Consulting Project Report

Variational Inference for Microbiome Survey Data

Department of Statistics
Ludwig-Maximilians-Universität München

Eesha Samir Chitnis and Iris Dania Jimenez

Munich, May 6th, 2025

Submitted in partial fulfillment of the requirements for Masters Statistics and Data Science (PO 2021)

Project Partners : Prof. Dr. Christian L. Müller and Viet Tran

Supervisor : Prof. Dr. David Rügamer

Abstract

In this study, we reproduced and extended the VI-MIDAS (Variational Inference for microbiome survey data) framework to model marine microbiome (mOTU) data from the Tara Oceans Expedition. As a proof of concept, we first replicated the original VI-MIDAS model—which incorporated environmental covariates via direct coupling—on the original dataset. We then introduced a novel variant, the no direct coupling model, which projected all covariates into a shared latent space. Both models were applied to two datasets: the original and an expanded version containing additional samples and satellite-derived covariates. Missing covariate values were imputed using a best-per-variable strategy via MICE. Exploratory analysis revealed distinct microbiome patterns in polar versus non-polar samples. The model comparison and ablation study showed that the no direct coupling variant consistently outperformed the direct coupling model across both datasets. Beyond predictive performance, the framework allowed for the visualization of the influence of environmental covariates on microbial taxa and the identification of taxon-taxon interactions, thus providing deeper insight into species abundance patterns.

Contents

1 Introduction

Marine microbial communities form the foundation of oceanic ecosystems, influencing global biogeochemical cycles and driving critical ecological processes such as nutrient recycling and primary production. These communities—composed of diverse microbial taxa including bacteria, archaea, viruses, and eukaryotes—exhibit immense taxonomic and functional diversity, shaped by a complex interplay of environmental, spatial, temporal, and biological factors. Modeling such communities is key to understanding how environmental gradients and species-species interactions regulate microbial composition and function in the world's oceans.

The advent of high-throughput sequencing and global expeditions such as the Tara Oceans project has enabled large-scale microbiome profiling across diverse oceanic regions. The Tara Oceans dataset provides microbiome abundance data (in the form of metagenomic taxonomic units, mOTUs), along with a rich set of environmental, satellite-derived, and spatiotemporal covariates. However, extracting meaningful associations from this high-dimensional, sparse, and overdispersed count data remains a significant statistical challenge.

Metagenomic Operational Taxonomic Units (mOTUs) are groups of microbes identified from metagenomic sequencing data. They provide species-level information about microbial communities without requiring complete genome assemblies, and are commonly used to quantify microbial abundance in large-scale studies such as the Tara Oceans project.

In this work, we reproduce and extend the VI-MIDAS framework using both the original Tara Oceans dataset and an expanded version enriched with satellite-derived covariates and additional samples. As part of our contributions, we also propose a novel model variant—the no direct coupling model—which projects all covariates into the latent space, rather than separating environmental effects via direct coupling. This enables a unified representation of contextual influences and offers greater flexibility in capturing latent ecological patterns.

We begin by addressing missing data through a principled imputation strategy using Multiple Imputation by Chained Equations (MICE), ensuring that the correlation structure is preserved. We then retain mOTUs based on empirical filtering criteria. Exploratory data analysis reveals consistent ecological differences between polar and non-polar regions.

2 Data

We have used two data sets as a part of this project. They will be referred to as 'original data' and 'new data'. The samples for both data sets were collected from the *'Tara Oceans Expedition'*. Each sample consisted of environmental, spatio-temporal, taxonomic abundance profiles of mOTUs and satellite-derived (only for the original data) covariates.

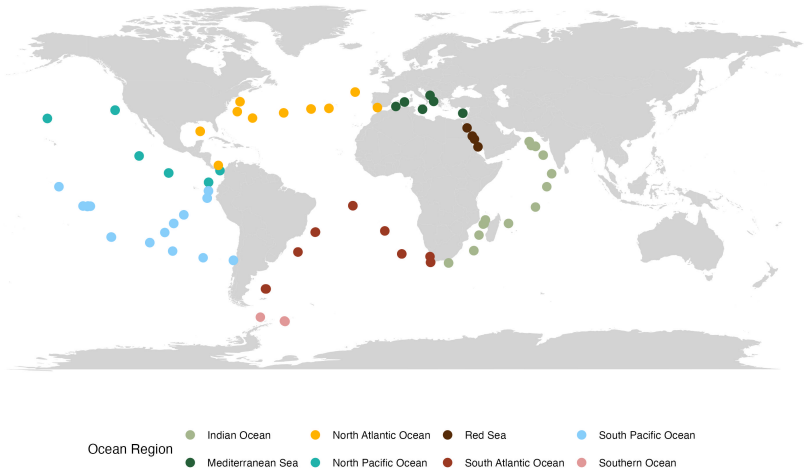

Figure 2.1: **Original Data**: n=139 samples from 8 Ocean regions

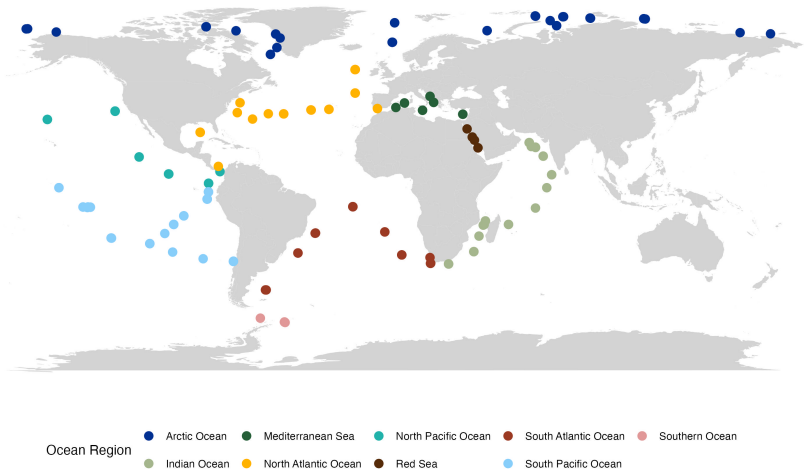

Figure 2.2: **New Data**: n=180 samples from 9 Ocean regions

2.1 Original Data

The original dataset in Mishra et al. (2024) consisted of $n = 139$ distinct water samples.

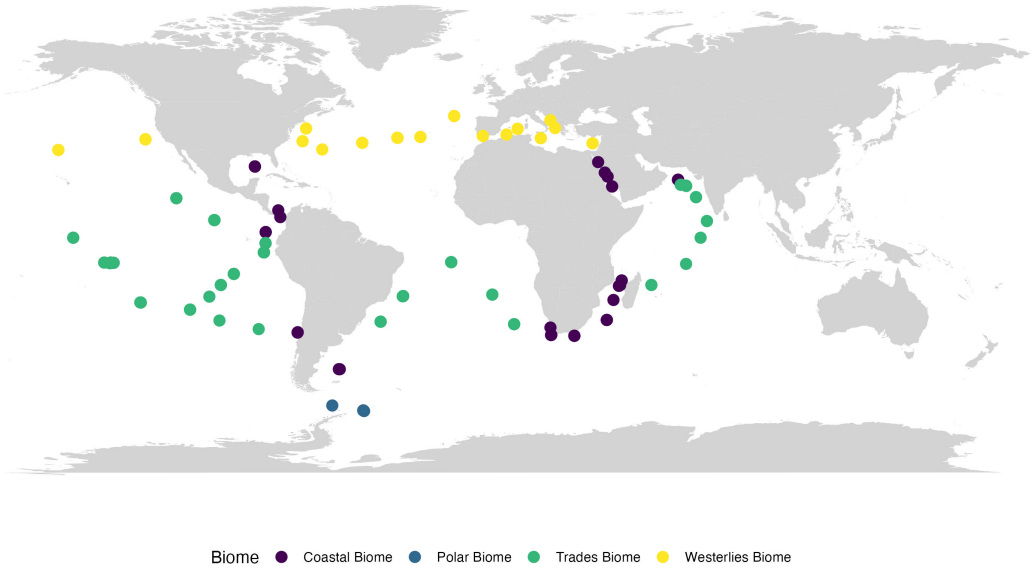

Figure 2.3: Original Data: n=139 samples by Biome

For each sample :

- There were $p = 9$ Environmental covariates:

Environmental Covariate	Unit of Measurement
Temperature	°C
Salinity	PSU
Oxygen	μmol/kg
NO_2	μmol/L
PO_4	μmol/L
NO_2NO_3	μmol/L
Si	μmol/L
SST (Sea Surface Temperature) Gradient	K/km
Nitrates	mg/L

Table 1: Original Data - Environmental Covariates

- Spatio-temporal covariates:
  - Depth Layer - SRF, DCM, MIX, MES (Figure 2.4)

– Province - Polar, Westerlies, Trades, Coastal (Biome, Figure 2.4)

– Season - 1, 2, 3, 4

Season	Quarter	Timeframe
1	Q1	January-March
2	Q2	April-June
3	Q3	July-September
4	Q4	October-December

Table 2: Original and New Data - Season

Figure 2.4: Spatio-temporal covariates – Depth Layer and Biome (Longhurst Province) classifications as used in the analysis. The descriptions are adapted from Mishra et al. (2024).

Across the 139 samples, there were more than 35,000 mOTUs. To reduce sparsity and focus on the most abundant and informative taxa, only those contributing to the top 40% of the total library size per sample were retained (Mishra et al., 2024). This resulted in the retention of q = 1,378 mOTUs.

2.2 New Data

The dataset was constructed to expand on the original dataset by incorporating 41 additional samples, which primarily came from the Polar Biome (Artic Ocean region). Satellite-derived covariates were also incorporated into the model along with the environmental and spatio-temporal covariates.

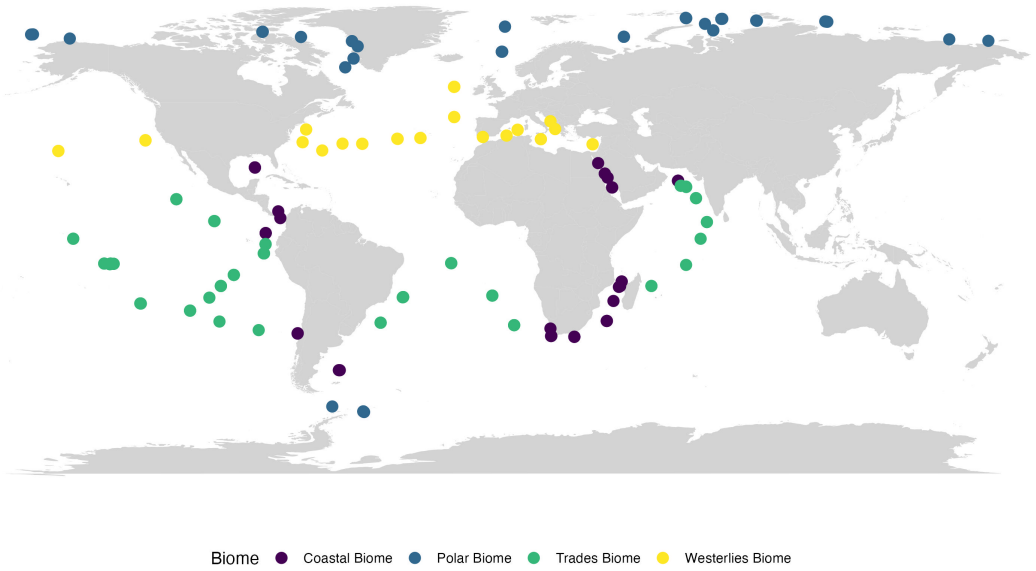

Figure 2.5: New Data: n=180 samples by Biome

The new dataset consisted of $n = 180$ distinct water samples across nine ocean regions. For each sample :

- There were $p = 10$ Environmental covariates:

Environmental Covariate	**Unit of Measurement**
Temperature	°C
Salinity	PSU
Oxygen	μmol/kg
NO_2	μmol/L
PO_4	μmol/L
NO_2NO_3	μmol/L
Si	μmol/L
SST (Sea Surface Temperature)	°C
ChlorophyllA	mg/m^3
Carbon.total	μmol/L

Table 3: New Data – Environmental Covariates

- Spatio-temporal covariates:
  - Depth Layer - SRF, DCM, MIX, MES (Figure 2.4)
  - Province - Polar, Westerlies, Trades, Coastal (Biome, Figure 2.4)

  - Season - 1, 2, 3, 4 (Table 2)

- There were 9 Satellite-derived covariates:

Satellite Covariate	Unit of Measurement
Fluorescence	RFU
Chl (Chlorophyll-a concentration)	mg/m^3
PAR (Photosynthetically Active Radiation)	μmol photons m^{-2} s^{-1}
MLD (Mixed Layer Depth)	m
Wind	m/s
EKE (Eddy Kinetic Energy)	m^2/s^2
Rrs490 (Remote Sensing Reflectance at 490 nm)	sr^{-1}
Rrs510 (Remote Sensing Reflectance at 510 nm)	sr^{-1}
Rrs555 (Remote Sensing Reflectance at 555 nm)	sr^{-1}

Table 4: Satellite covariates and their units of measurement

2.2.1 Imputation

Missing data is a pervasive issue in real-world datasets and we had missing values present in the Environmental and Satellite-derived covariates. As shown in Figure 2.6 the proportion of missing values varied significantly across variables — with some features such as EKE and Carbon.total having almost 30% missingness, while others such as Temperature and Salinity have less than 2% missingness. To address this, we implemented Multiple Imputation by Chained Equations (MICE) using 'IterativeImputer' from 'scikit-learn'. Our imputation pipeline consists of data preparation, multiple imputations, simulation-based evaluation using Root Mean Squared Error (RMSE), and a hybrid strategy that selects the best imputation per variable to construct the final dataset.

