

When developing a PCR model for either dataset, the number of metabolites used played a significant role in the accuracy of each one. Implementing every metabolite resulted in a lower accuracy of prediction, most likely as a result of the lack of a linear relationship between most metabolites and inorganic nitrogen concentrations. Some metabolites have a negatively correlated relationship with inorganic nitrogen concentrations, while others may have an exponential relationship. Selecting for the strongest indicators mentioned in Figure 2.6 allowed for increasing accuracy of both models.

# **Reproducibility**

The reproducible code can be found at the following link: <https://github.com/UW-ESS-DS/MLGEO2023_ESS569_MICROBIAL_METABOLITES>

The next steps for publication would be to define the hyperparameters more to lower the RMSE or MSE of each model. This could be done by prioritizing metabolites that are more powerful in determining the inorganic nitrogen concentrations as well as refining or normalizing the data to additional metadata beyond particulate carbon. This data could be of interest to a variety of journals, including ones such as *Nature*, which has a broad scope of interests, as well as journals with an emphasis on marine science, such as *Frontiers in Marine Science*.

# **Conclusion**

The Multiple Linear Regression models performed better when they were trained and tested on the same dataset. The models struggled to capture the patterns in the data correctly when trained with one dataset and then tested with the other. The random forest regression models performed best when tested on the same dataset as opposed to the other dataset. The Principal Component Regression performed significantly better when trained and tested on the same dataset, and the RMSE values were much higher when training and testing on opposing datasets. The Random Forest Regression minimized the MSE most significantly when training and testing on the same datasets (Gradients 1 model and PERI-DICE model had MSEs of 1.04 and 1.14 respectively), however the Principal Component Regression minimized RMSE most significantly when training on the Gradients 1 dataset and testing on PERI-DICE (RMSE of 3.317 compared to MSE of 11.71 when the Gradients 1 model was applied to PERI-DICE using the RFR). This is most likely a result of the Principal Component Regression being capable of minimizing variance and disregarding metabolite concentrations that are not related to inorganic nitrogen concentrations. The Gradients 1 dataset also reduces the error most significantly when predicting PERI-DICE, as compared to the PERI-DICE dataset maximizing error when applied to the Gradients 1 dataset.

The lack of ability to accurately predict inorganic nitrogen concentrations most likely originates from the simple fact that microbial communities (and subsequently their metabolites) are some of the most diverse components of the marine environment. No matter the machine learning method, making generalizations about inorganic nitrogen concentrations from such a complex system is likely to result in error and bias based on sampling methods and density of sampling in each region. The improvement of machine learning models for marine microbial communities will most likely need significant refinement in regards to the metabolites selected for sampling, developing known correlations between specific microorganisms and their metabolites in nitrogen-limited marine environments. Additionally, a machine learning model that is able to fully comprehend all of the diverse relationships between these metabolites and inorganic nitrogen concentrations needs to be able to understand the complexity of these relationships, especially when they might not necessarily have a linear relationship.

**References**

Cloern, J. E., Foster, S. Q., and Kleckner, A. E.: Phytoplankton primary production in the world's estuarine-coastal ecosystems, Biogeosciences, 11, 2477–2501, https://doi.org/10.5194/bg-11-2477-2014, 2014.

Martin, J., Fitzwater, S. Iron deficiency limits phytoplankton growth in the north-east Pacific subarctic. *Nature* 331, 341–343 (1988). https://doi.org/10.1038/331341a0

Passow U, Carlson CA (2012) The biological pump in a high CO2 world. Mar Ecol Prog Ser 470:249-271. <https://doi.org/10.3354/meps09985>

Li, G., Cheng, L., Zhu, J. *et al.* Increasing ocean stratification over the past half-century. *Nat. Clim. Chang.* 10, 1116–1123 (2020). https://doi.org/10.1038/s41558-020-00918-2

Moran, M.A., Kujawinski, E.B., Schroer, W.F. *et al.* Microbial metabolites in the marine carbon cycle. *Nat Microbiol* 7, 508–523 (2022). https://doi.org/10.1038/s41564-022-01090-3

Liland K, Mevik B, Wehrens R (2023). \_pls: Partial Least Squares and Principal Component Regression\_. R package version 2.8-3,

<[https://CRAN.R-project.org/package=pls](https://cran.r-project.org/package=pls)>

Heal, Katherine R., et al. "Marine community metabolomes carry fingerprints of phytoplankton community composition." *Msystems* 6.3 (2021): 10-1128. <https://doi.org/10.1128/msystems.01334-20>

Boysen, Angela K., et al. "Correction for Boysen et al.,“Particulate Metabolites and Transcripts Reflect Diel Oscillations of Microbial Activity in the Surface Ocean”." *Msystems* (2023): e01083-22. <https://doi.org/10.1128/msystems.00896-20>

Boysen, Angela K., et al. "Best-matched internal standard normalization in liquid chromatography–mass spectrometry metabolomics applied to environmental samples." *Analytical chemistry* 90.2 (2018): 1363-1369. <https://doi.org/10.1021/acs.analchem.7b04400>

Mojzeš, Peter, et al. "Guanine, a high-capacity and rapid-turnover nitrogen reserve in microalgal cells." *Proceedings of the National Academy of Sciences* 117.51 (2020): 32722-32730. <https://doi.org/10.1073/pnas.2005460117>

Durham, Bryndan P., et al. "Chemotaxonomic patterns in intracellular metabolites of marine microbial plankton." *Frontiers in Marine Science* 9 (2022): 864796. <https://doi.org/10.3389/fmars.2022.864796>

[Scikit-learn: Machine Learning in Python](https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.