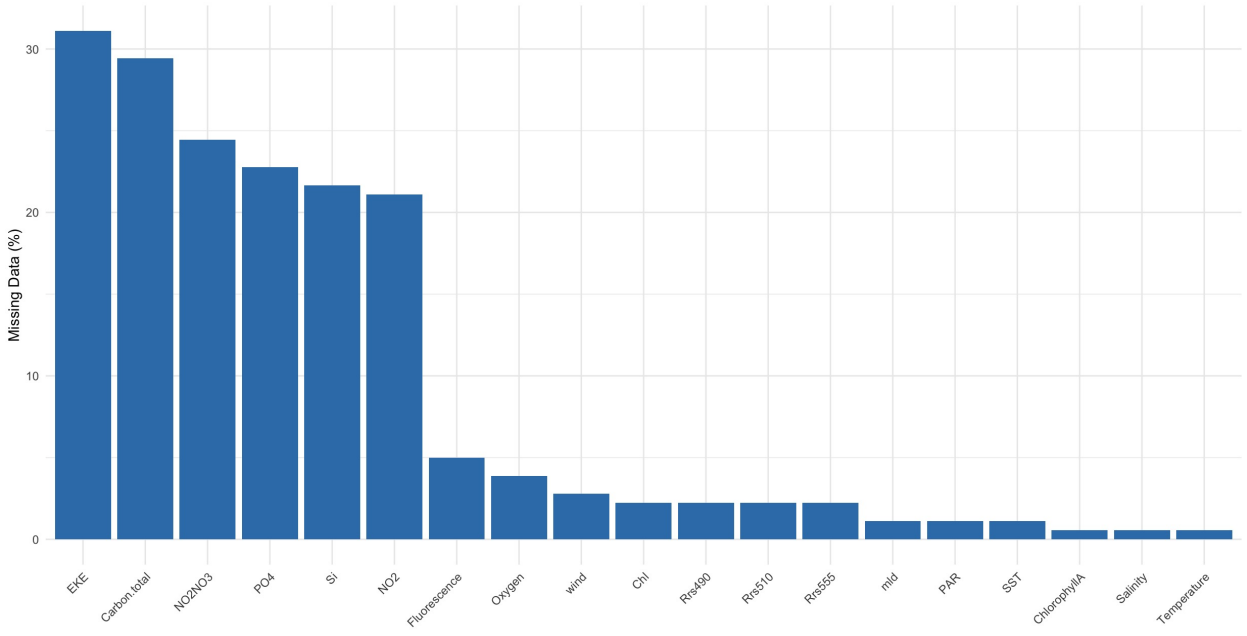

Figure 2.6: Missing Percentage of the new data's Environmental and Satellite covariates

Step 1: Preprocessing and Feature Selection

Given a dataset $\mathbf{T} \in \mathbb{R}^{n \times p}$ comprising the covariates which are numeric.

$$\mathbf{T}_{\text{sub}} = \left[t_1, t_2, \ldots, t_p\right], \quad \text{where } t_i \in \mathbb{R}^n$$

Only numeric features t_i are retained for imputation, denoted by $\mathcal{T}_{\text{num}} \subseteq \{t_1, \ldots, t_p\}$.

To improve imputation stability, especially in the presence of outliers, we apply `RobustScaler` from `scikit-learn`, which performs:

$$t_i^{(\text{scaled})} = \frac{t_i - \text{median}(t_i)}{\text{IQR}(t_i)} \quad \text{where IQR} = Q_3 - Q_1$$

Step 2: MICE Imputation (per run)

MICE is an iterative imputation technique where each feature with missing values is modeled as a function of the other variables. In our implementation, we use `IterativeImputer` from `scikit-learn`, which automates this process.

At the start of imputation, missing values are not left empty. Instead, `IterativeImputer` initializes them using the mean of the observed values in each column. This forms an initial imputed matrix $T^{(0)}$ internally, although we do not compute it manually in our code. Conceptually:

$$t_{ij}^{(0)} = \begin{cases} t_{ij}, & \text{if observed} \\ \text{mean of column } j, & \text{if missing} \end{cases}$$

Once this initial guess is in place, the algorithm proceeds iteratively. At each iteration:

1. For every feature t_j with missing values, a regression model f_j is trained using the other features T_{-j} as predictors.
2. The model f_j is then used to predict the missing entries in t_j, using the most recent values of the other variables.
3. The predictions replace the missing values, and the process moves to the next variable.

This cycle is repeated for a fixed number of iterations (we use $T = 20$), gradually refining the imputed values. The model used for regression is a Bayesian Ridge Regressor by default, which handles multicollinearity well and helps stabilize the imputations.

After looping through all features once, a single iteration is complete. This process is repeated for 20 iterations, after which one fully imputed dataset is obtained.

Step 3: Multiple Imputations, Simulated Missingness, and RMSE Evaluation

We perform the MICE process $K = 5$ times using `IterativeImputer`, each time with a different random seed. This introduces variability in the regression models and imputation paths, producing K slightly different complete datasets:

$$\left\{\widehat{T}^{(1)}, \widehat{T}^{(2)}, \ldots, \widehat{T}^{(K)}\right\}$$

Each of these imputed datasets represents a plausible version of the complete data and is retained for downstream evaluation.

To assess imputation accuracy, we begin with a version of the dataset that has no missing values in its numeric features. We then simulate missingness under the Missing Completely At Random (MCAR) assumption by randomly masking 10% of the entries. This creates a controlled test case where the true values are known, enabling direct evaluation of imputation accuracy.

In practice, we generate a boolean mask to simulate missing values. For each entry (i, j) in the dataset, we sample a uniform random number from $[0, 1]$ and set:

$$\text{mask}_{ij} = \begin{cases} \text{True}, & \text{if Uniform}(0,1) < \alpha \\ \text{False}, & \text{otherwise} \end{cases} \quad \text{with } \alpha = 0.1$$

All entries where the mask is `True` are replaced with `NaN`, simulating 10% missingness across the dataset.

We then apply the MICE procedure to this artificially incomplete dataset using the same K random seeds, resulting in another set of K imputed datasets:

$$\left\{\widehat{T}^{(1)}_{\text{sim}}, \widehat{T}^{(2)}_{\text{sim}}, \ldots, \widehat{T}^{(K)}_{\text{sim}}\right\}$$

Since the original values are known, we compute the Root Mean Squared Error (RMSE) for each imputation and each numeric variable over the artificially missing entries:

$$\text{RMSE}_j^{(k)} = \sqrt{\frac{1}{|\mathcal{M}_j|} \sum_{i \in \mathcal{M}_j} \left(\hat{t}_{ij}^{(k)} - t_{ij}\right)^2}$$

Where:

- $\mathcal{M}_j$ is the set of row indices where feature j was masked (i.e., set to `NaN`) during the simulated missingness step.
- t_{ij} is the true (original) value from the complete dataset at row i, column j.
- $\hat{t}_{ij}^{(k)}$ is the imputed value for the same entry, predicted by the k-th MICE run.

We aggregate the results into an RMSE matrix $\mathbf{R} \in \mathbb{R}^{K \times p}$, where $\mathbf{R}_{kj}$ represents the RMSE for variable j in imputation run k. This evaluation allows us to compare imputation quality both across variables and across different MICE runs.

Step 4: Best-Per-Variable Imputation

Rather than selecting one of the K imputed datasets as a whole, we construct a hybrid dataset by selecting, for each variable, the imputation run that achieved the lowest Root Mean Squared Error (RMSE) during evaluation.

For each variable t_j, we identify the best-performing imputation run:

$$k_j^* = \arg\min_k \mathrm{RMSE}_j^{(k)}, \quad \forall j \in \{1, \ldots, p\}$$

We then use the imputed values from run k_j^* for all originally missing entries in variable t_j. The final hybrid dataset $\mathbf{T}_{\text{best}}$ is constructed as:

$$t_{ij}^{(\text{best})} = \begin{cases} t_{ij}, & \text{if observed in original data} \\ \hat{t}_{ij}^{(k_j^*)}, & \text{if missing in original data} \end{cases}$$

This column-wise selection strategy ensures that, for each variable, the imputed values come from the run that minimized error in the simulated evaluation. As a result, the final dataset benefits from variable-wise optimized imputations, rather than being limited by the performance of any single imputation run.

Step 5: Visualization of RMSE results

To communicate performance:

- A grid of bar plots visualizes RMSE for each variable across all K imputations.
- The bar corresponding to k_j^* is highlighted to indicate the best imputation for feature j.
- RMSE values are plotted as:

$$\mathrm{Bar}_j = \left[\mathrm{RMSE}_j^{(1)}, \ldots, \mathrm{RMSE}_j^{(K)}\right]$$

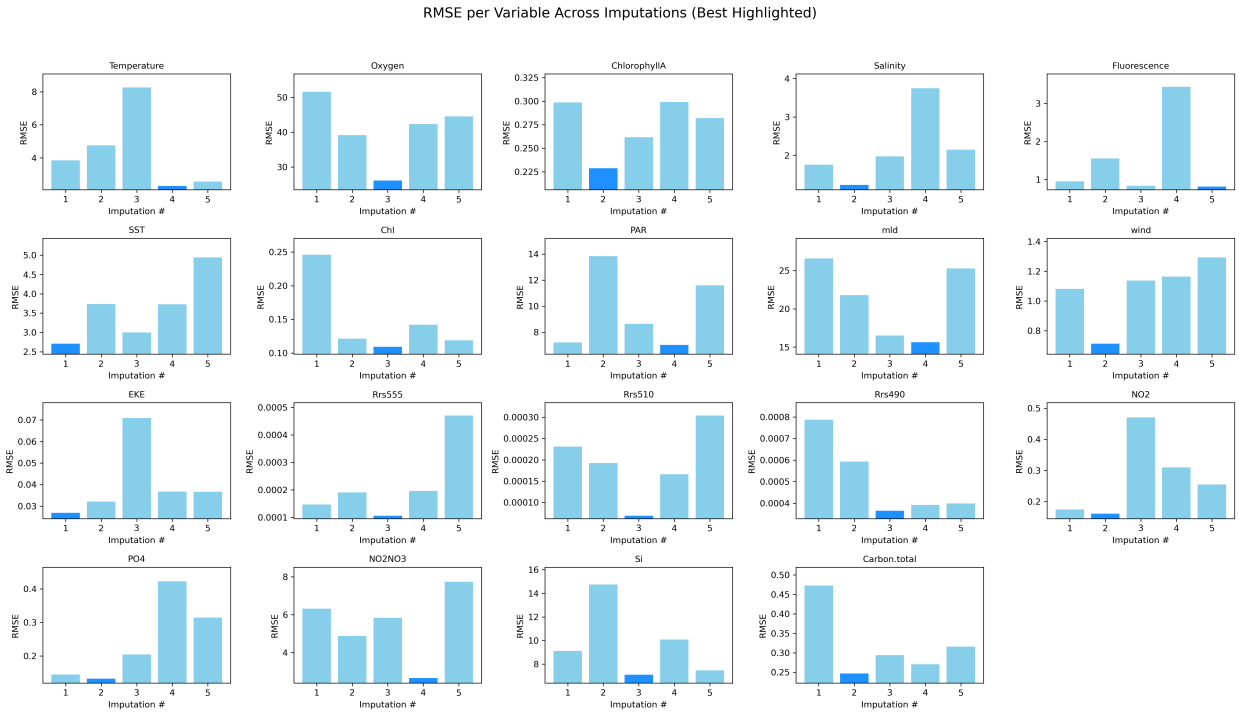

Figure 2.7: RMSE for simulated missing values across five MICE imputations per variable. Each bar represents the RMSE for one imputation run. The dark blue bar indicates the lowest RMSE for that variable, which was used to select the imputed values in the final dataset.

Correlation Structure Before and After Imputation

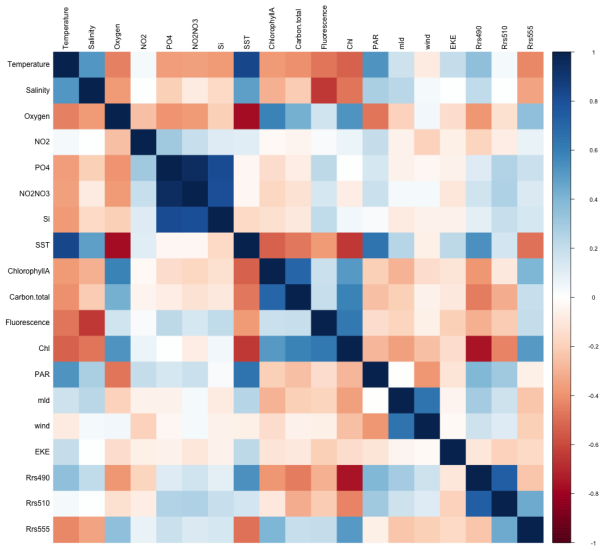

Figure 2.8: Correlation before imputation

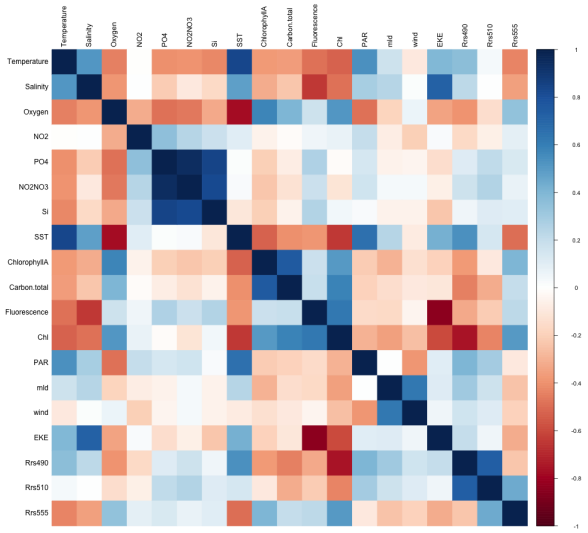

Figure 2.9: Correlation after imputation

To evaluate whether imputation preserved the underlying data structure, we compared the pairwise correlations among environmental and satellite-derived variables before and after imputation. Figure 2.8 shows the correlation matrix computed from complete cases (i.e., only rows with no missing values), while Figure 2.9 shows the correlation structure after imputation using the best-per-variable strategy.

Overall, the correlation structures are consistent across both plots, suggesting that imputation did not introduce spurious relationships or distort the underlying associations among variables. Some pairwise correlations appear slightly stronger post-imputation, likely due to the increased sample size when missing entries are filled in.

2.2.2 Understanding Environmental Covariates - Polar vs Non-polar

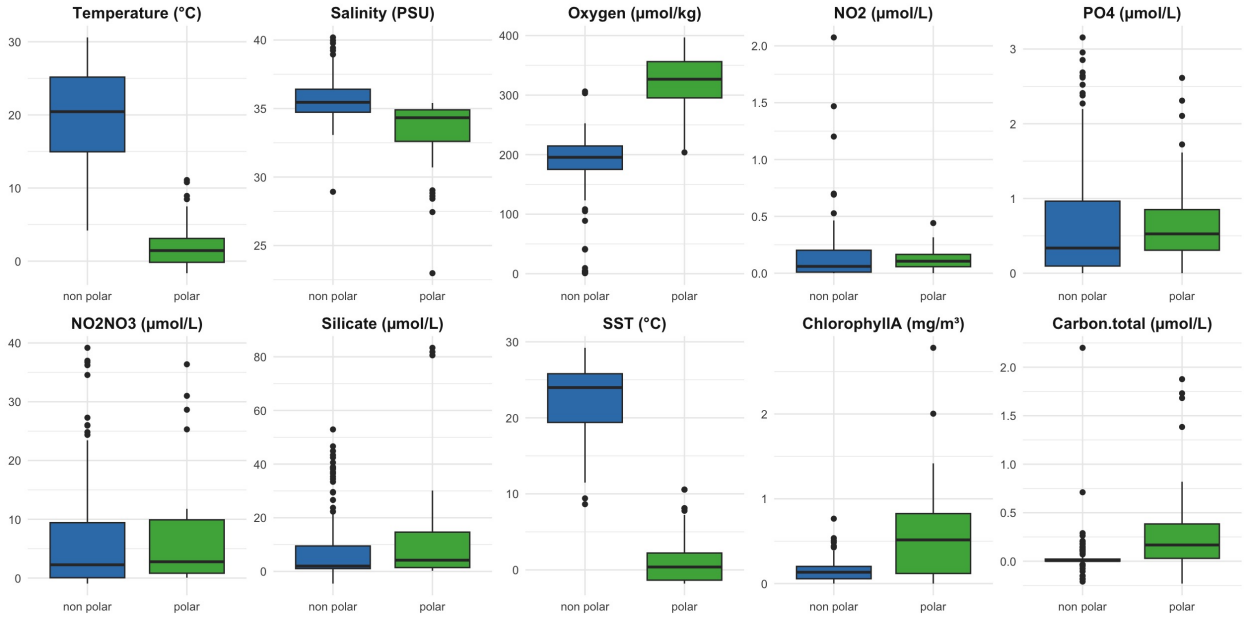

Figure 2.10: Environmental Covariates - Polar vs Non-polar

Given that the additional 41 samples were primarily collected from Polar regions, we assessed environmental covariate differences between Polar and Non-Polar samples after imputation.

Our key observations were:

- **Temperature and SST:** Non-Polar regions exhibit significantly higher temperatures, while Polar regions are mostly constrained to the 0–5°C range.
- **Salinity:** While median surface salinity is slightly higher in Non-Polar regions, Polar samples show broader variability. This pattern is consistent with known oceanographic dynamics: at the poles, melting ice during warmer months dilutes surface waters, reducing salinity, whereas in winter, ice formation expels salt, increasing subsurface

salinity (NOAA, 2023). These seasonal processes may contribute to the wider salinity range observed in Polar samples. In contrast, Non-Polar regions, especially in subtropical zones, experience high evaporation and limited freshwater input, concentrating salts and elevating salinity levels.

- **Oxygen:** Polar waters exhibit higher dissolved oxygen concentrations due to the increased solubility of oxygen in colder temperatures. This aligns with the physical principle that colder water holds more dissolved oxygen. The wider spread in polar oxygen values may reflect environmental variability such as seasonal mixing, sea ice dynamics, or differing sampling depths, whereas lower and more tightly clustered oxygen values in non-polar regions could result from warmer temperatures and greater stratification, which limits oxygen replenishment in surface waters (Webb, 2023).

- **PO_4:** While median PO_4 levels appear slightly higher in polar regions, non-polar regions exhibit a wider range and more high-value outliers. This suggests greater variability in phosphate availability across non-polar waters, which could be influenced by regional factors such as coastal upwelling zones, varying rates of phytoplankton uptake, or anthropogenic nutrient inputs in more temperate regions (Webb, 2023).

- **Chlorophyll-a:** The observed Chlorophyll-a levels are notably higher in polar regions, which aligns with known oceanographic patterns where sea ice retreat (seasonal ice melts) enhances light availability and nutrient upwelling, lead to phytoplankton blooms (NOAA, 2019).

- **Carbon.total:** The wider spread of carbon.total in polar regions reflects increased carbon uptake driven by sea ice loss. As sea ice melts, it exposes the ocean surface, allowing greater atmospheric carbon absorption. This process accelerates ocean acidification, occurring up to four times faster in the Arctic (Cai and Ouyang, 2022). The variability seen in the polar carbon values likely captures spatial and temporal differences in melt timing, however, these dynamics are largely absent and less pronounced in non-polar regions.

2.2.3 mOTU Filtering

Microbial abundance data derived from mOTUs are inherently sparse, meaning that most taxa are observed in only a limited subset of samples. To address this, we applied a two-step preprocessing procedure. First, for each sample, raw counts were normalized by the total library size to obtain relative abundances. Then, a cumulative contribution filtering was performed, retaining only those mOTUs that together accounted for up to 40% of the total relative abundance within each sample.

To determine the final set of mOTUs for analysis, we applied a union-based approach: the mOTUs retained in each sample were aggregated across samples via set union. This ensures

that any mOTU contributing substantially in at least one sample is preserved in the final dataset, even if it is absent or low in others. This approach reduces sparsity by selecting only the most abundant mOTUs within each sample, while the union step ensures their inclusion across the dataset.

Let $W = [w_{ij}] \in \mathbb{R}^{n \times q_0}$ denote the unfiltered mOTU count matrix, where w_{ij} is the raw count of mOTU j in sample i, n is the number of samples, and q_0 is the number of mOTUs before filtering.

Step 1: Relative Abundance Normalization

Relative abundance is the proportion of each mOTU's count compared to the total count in a sample (i.e., the row sum). This normalization makes samples comparable despite differences in sequencing depth. The total library size refers to the sum of all mOTU counts in a sample and represents the total number of reads for that sample.

We compute the relative abundance r_{ij} for each mOTU j in sample i, by normalizing with the total library size:

$$r_{ij} = \frac{w_{ij}}{\sum_{k=1}^{q_0} w_{ik}}$$

Let $R = [r_{ij}] \in \mathbb{R}^{n \times q_0}$ denote the resulting relative abundance matrix.

Step 2: Sample-wise Cumulative Contribution Filtering

For each sample $i \in \{1, \ldots, n\}$:

1. Rank the mOTUs in descending order of their relative abundances r_{ij}
2. Compute the cumulative sum of the ranked r_{ij} values:

$$c_{ij} = \sum_{k=1}^{j} r_{is_k}$$

   where f is a permutation of the indices such that $r_{if_1} \geq r_{if_2} \geq \cdots \geq r_{if_{q_0}}$.

3. Select the subset of mOTUs $\mathcal{J}^i \subseteq \{1, \ldots, q_0\}$ whose cumulative contribution does not exceed c:

$$\sum_{j \in \mathcal{J}^i} r_{ij} \leq c$$

   where $c = 0.4$ is the chosen cumulative contribution threshold.

If no mOTUs met the threshold due to sparsity, the highest-abundance mOTU is retained for that sample

Step 3: Union Across Samples

After identifying the subset of mOTUs $\mathcal{J}^i$ for each sample i, we construct the final set of

retained mOTUs by taking the union across all samples:

$$\mathcal{J}_{\text{final}} = \bigcup_{i=1}^{n} \mathcal{J}^i$$

Step 4: Construction of the Filtered mOTU matrix

We subset the original count matrix $W \in \mathbb{R}^{n \times q_0}$ to retain only the columns corresponding to the mOTUs in $\mathcal{J}_{\text{final}}$, resulting in a filtered matrix $W \in \mathbb{R}^{n \times q}$, where $q = |\mathcal{J}_{\text{final}}|$ is the number of retained mOTUs after filtering.

We began with a total of 18,700 mOTUs across 180 samples. After applying this filtering algorithm, q = 1077 mOTUs were retained. Between the new and the original data, 340 mOTUs are common to both datasets.

2.3 mOTU Abundance Profile

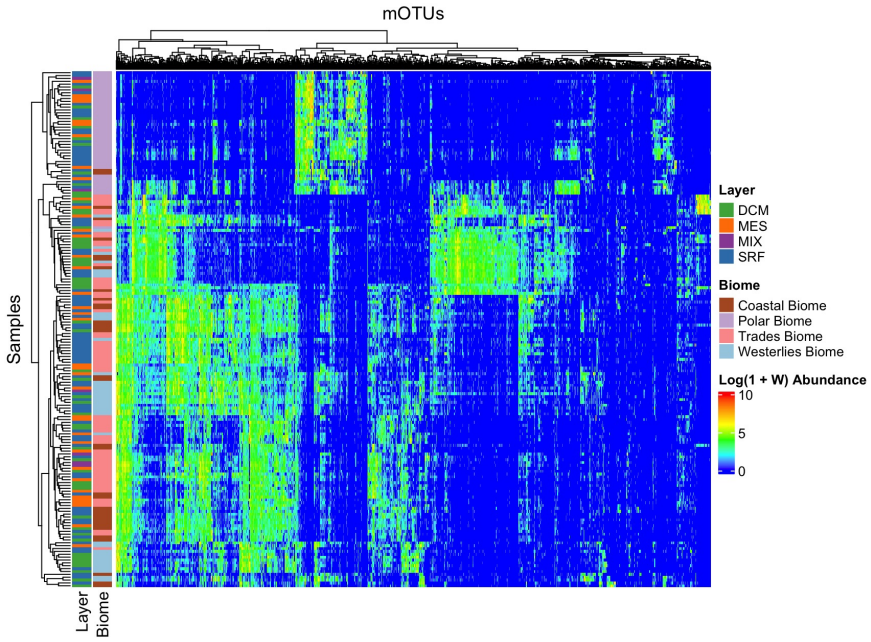

Figure 2.11: Heatmap of log abundance values $\log(1 + W)$ of $q = 1077$ mOTUs across $n = 180$ samples. Rows represent samples and columns represent mOTUs

Figure 2.11 shows the heatmap of the 1077 mOTUs across 180 samples. Sample are hierarchically clustered with average linkage and Bray–Curtis dissimilarity/distance. The color scale reflects the log(1 + W) abundance, where blue indicates absence or very low abundance while warmer colors reflect increasing abundance. The vertical bars for Layer and Biome are row annotations representing categorical metadata for each sample. Each color-coded bar

corresponds to the sample's Depth Layer and Biome.

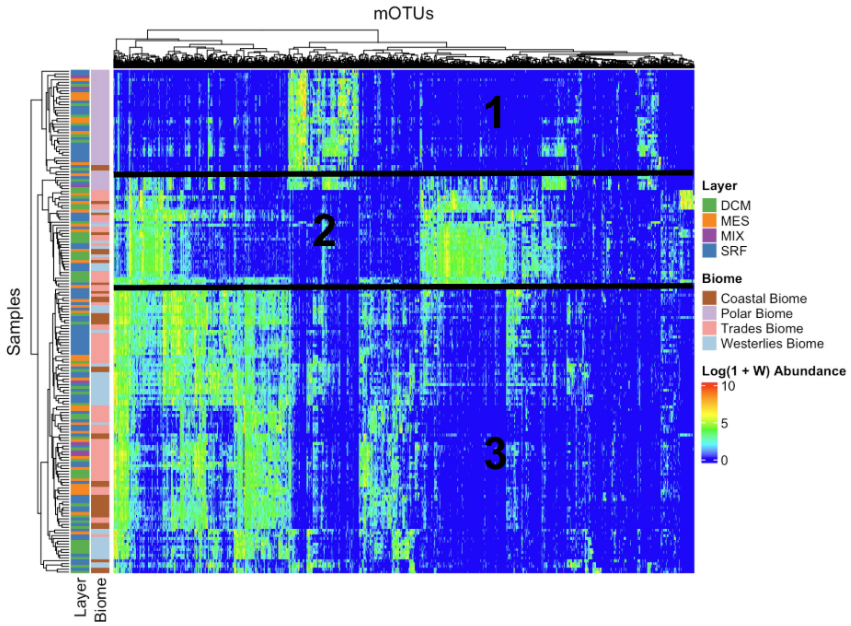

Figure 2.12: Heatmap of log abundance values $\log(1 + W)$ of $q = 1077$ mOTUs across $n = 180$ samples. Rows represent samples and columns represent mOTUs

Figure 2.12 shows three clear blocks of samples based on observed mOTU abundance with visible alignment to the Layer and Biome annotations.

- Block 1 is mostly composed of samples from the Polar Biome. Samples come from all four layers, though about half appear to be from the SRF layer. This group shows consistently low abundance overall, except for a distinct patch of very high abundance that appears only within this block. This sharp, localized region suggests a population of mOTUs that is uniquely abundant in the Polar Biome. While some Polar Biome samples are present in Block 3, they do not exhibit the same localized peak abundance.
- Block 2 includes mostly DCM layer samples and is dominated by the Trades biome, with a few samples from the Polar and Coastal biomes. This block shows the highest concentration of abundant mOTUs (yellow-green region), indicating that DCM layers in these biomes are perhaps associated with more active microbial communities.
- Compared to Blocks 1 and 2, Block 3 shows the widest spread of high abundance mOTUs across more samples. It includes a mix of all 4 layers and is dominated by the Trades and Westerlies Biome with some samples also from the Coastal Biome. This block stands out for its broad, consistent abundance across many samples and taxa.

Richness and Evenness

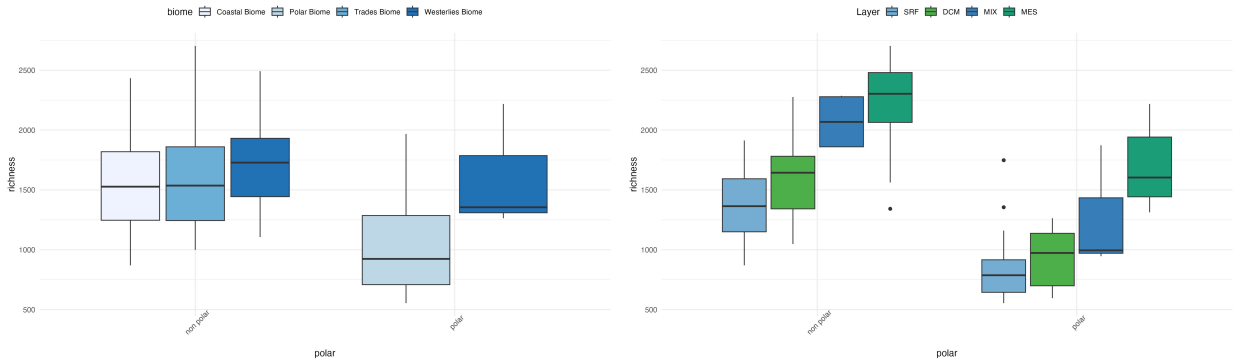

Figure 2.13: Left: (a) Richness by of polar and non-polar mOTUs across Biomes Right: (b) Richness by of polar and non-polar mOTUs across Depth Layers

Richness tells us the number of distinct mOTUs detected per sample and is a simple measure of microbial diversity, without accounting for relative abundance. The richness plots are based on the full set of 18,700 mOTUs prior to any filtering.

Figure 2.13 (a) shows that in non-polar regions, richness appears relatively consistent across biomes, with the Westerlies biome showing a slightly higher median richness. In contrast, richness in polar regions is comparatively lower. Interestingly, within the polar category, the Westerlies biome shows higher richness than the Polar biome itself. This contrast may reflect the influence of oceanic fronts and mixing zones at the boundary of polar and temperate waters, which can enhance nutrient cycling and biological activity. It also suggests that polar environments are not ecologically uniform—biome-specific oceanographic features such as current systems (e.g., the Antarctic Circumpolar Current), upwelling zones, and ice cover dynamics can create localized biodiversity hotspots. This highlights that polar samples are not homogeneous and that biome-specific differences persist even within the broader polar category.

Figure 2.13 (b) shows that in both polar and non-polar regions, mOTU richness increases with depth, from the surface (SRF) through the deep chlorophyll maximum (DCM) and mixed (MIX) layers, peaking in the mesopelagic (MES) zone. This consistent vertical pattern likely reflects increasing habitat complexity, reduced environmental fluctuations, and enhanced ocean particle flux (i.e., the transport of organic and inorganic material from the surface to the deep ocean) with depth. The MES zone, in particular, experiences less predation and accumulates organic matter from surface production, creating a resource-rich environment that supports a wide diversity of microbes—consistent with the observed high median richness in this layer across both polar and non-polar regions.

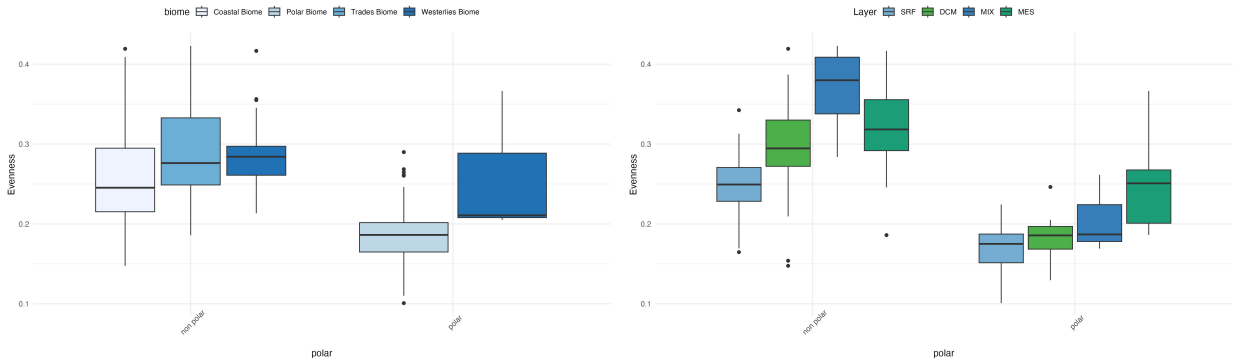

Figure 2.14: Left: (a) Evenness by of polar and non-polar mOTUs across Biomes
Right: (b)Evenness by of polar and non-polar mOTUs across Depth Layers

Evenness describes how equally individual mOTUs are represented in a community, reflecting whether a few taxa dominate or if abundances are more balanced across many taxa. Unlike richness, which counts how many mOTUs are present, evenness considers their relative abundances. On the y-axis, we have the evenness scale: values closer to 0.1–0.2 suggest that a few dominant taxa make up most of the abundance, while values closer to 0.4 indicate a more even distribution, where more taxa coexist without any single group dominating.

Figure 2.14 (a) shows that, similar to richness, evenness is generally higher in non-polar regions across biomes. Within the polar region, the Westerlies biome again shows higher evenness than the Polar biome. Tthe Polar biome is characterized by low evenness, indicating that a few mOTUs dominate these communities.

Figure 2.14 (b) shows that evenness increases with depth. In the polar region, evenness peaks in the MES layer, though the overall values remain low (median between 0.1–0.2), indicating dominance by a limited number of taxa. In contrast, in the non-polar region, the MIX layer has the highest median evenness, suggesting a more balanced microbial community structure at that depth.

mOTU Non-Metric Multidimensional Scaling (NMDS)

NMDS (Non-Metric Multidimensional Scaling) is an ordination technique that reduces complex mOTU data into a few dimensions based on rank dissimilarities, allowing us to visualize patterns in community composition. Unlike PCA, NMDS doesn't assume linear relationships or Euclidean distances, making it more suitable for mOTU count data such as ours.

To assess the influence of Depth Layer on microbial community structure in Polar and Non-polar samples, we performed an NMDS analysis followed by a PERMANOVA (Permutational Multivariate Analysis of Variance) test. NMDS reduced the complex community data into two axes (NMDS1 and NMDS2), which reflect how different the samples are from each other in terms of community composition. Each point in the NMDS plot represents a sample, and

the distance between points reflects the degree of difference in their microbial communities. Closer points are more similar. We then calculated a *stress* value which falls between 0 and 1 and reflects how well the reduced-dimensional ordination (e.g., 2D) represents the observed dissimilarities in the data. A lower stress value indicates a better fit. NMDS stress is commonly used to judge whether the plot accurately captures the structure of the original data without distortion Birdeau (2023).

Before calculating dissimilarities, we first normalized the mOTU count data to relative abundances. This means each mOTU count was divided by the total number of mOTUs in that sample, so the data represent proportions instead of raw counts. This step ensures that samples with different sequencing depths are placed on the same scale and can be compared fairly.

We then calculated pairwise dissimilarities between samples using the Bray-Curtis index. Bray-Curtis dissimilarity is a widely used ecological distance measure that takes into account both the presence/absence and abundance of species. It is calculated as:

$$BC_{ij} = \frac{\sum |x_{ik} - x_{jk}|}{\sum (x_{ik} + x_{jk})}$$

where x_{ik} and x_{jk} are the relative abundances of species k in samples i and j, respectively. The numerator sums the absolute differences in abundances, while the denominator sums the total abundances across both samples. This results in values ranging from 0 (identical communities) to 1 (completely different). Bray-Curtis is well-suited for ecological data because it is sensitive to differences in species abundance, and it does not assume a linear relationship.

We used PERMANOVA (Permutational Multivariate Analysis of Variance), a non-parametric statistical method, to assess whether microbial communities differ significantly across Depth Layers and between Polar and Non-polar regions. The test compares group centroids (i.e., multivariate means) within a dissimilarity matrix—in our case, based on Bray-Curtis distances—to evaluate whether observed differences exceed those expected by chance Ebner (2018).

The PERMANOVA test (Table 7) shows that differences in depth layer and whether a sample is from a polar or non-polar region explain about 36% of the variation in microbial community composition.

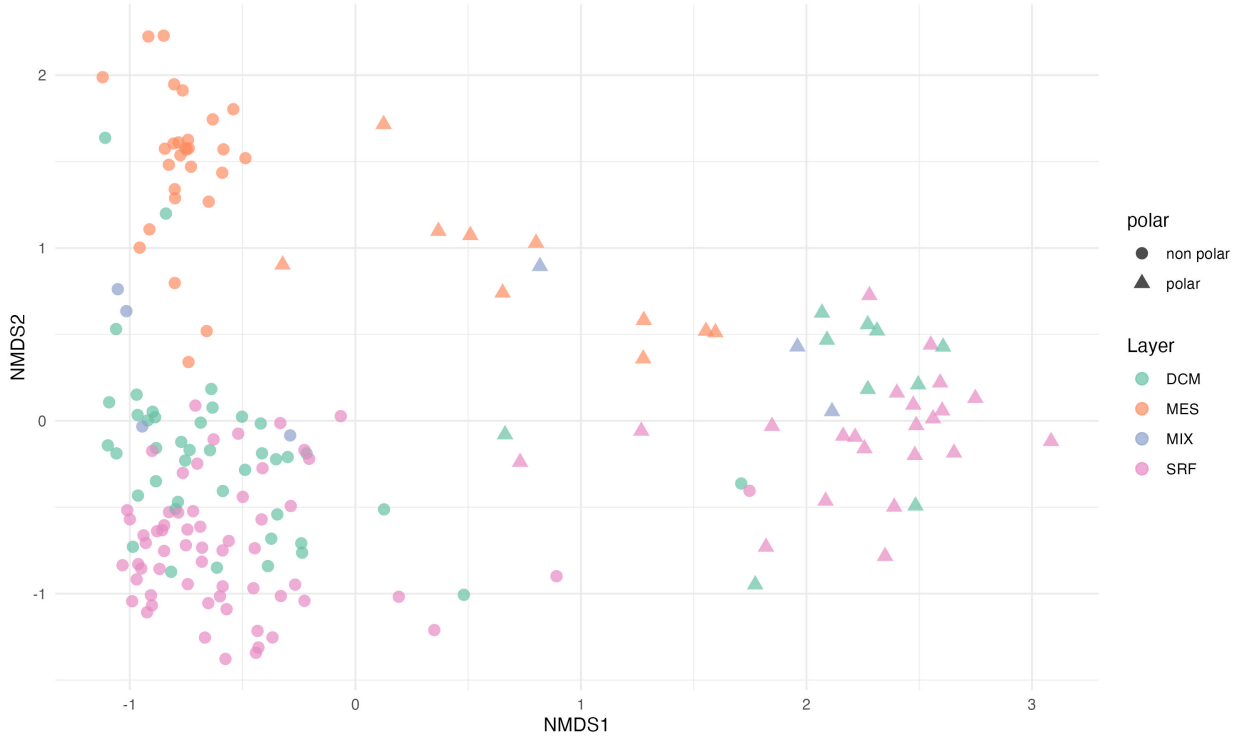

Figure 2.15: NMDS ordination of mOTU relative abundances, colored by Depth Layer and shaped by Polar/Non-polar

The NMDS plot helps us see how similar or different the samples are based on their community composition. The exact values or directions of the axes don't have a specific meaning — what matters is how close or far apart the points are from each other. Closer points imply more similar communities.

In the plot, we observe that Non-polar samples are generally clustered on the left, across all four Depth Layers, while Polar samples are more spread out but tend to cluster on the right. Among the Non-polar samples, those from the SRF and DCM layers cluster tightly together, suggesting higher similarity in microbial composition. A similar layering pattern is visible in the Polar samples, where SRF and DCM samples also group more closely, while MES samples appear more distinct. This suggests that depth layer influences community composition within both Polar and Non-polar regions.

In our analysis, the stress value was 0.083, which falls within the "good" range. This indicates that the NMDS plot provides a reliable representation of the dissimilarity structure in our mOTU data.

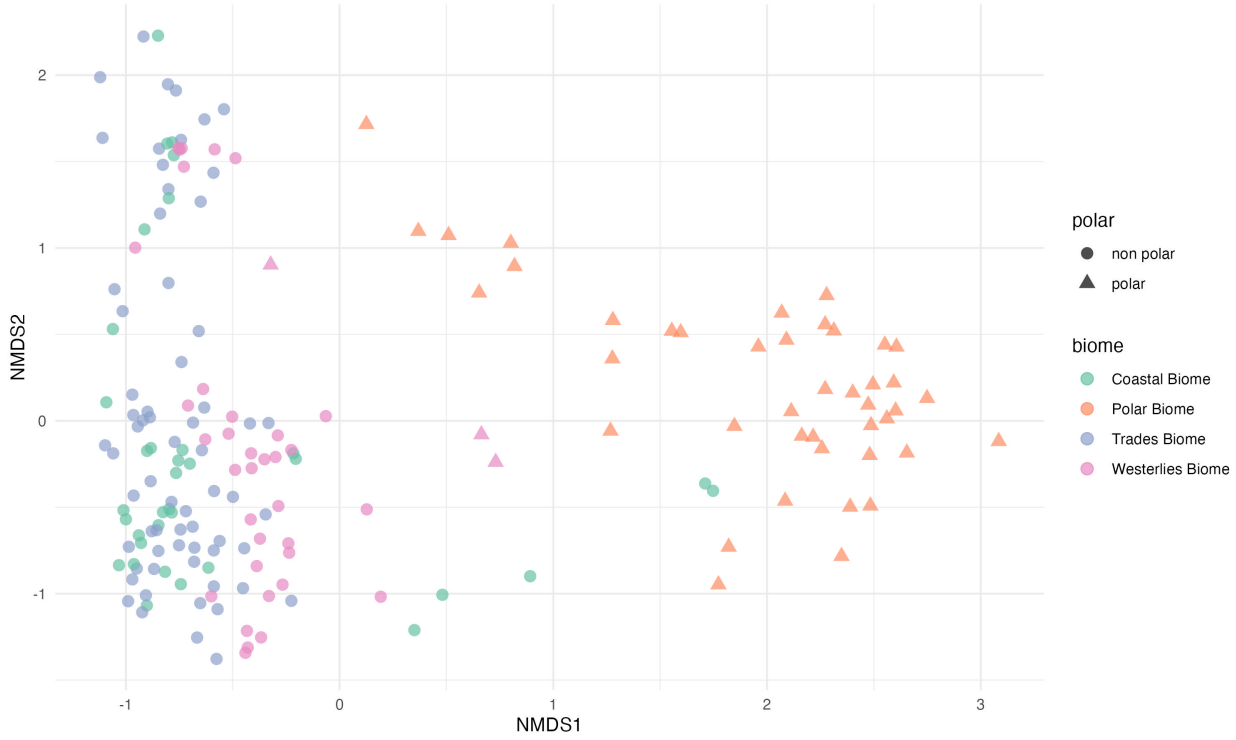

Figure 2.16: NMDS by Biome and Polar vs Non polar

We also did an NMDS for Biome and Polar vs Non-polar samples and it is evident from the plot that Polar Biome samples are not similar to samples from the other three Biomes.

The PERMANOVA test (Table 8) shows that differences in biome and whether a sample is from a polar or non-polar region explain about 24% of the variation in microbial community composition.

3 Methods and Modeling

mOTU data presents a unique statistical challenge due to over-dispersion and complex dependencies driven by both environmental conditions and species interactions. In this section, we present the probabilistic generative model VI-MIDAS introduced by (Mishra et al., 2024), along with a novel variant we developed—the no direct coupling model. Unlike VI-MIDAS, which incorporates environmental covariates through direct coupling, the no direct coupling model projects these covariates into the latent space, alongside spatiotemporal covariates and taxon-taxon interaction. We apply both models to the original and new datasets. The new dataset includes additional environmental covariates and a novel integration of satellite features, which are incorporated into both the direct coupling (original VI-MIDAS) and no direct coupling models. To establish a proof-of-concept, we have also reproduced the original VI-MIDAS results using the original dataset.

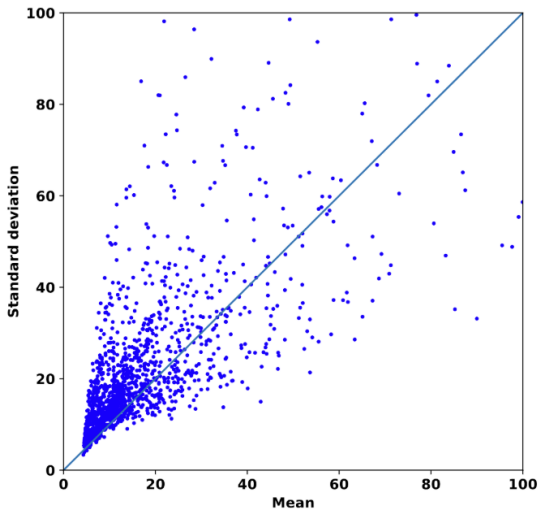

Figure 3.1: **Original Data**- Over-dispersion of mOTU abundance

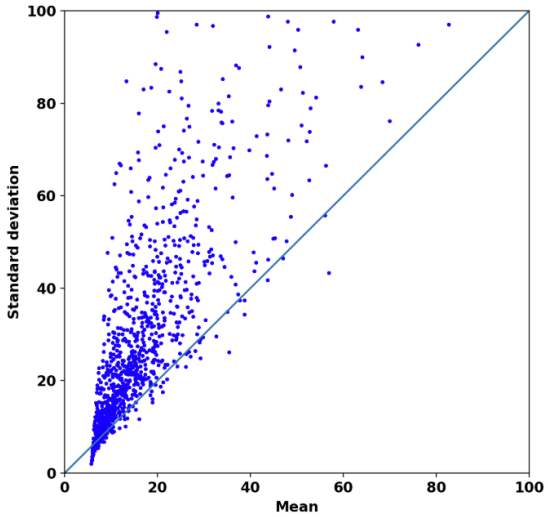

Figure 3.2: **New Data**- Over-dispersion of mOTU abundance

3.1 VI-MIDAS - Original Data

The goal of VI-MIDAS is to model the abundance profiles of q mOTUs (microbial taxa) where they are denoted as $W = [w_{ij}]^{n \times q} \in \mathbb{R}^{n \times q}$ (Mishra et al., 2024). For a single sample, the abundance profile is denoted as $w_i \in \mathbb{R}^q$, representing the observed counts of all q taxa in sample i. Figure 3.1 and Figure 3.2 show that the mOTU abundance for both datasets exhibits over-dispersion, with variance significantly exceeding the mean. VI-MIDAS aims to capture overdispersed data in a manner that enables interpretation of the effects of environmental and latent space covariates on microbial taxa, as well as the species-species interaction. This observation motivates the use of a Negative Binomial distribution in the generative model.

3.1.1 Generative Modeling

Let w_{ij} denote the observed count of microbial taxon j in sample i. These counts are modeled using a Negative Binomial distribution to account for overdispersion.

$$\begin{aligned} \mathrm{p}(w_{ij};\ \tau_j\mu_{ij},\ \phi_j) &= \mathrm{NB}(w_{ij};\ \tau_j\mu_{ij},\ \phi_j) \\ &= \binom{w_{ij}+\phi_j-1}{w_{ij}} \left(\frac{\tau_j\mu_{ij}}{\tau_j\mu_{ij}+\phi_j}\right)^{w_{ij}} \left(\frac{\phi_j}{\tau_j\mu_{ij}+\phi_j}\right)^{\phi_j} \end{aligned}$$

where,

- w_{ij}: the observed count of taxon j in sample i
- μ_{ij}: the mean parameter of taxon j in sample i; $\mu_{ij} \in \mathbb{R}^+$
- τ_j: taxa level scaling factor, specific to taxon j, that adjusts the mean μ_{ij} ; $\tau_j \in (0, 1)$
- ϕ_j: the dispersion parameter specific to taxon j; it captures overdispersion ; $\phi_j \in \mathbb{R}^+$

The shape parameter τ accounts for the disparity in abundance among the mOTUs. This model implies that the expected value and variance of w_{ij} are given by $\mathbb{E}(w_{ij}) = \tau_j\mu_{ij}$ and $\mathrm{Var}(w_{ij}) = \tau_j\mu_{ij} + \frac{(\tau_j\mu_{ij})^2}{\phi_j}$, respectively and $\mathrm{Var}(w_{ij}) > \mathbb{E}(w_{ij})$ (Mishra et al., 2024).
We model the mean parameter, μ_{ij} using a log-linear link function:

$$\log(\mu_{ij}) = \log(t_i) + \eta_{ij}$$

where:

- η_{ij}: models the covariates
- t_i : sample-specific scaling factor

The offset term $\log(t_i)$ is sample-specific and computed as a zero-aware geometric mean of the mOTU counts in each sample, where a small pseudocount d is added to avoid taking the logarithm of zero. This is conceptually similar to the denominator in CLR transformation (which centers each log-count by the geometric mean).

For sample i :

$$t_i = \exp\left(\frac{1}{q}\sum_{j=1}^{q}\log(w_{ij}+g)\right)$$

where,

- w_{ij}: the observed count of taxon j in sample i
- g: is a small constant added to each count to avoid taking the log of zero
- q : total number of taxa (mOTUs)

Thus, $log(t_i)$ is the offset that ensures that the model does not treat mOTUs with higher counts as evidence of higher biological abundance but rather, adjusts for different 'library sizes'. The use of a geometric mean allows us to stabilize the influence of extreme values.

This approach follows the procedure described in Section 3.1 of the supplementary material of (Mishra et al., 2024) which introduces the use of a zero-aware geometric mean as an offset.

3.1.2 Direct Coupling Model

The component η_{ij} encompasses - the direct coupling covariates (environmental covariates such as Temperature, Salinity, etc.) and the latent-space covariates - Spatio-temporal covariates (such as Depth Layer, etc.) and a species-species (taxon-taxon) interaction term.

$$\eta_{ij} = \underbrace{\eta_{ij}^{[E]}}_{\text{(environmental covariates)}} + \underbrace{(\eta_{ij}^{[P]} + \eta_{ij}^{[D]} + \eta_{ij}^{[S]})}_{\text{(spatio-temporal covariates)}} + \underbrace{\eta_{ij}^{[I]}}_{\text{(taxon-taxon interaction)}}$$

$$\eta_{ij} = \underbrace{\eta_{ij}^{[E]}}_{\text{direct coupling}} + \underbrace{(\eta_{ij}^{[P]} + \eta_{ij}^{[D]} + \eta_{ij}^{[S]} + \eta_{ij}^{[I]})}_{\substack{\text{latent space covariates} \\ (\beta)}}$$

Direct Coupling of Environmental Covariates $(\boldsymbol{\eta_{ij}})^{[\boldsymbol{E}]}$

We denote the matrix of p environmental covariates as $X = [x_1, \ldots, x_n]^\top = [x_{ij}] \in \mathbb{R}^{n \times p}$, where each row x_i corresponds to the covariates for sample i. The environmental covariates for the j^{th} taxa in the i^{th} sample is modeled as:

$$\eta_{ij}^{[E]} = x_i^\top \gamma_{\cdot j}$$

where $\gamma \in \mathbb{R}^{p \times q}$ is the matrix of regression coefficients, and $\gamma_{\cdot j}$ denotes the column of γ corresponding to taxon j.

$p = 9$ environmental covariates of the original data (as listed in Table 1) are modeled as a part of $(\eta_{ij})^{[E]}$.

Latent Space Coupling Covariates $(\boldsymbol{\eta_{ij}})^{[\boldsymbol{P}]}$, $(\boldsymbol{\eta_{ij}})^{[\boldsymbol{D}]}$, $(\boldsymbol{\eta_{ij}})^{[\boldsymbol{S}]}$, $(\boldsymbol{\eta_{ij}})^{[\boldsymbol{I}]}$

In addition to the direct environmental covariates, VI-MIDAS models the influence of spatiotemporal covariates through a latent space representation. While environmental covariates

such as salinity, oxygen, and temperature are modeled through direct coupling, spatiotemporal covariates and species-species interactions are captured via taxa-specific latent variables β.

In our analysis, these latent variables help uncover ecological patterns—particularly how taxa are organized by community role, such as mutualistic or competitive interactions.

We represent each of the q taxa using a shared set of latent variables arranged in a matrix $\beta \in \mathbb{R}^{k \times q}$, where each column $\beta_{\cdot j}$ is a vector of length k associated with taxon j. These latent vectors are learned from the data and reflect how each taxon responds to spatiotemporal covariates and to other taxa. The size factor k determines the dimensionality of these vectors and is an application-specific hyperparameter that controls the expressiveness of the latent space—i.e., how well the model can understand the underlying ecological relationships. This latent structure forms the basis for modeling both taxon–taxon interactions and the effects of Depth Layer (Figure 2.4), Province (Biome, Figure 2.4), and Season (Table 3).

Latent Space Coupling Covariates - Province (Biome), $(\boldsymbol{\eta_{ij}})^{[P]}$

We denote the indicator matrix of r=4 Provinces (Biome) - Polar, Westerlies, Trades, Coastal (Figure 2.4) of the n samples by $R = [r_1, \ldots, r_n]^\top \in \mathbb{R}^{n \times r}$, and connect it to the latent space β via the coefficient matrix $\alpha = [\alpha]^{r \times k} \in \mathbb{R}^{r \times k}$ (Mishra et al., 2024). $\eta_{ij}^{[P]}$ is given by:

$$\eta_{ij}^{[P]} = r_i \alpha \, \beta_{\cdot j}$$

Latent Space Coupling Covariates - Depth Layer, $(\boldsymbol{\eta_{ij}})^{[D]}$

We denote the indicator matrix of b=4 Depth Layer - SRF, DCM, MIX, MES (Figure 2.4) of the n samples by $D = [d_1, \ldots, d_n]^\top = [x_{ij}] \in \mathbb{R}^{n \times d}$, and connect it to the latent space β via the coefficient matrix $\delta = [\delta]^{b \times k} \in \mathbb{R}^{b \times k}$ (Mishra et al., 2024). $\eta_{ij}^{[D]}$ is given by:

$$\eta_{ij}^{[D]} = d_i \delta \, \beta_{\cdot j}$$

Latent Space Coupling Covariates - Season, $(\boldsymbol{\eta_{ij}})^{[S]}$

We denote the indicator matrix of m=4 Seasons - 1, 2, 3, 4 (column 1 of Table 3) of the n samples by $S = [s_1, \ldots, s_n]^\top \in \mathbb{R}^{n \times s}$, and connect it to the latent space β via the coefficient matrix $\vartheta = [\vartheta]^{m \times k} \in \mathbb{R}^{m \times k}$ (Mishra et al., 2024). $\eta_{ij}^{[S]}$ is given by:

$$\eta_{ij}^{[S]} = s_i \vartheta \, \beta_{\cdot j}$$

Projecting spatiotemporal covariates into the latent space using the coefficient matrices α, δ, and ϑ allows the model to quantify how much these covariates contribute to each taxon's abundance, independent of the effects already explained by the environmental covariates.

Latent Space Coupling Covariates - Taxon-Taxon Interaction, $(\boldsymbol{\eta_{ij}})^{[I]}$

Microbial abundances are influenced not only by environmental and spatiotemporal factors but also by interactions among the taxa themselves. These interactions can represent positive biological relationships (mutualism) or negative relationships (competitive). VI-MIDAS accounts for such influences through a latent interaction component, which models associations among the taxa that arise from their patterns of co-occurrence.

The interaction between taxon j and any taxon m is computed as the inner product $\rho_{\cdot j}^{\top}\beta_{\cdot m}$, where $\rho = [\rho]^{k\times q} \in \mathbb{R}^{k\times q}$ encodes a length-k latent vector for each of the q taxa (Mishra et al., 2024). The interaction term for the j^{th} taxon in the i^{th} sample is computed as:

$$\eta_{ij}^{[I]} = \begin{cases} 0, & w_{ij} = 0 \\ \frac{1}{a_i - 1}\boldsymbol{\rho}_{\cdot j}^{\mathrm{T}} \sum_{m\neq j} \mathbf{1}_{w_{im}\neq 0}\boldsymbol{\beta}_{\cdot m}, & w_{ij} \neq 0, \end{cases}$$

where $a_i = \sum_{m=1}^{q} 1(w_{im} \neq 0)$ denotes the total number of taxa present in the i^{th} sample. The interaction term $\rho^{\top}\beta$ is not symmetric, however, a symmetrized version $I = [I_{i,j}] \in \mathbb{R}^{q\times q}$ is defined as:

$$I_{i,j} = \frac{1}{2}\left(\rho_{\cdot j}^{\top}\beta_{\cdot m} + \rho_{\cdot m}^{\top}\beta_{\cdot j}\right)$$

To bring together the different covariates influencing microbial abundances, we decompose η_{ij} additively such that :

$$\eta_{ij} = \underbrace{x_i^{\top}\gamma_j}_{\eta_{ij}^{[E]}} + \underbrace{r_i\alpha\beta_{\cdot j}}_{\eta_{ij}^{[P]}} + \underbrace{d_i\delta\beta_{\cdot j}}_{\eta_{ij}^{[D]}} + \underbrace{s_i\vartheta\beta_{\cdot j}}_{\eta_{ij}^{[S]}} + \underbrace{I\beta_{\cdot j}}_{\eta_{ij}^{[I]}}$$

For the 'original data', η_{ij} for the **direct coupling** model is given by :

$$\eta_{ij} = x_i^{\top}\gamma_j + (r_i\alpha + d_i\delta + s_i\vartheta + I)\cdot\beta_{\cdot j} \tag{1}$$

3.1.3 No Direct Coupling Model

In the original VI-MIDAS framework, environmental covariates are incorporated into the model via direct coupling. The novelty we bring to VI-MIDAS is this modified version of the model where we implement " no direct coupling". In this framework, the environmental covariates are also projected into the latent space. This experiment served two principal objectives :

- Unifying the Model Structure: By placing all covariates (environmental, spatiotemporal, and interaction-driven) in the same latent representation, the model is encouraged to learn shared ecological patterns more holistically. It avoids assuming that environmental factors act independently of the latent ecological traits encoded in β.

- Testing Representational Power: The no direct coupling model allows us to assess whether latent representations alone are sufficient to explain mOTU abundance. If performance is comparable to the original VI-MIDAS model with direct coupling, it would suggest that latent embeddings can implicitly encode environmental effects—a valuable property in settings where model interpretability is secondary.

For the "original data", the **no direct coupling** model, redefines η_{ij} as :

$$\eta_{ij} = (x_i^\top \gamma_j + r_i \alpha + d_i \delta + s_i \vartheta + I) \cdot \beta_{\cdot j} \tag{2}$$

where, the environmental and spatiotemporal covariates and taxon-taxon interaction are integrated into the model via projection into the latent space.

3.2 VI-MIDAS - New Data

The new dataset consists of the environmental covariates as listed in Table 3 and the satellite derived covariates as listed in Table 4.

These enhancements provide an opportunity to evaluate VI-MIDAS under a richer covariate space and to assess whether the model's coupling mechanisms generalize effectively. As with the original dataset, we apply both the direct coupling ((Mishra et al., 2024)) and no direct coupling variants to the new data to assess the generalizability of both models.

3.2.1 Direct Coupling Model

In the direct coupling setup, environmental covariates are incorporated in the same manner as in the original dataset. All other components of the model, including spatiotemporal features and taxon–taxon interactions, remain unchanged.

Latent Space Coupling Covariates – satellite Covariates, $(\boldsymbol{\eta_{ij}})^{[\text{sat}]}$

Let $Z = [z_1, \ldots, z_n]^\top \in \mathbb{R}^{n \times h}$ denote the matrix of $h = 9$ satellite covariates. These covariates are incorporated into the model through the latent space , β, using a coefficient matrix $\pi = [\pi]^{h \times k} \in \mathbb{R}^{h \times k}$, leading to:

$$\eta_{ij}^{[\text{sat}]} = z_i \pi \, \beta_{\cdot j}$$

For the new dataset, η_{ij} in the **direct coupling** model takes the form:

$$\eta_{ij} = \eta_{ij}^{[E]} + \left(\eta_{ij}^{[P]} + \eta_{ij}^{[D]} + \eta_{ij}^{[S]}\right) + \eta_{ij}^{[I]} + \underbrace{\eta_{ij}^{[\text{sat}]}}_{\text{satellite covariates}}$$

$$\eta_{ij} = x_i^\top \gamma_j + (r_i\alpha + d_i\delta + s_i\vartheta + I + z_i\pi) \cdot \beta_{\cdot j}$$

3.2.2 No Direct Coupling Model

For the 'new data' in the **no direct coupling** case, the environmental component is moved into the latent space. Thus, η_{ij} is modeled as :

$$\eta_{ij} = (x_i^\top \gamma_j + r_i\alpha + d_i\delta + s_i\vartheta + I + z_i\pi) \cdot \beta_{\cdot j}$$

3.3 VI-MIDAS - Variational Inference

Irrespective of whether we are doing direct coupling or no direct coupling for the 'original data' or the 'new data', the modeling process from this point further will remain the same.

For the old data, we denote all the parameters as $\ell = (\alpha, \vartheta, \beta, \gamma, \rho, \tau, \phi, \delta)$ (Mishra et al., 2024). We place priors on all the aforementioned model parameters. We place a Laplace prior with parameters $(0,\ \lambda)$ on $\alpha, \delta, \beta, \gamma, \rho, \vartheta$. We place the following priors on $\tau_j \sim \text{Beta}(1, 1)$ and $\phi_j \sim \text{Inverse} - \text{Cauchy}(0, \nu)$ respectively (Mishra et al., 2024).

Similarly, for the new data, we denote all the parameters as $\ell = (\alpha, \vartheta, \beta, \gamma, \rho, \tau, \phi, \delta, \pi)$ (Mishra et al., 2024). We place priors on all the aforementioned model parameters. We place a Laplace prior with parameters $(0,\ \lambda)$ on $\alpha, \delta, \beta, \gamma, \rho, \vartheta, \pi$. We place the following priors on $\tau_j \sim \text{Beta}(1, 1)$ and $\phi_j \sim \text{Inverse} - \text{Cauchy}(0, \nu)$ respectively (Mishra et al., 2024).

Given the microbial abundance data, $\mathbf{W}$, the direct coupling covariates $\mathbf{X}$ and the latent space parameters ℓ, we integrate the generative model into a Bayesian framework where the posterior distribution is :

$$p(\ell; W, X, t) = \frac{p(W; \ell, X, t)\, p(\ell)}{p(W; X, t)}$$

where,

- $p(W; \ell, X, t) = \prod_{i,j} \text{p}(w_{ij};\ \tau_j\mu_{ij},\ \phi_j) = \prod_{i,j} \text{NB}(w_{ij};\ \tau_j\mu_{ij},\ \phi_j)$
- $p(\ell) = p(\alpha) \cdot p(\delta) \cdot p(\beta) \cdot p(\gamma) \cdot p(\rho) \cdot p(\phi) \cdot p(\tau) \cdot p(\vartheta) \cdot p(\pi)$
- $p(W; X, t) =$ marginal distribution

The marginal distribution $p(W; X, t)$ is obtained by integrating out the latent parameters:

$$p(W; X, t) = \int p(W; \ell, X, t)\, p(\ell)\, d\ell$$

This integral is intractable due to the high dimensionality of the latent parameters ℓ, and their non-linear interactions through the mean term μ_{ij}, which is embedded inside a log-link function. Additionally, the Negative Binomial likelihood makes it impossible to simplify the integral analytically. As a result, we cannot compute the exact posterior, which motivates the use of variational inference to approximate the posterior.

The Bayesian setting allows us to incorporate prior knowledge about parameters and to effectively model the latent variables. By placing priors on the parameters ℓ, we regularize the model and reduce the risk of overfitting—particularly important when dealing with sparse or overdispersed microbial abundance data.

Since computing the exact posterior distribution $p(\ell \mid W, X, t)$ is intractable and computationally infeasible, in order to approximate the posterior, we use *Variational Inference (VI)* (Blei et al., 2017). The key idea is to define a simpler family of distributions $\mathcal{Q}$, and then find the member $q(\ell; \nu) \in \mathcal{Q}$ that is closest to the true posterior. We then seek the member $q(\ell; \nu) \in \mathcal{Q}$ that minimizes the Kullback–Leibler (KL) divergence to the true posterior:

$$\min_{(\nu)} \mathrm{KL}(q(\ell; \nu) \,\|\, p(\ell \mid W, X, t))$$

To summarize, variational inference is a method that approximates complex posterior distributions using optimization - it finds the closest distribution by minimizing the KL-Divergence to the true posterior from within a family of candidate distributions. Minimizing the KL-Divergence is equivalent to maximizing the 'Evidence Lower Bound (ELBO)':

$$\mathcal{L}(\nu) = \mathbb{E}_{q(\ell;\nu)}[\log p(W, \ell; X, t)] - \mathbb{E}_{q(\ell;\nu)}[\log q(\ell; \nu)]$$

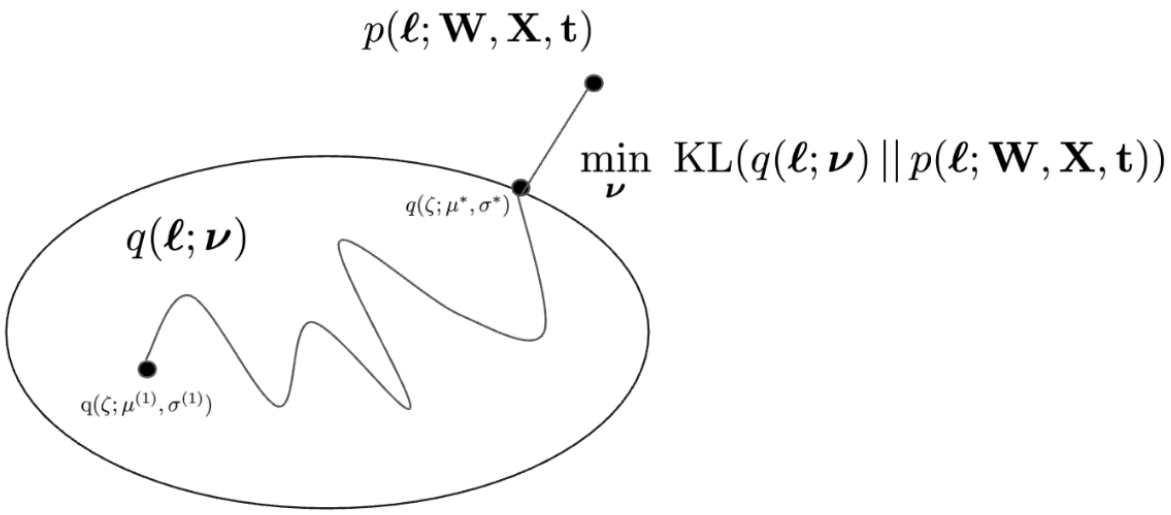

Figure 3.3: Variational Inference Schematic. Adapted from (Blei, 2017)

This allows us to use optimization to approximate the posterior without needing to evaluate the marginal. Maximizing the ELBO is equivalent to minimizing the KL divergence between the variational distribution $q(\ell;\nu)$ and the true posterior $p(\ell \mid W, X, t)$ up to a constant (Blei et al., 2017).

The ELBO can be rewritten as:

$$\mathcal{L}(\nu) = \mathbb{E}_{q(\ell;\nu)}[\log p(W, \ell; X, t)] - \mathbb{E}_{q(\ell;\nu)}[\log q(\ell;\nu)]$$

- The first term favors variational distributions $q(\ell;\nu)$ that place higher probability on those values of ℓ for which the model likelihood $p(W \mid \ell, X, t)$ is high
- The second term corresponds to the entropy of $q(\ell;\nu)$. Maximizing the ELBO therefore favors higher-entropy distributions, unless the likelihood sharply constrains the posterior

This balance reflects the Bayesian tradeoff between data fit and model complexity. It also ensures that the approximation is both informative and well-regularized (Blei et al., 2017).

In (Mishra et al., 2024), the variational posterior $q(\ell;\nu)$ is defined over a mean-field Gaussian variational distribution family in a transformed space. Detailed explanation of the optimization and estimation procedure, is provided in Section 3.2 of the Supplementary Material of Mishra et al. (2024).

The schematic diagram (adapted from (Mishra et al., 2024)) of VI-MIDAS for the direct coupling and no direct coupling model, please refer to Appendix A.

3.4 VI-MIDAS - Hyperparameter Tuning and LLPD

Hyperparameter Tuning

In order to estimate the model parameters, (Mishra et al., 2024) identify three key hyperparameters :

- k : dimension of the latent space, β
- λ : scale of the sparsity-inducing Laplace prior
- ν : scale of the inverse-Cauchy prior (dispersion)

These parameters are tuned via random search over 50 different hyperparameter initializations :

- $k \in \{10, 16, 30, 50, 80, 100, 150, 200, 500\}$

- $\lambda \in \{0.01, 3000\}$
- $\nu \in \{0.03125, 0.5\}$

We select the hyperparameter value which gives the best averaged LLPD. To evaluate the performance, the data is split into five folds, with 90% used for training and 10% for testing.

Log Pointwise Predictive Density (LLPD)

To evaluate model performance we use the (out-of-sample) LLPD:

$$\text{LLPD} = -\frac{1}{n}\sum_{i,j}\log\left(\text{NB}(w_{ij};\ \tau_j\mu_{ij},\ \phi_j)\right)$$

The LLPD for VI-MIDAS is the the negative average log-likelihood of the observed microbial abundance counts w_{ij}, under the Negative Binomial distribution with parameters $\tau_j\mu_{ij}$ (mean) and ϕ_j (dispersion). Higher LLPD values indicate better predictive performance (Mishra et al., 2024) and (Gelman et al., 2013).

4 Results

4.1 Hyperparameter Tuning

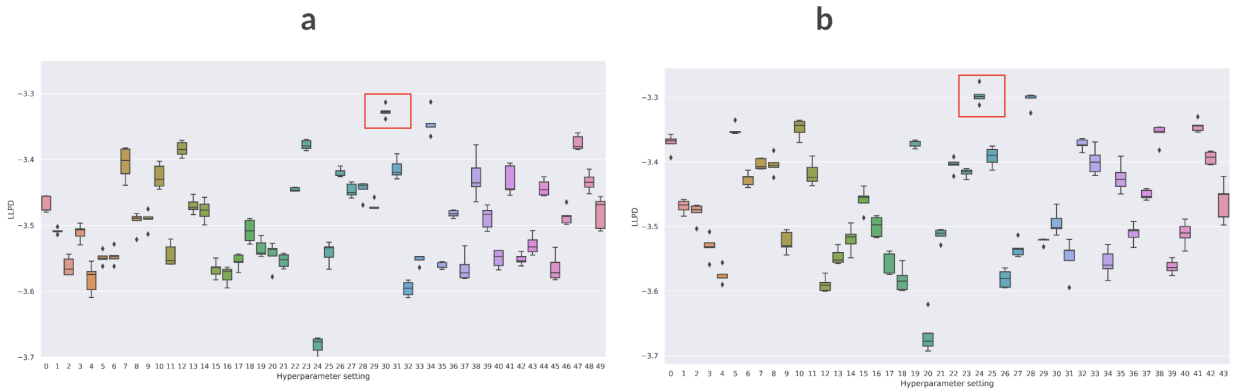

Figure 4.1: LLPD across different hyperparameter settings - red box indicates the best model
a. Direct Coupling - Original Data b. No Direct Coupling - Original Data

Figure 4.1 presents the distribution of out-of-sample log-likelihood posterior predictive density (LLPD) values across 50 different hyperparameter settings tested during model tuning. Each boxplot summarizes the out-of-sample LLPD values for the 50 models, separately, for the two model types: (a) Direct Coupling and (b) No Direct Coupling.

Higher LLPD values indicate better hyperparameter settings and, consequently, improved model performance. The best models—highlighted with red boxes—correspond to the highest LLPD values. These results demonstrate that certain combinations of the latent dimension k, λ, and ν yield significant improvements in model fit. This confirms that LLPD is a reliable criterion for hyperparameter selection. The best hyperparameter setting is discussed below.

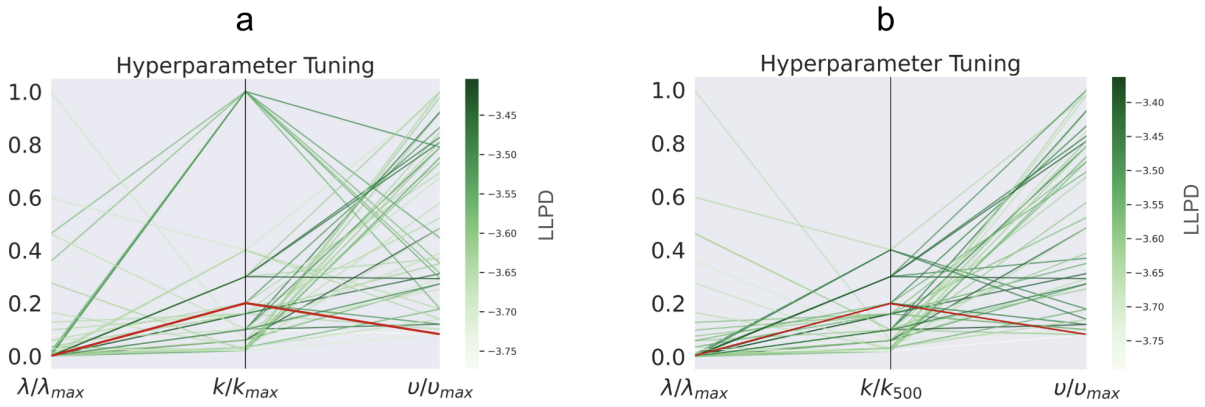

Figure 4.2: Parallel Coordinates Plot of Hyperparameter Tuning Performance - dark red line indicates the best hyperparameter setting
a. Direct Coupling - Original Data b. No Direct Coupling - Original Data

Figure 4.2 shows the parallel coordinates plots of hyperparameter tuning results using the original data, for the direct coupling and no direct coupling models. All the lines in green are hyperparameter combinations. So each line represents 1 hyperparameter combination. The plot shows 50 lines in total for 50 models. The color of each line indicates the performance of that setting based on the out-of-sample log-likelihood posterior predictive density (LLPD), with darker green corresponding to better model performance. Line in red indicates the best hyperparameter combination.

The three hyperparameters tuned were:

- k : corresponds to the dimension of the latent space, controls the complexity of the species interactions and abundance patterns
- λ : the scale of the sparsity-inducing Laplace prior
- ν : the scale of the inverse-Cauchy prior on the dispersion parameter ϕ, which regulates the allowed variance in species count predictions

The optimal hyperparameter settings (values) selected were:

- Direct Coupling : k = 100 , $\lambda = 0.02164$, $\nu = 0.050383$
- No Direct Coupling : k = 100 , $\lambda = 0.060596$, $\nu = 0.040816$

In subplot (a), representing the Direct Coupling scenario, the best-performing setting (red line) corresponds to intermediate values of k and low values of λ, suggesting that moderate latent dimensionality and stronger sparsity performed well. Notably, darker lines are clustered at lower values of λ, indicating that excessive regularization (higher λ) degraded

performance. Similarly, lower ν values tended to perform better, indicating that tighter variance control improved prediction.

In subplot (b), for the No Direct Coupling case, a similar trend is observed with respect to the different hyperparameter values. Again, moderate k values and relatively low λ and ν yielded better LLPD.

Due to the lack of computational resources, we were unable to tune the hyperparameters for the direct coupling and no direct coupling models which were run on the new data. Therefore, we used the best hyperparameter values of the original data on the respective models for the new data situation as well.

4.2 Original Data

Model comparison and ablation study

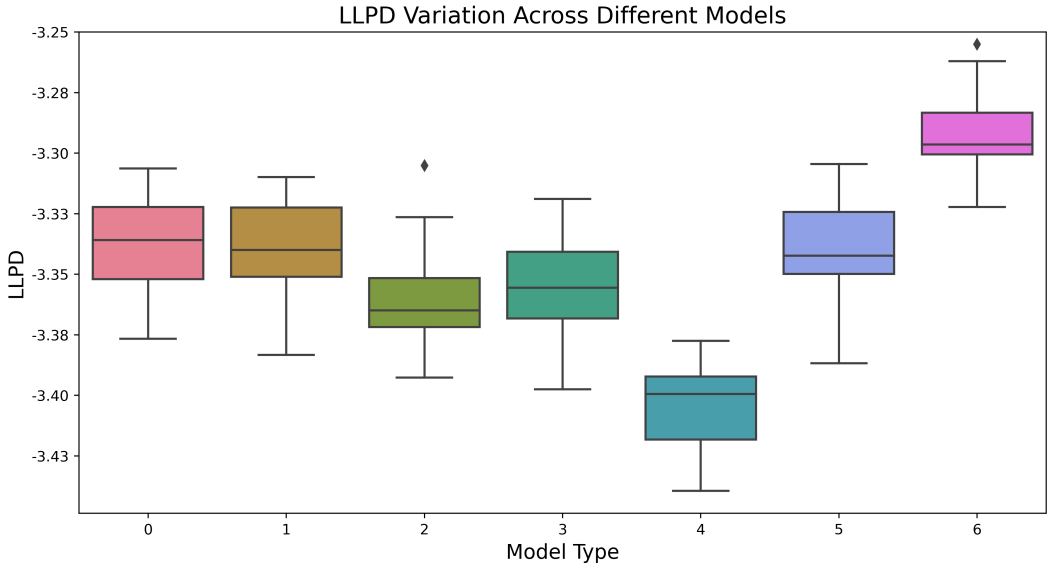

Figure 4.3: LLPD Distribution of the models. The Models (0-6) are explained in the table below

Model Type	Excluded Component
0	No component excluded
1	Spatial - Province (Biome)
2	Spatial - Depth Layer
3	Seasonal
4	Environmental
5	Species-Species Interaction
6	All covariates in the latent space (no direct coupling model)

Table 5: Model Types and their excluded components

Given the availability of both environmental and spatiotemporal data, a key objective of this study was to assess how each component contributed towards explaining the observed species (mOTU) abundance pattern. To this end, we evaluated seven model types. Model 0 corresponds to the full VI-MIDAS model with all covariates included via direct coupling. Model 6 is the no direct coupling model wherein the environmental covariates are moved into the latent space. Models 1 through 5 are component-excluded models, each omitting a specific covariate(s). For example - Model 2 excludes the 'Spatial - Depth Layer' component. Table 5 lists what each of the Models 0-6 are.

Figure 4.3 presents the out-of-sample LLPD distribution for the seven models. Model 6 - the 'no direct coupling' model - achieves the highest LLPD, indicating the best performance. In contrast, Model 4, which excludes environmental covariates, performs the worst. This result highlights the critical role that environmental covariates play in explaining abundance pattern.

Among the 'direct coupling' models (Models 0-5), Model 0 performed the best where no component was excluded. Model 1, which excludes the Spatial - Province (Biome) component performs only marginally worse than Model 0. This suggests that, relative to other components, it contributes lesser towards explaining the abundance pattern.

Species Abundance

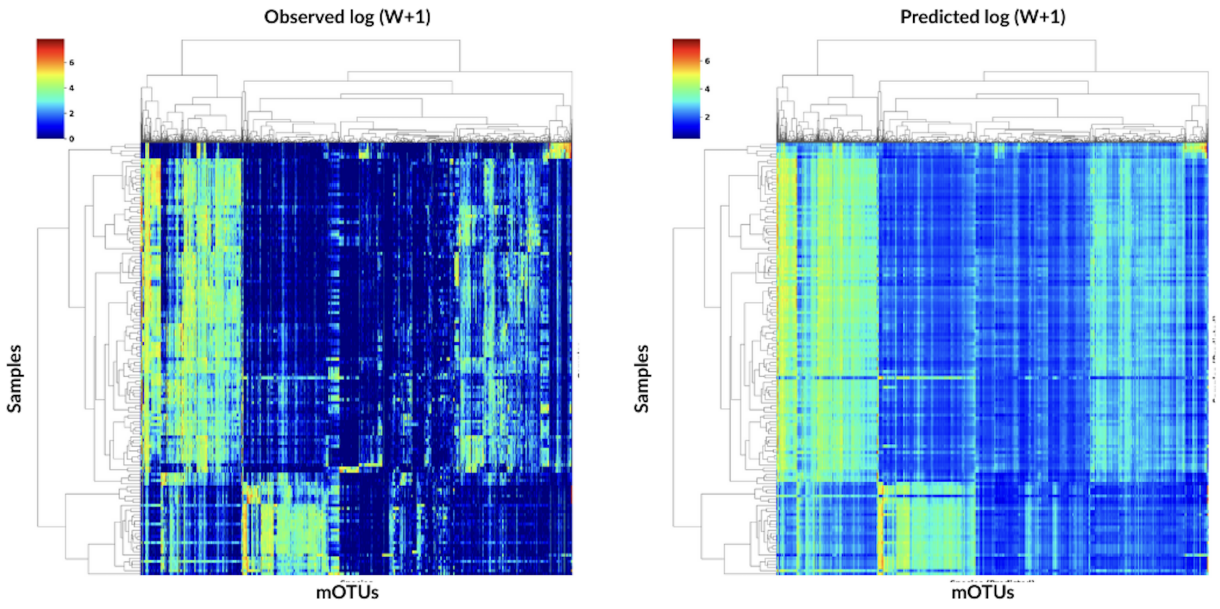

Figure 4.4: **Direct Coupling**- Abundance profile log(W+1) of 1378 mOTUs for n = 139 samples

Figure 4.4 compares the observed (left) and predicted (right) species abundance (log(W+1)) of 1378 mOTUs for 139 samples under the direct coupling model, using the hyperparameters

corresponding to the best model fit. Dark blue indicates absence of mOTUs, while warmer colors reflect increasing abundance.

The model broadly predicts regions of high abundance, demonstrating its ability to capture the structure of the mOTU abundance. However, it tends to overestimate presence in areas where the true abundance is near zero, as seen in the light green and blue regions of the predicted heatmap.

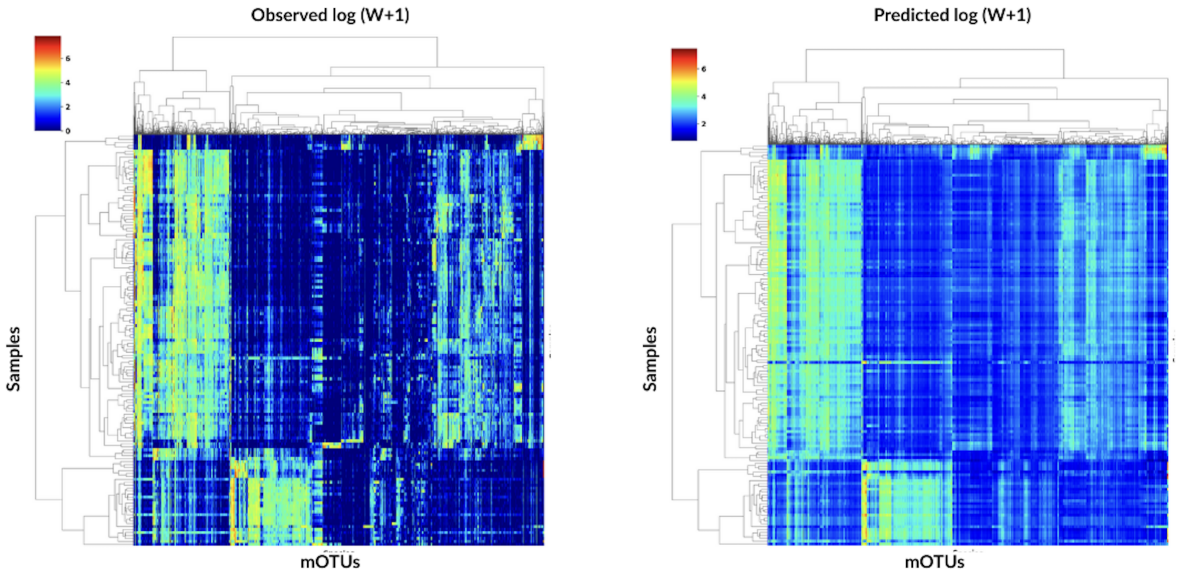

Figure 4.5: **No Direct Coupling**- Abundance profile log(W+1) of 1378 mOTUs for n = 139 samples

Figure 4.5 compares the observed (left) and predicted (right) species abundance (log(W+1)) of 1378 mOTUs for 139 samples under the no direct coupling model, using the hyperparameters corresponding to the best model fit.

The predicted heatmap captures the primary pattern of the high-abundance regions, it does a better job than the direct coupling model at predicting absence/low abundance pattern of the microbial abundance. The brightest (yellow) regions in the observed heatmap, which indicate very high mOTU abundance, appear less pronounced in the prediction, suggesting that the model underestimates peak abundances and compresses the dynamic range of high abundance.

Purely from the heatmaps, the direct coupling model does a better job at capturing the observed (presence) mOTU abundance structure, especially in the high-abundance (yellow/green) regions whereas, the no direct coupling model is better at predicting the absence/low abundance (dark blue).

Q-Q plot of Abundance profile of species

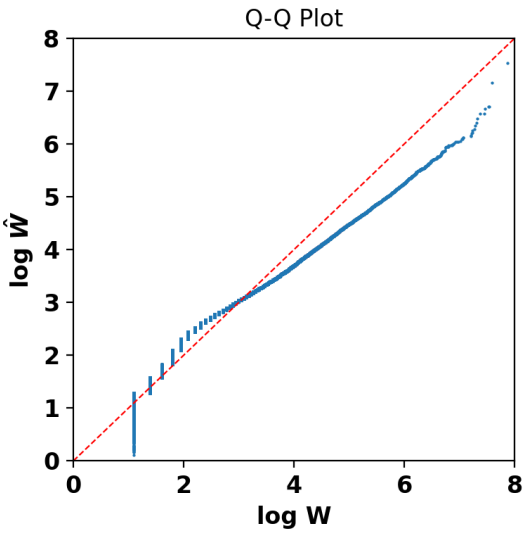

Figure 4.6: **Direct Coupling**

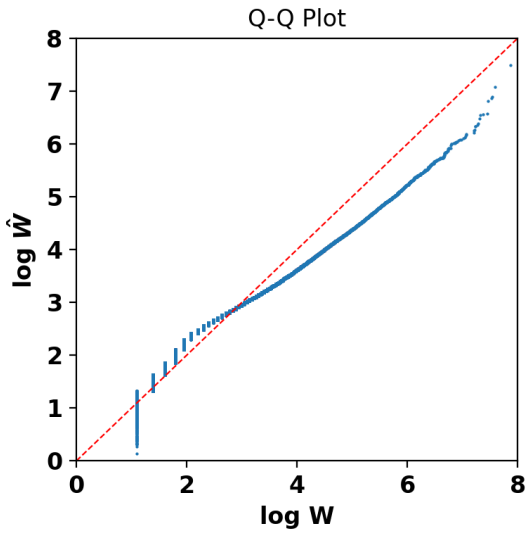

Figure 4.7: **No Direct Coupling**

Figure 4.6 and Figure 4.7 are the Q-Q plots comparing the log-transformed observed $\log W$ and predicted $\log \hat{W}$ mOTU abundances under the direct and no direct coupling models, respectively. In both cases, the predicted values closely follow the observed values up to moderate abundance levels. Both models show deviation from the red line (observed) in the upper tail, indicating that predicted values are systematically lower than the observed values at high abundance levels. The no direct coupling model shows slightly closer alignment with the observed data at higher abundance levels.

Environmental Covariates and ERCs

The Ecologically Relevant Classes (ERCs) are defined by Mishra et al. (2024), who manually curated these taxonomic groupings to enhance ecological interpretability.

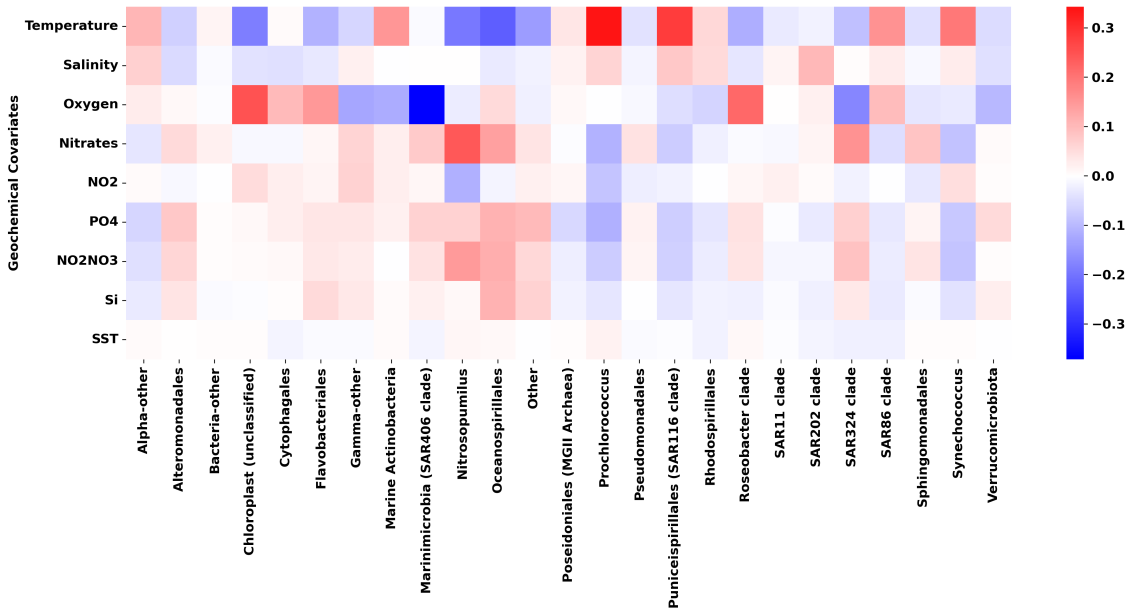

Figure 4.8: **Direct Coupling**- estimated effect of Environmental covariates on all ERCs

Figure 4.8 provides a summary of the estimated average effect side of the influence of the environmental covariates on the ERCs for the direct coupling model. The heatmap displays the VI-MIDAS model component, γ, with red indicating positive influence on abundance of the ERCs and blue indicating negative influence. Temperature and Oxygen appear as strong drivers of abundance for several ERCs.

Temperature is the primary positive factor influencing the abundance of Oceanospirillales, Nitrosopumilus and Chloroplast while it negatively influences the abundance of Prochlorococcus and SAR116 clade. It has a near zero influence on the abundance of SAR406 clade

Oxygen is the primary positive driver for the abundance of SAR406 clade and negative driver of abundance for Chloroplast and Roseobacter clade. It has a near zero influence on the abundance of Prochlorococcus and SAR11 clade.

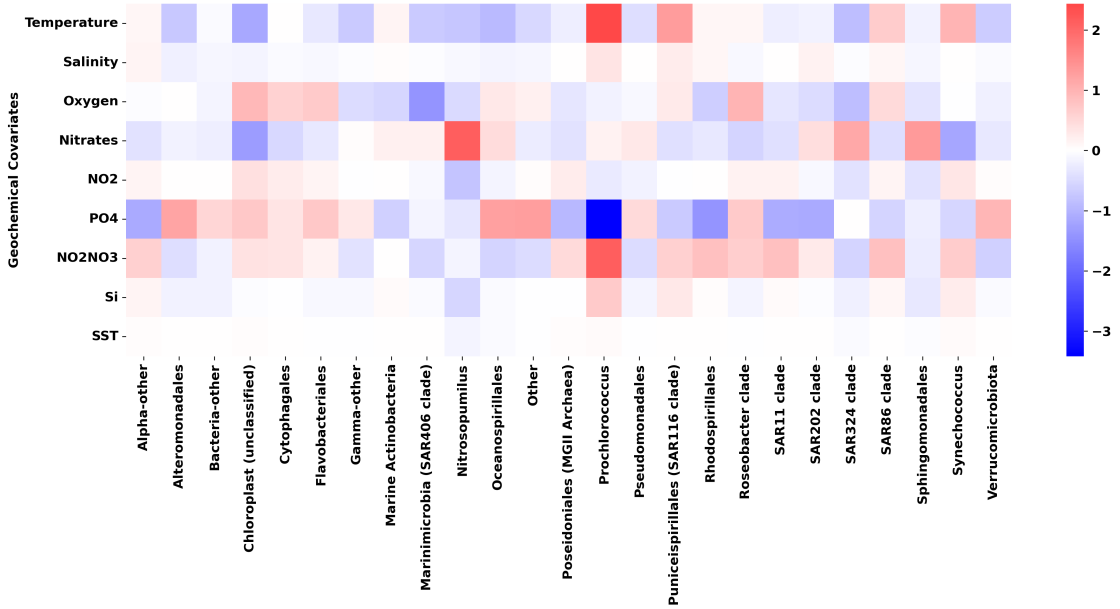

Figure 4.9: **No Direct Coupling** - estimated effect of Environmental Covariates on all ERCs

Figure 4.9 provides a summary of the estimated average effect side of the influence of the environmental covariates on the ERCs for the no direct coupling model. Here, PO4 and Temperature, NO_2NO_3 appear as strong drivers of abundance for several ERCs.

PO4 is the primary positive factor influencing the abundance of Prochlorococcus and Rhodospirillales. Temperature and NO_2NO_3 negatively influences the abundance of Prochlorococcus.

Positive and Negative Taxonomic Interactions

In VI-MIDAS, species-species interactions are captured through an interaction matrix computed as the inner product $\rho_{\cdot j}^{\top}\beta_{\cdot m}$. This term quantifies how much the abundance of taxon i is influenced by the latent ecological signature of taxon m.

A positive value of $\rho_{\cdot j}^{\top}\beta_{\cdot m}$ indicates that the presence of taxon m facilitates or supports taxon j, corresponding to a mutualistic or co-occurrence relationship. A negative value implies that taxon m inhibits or suppresses taxon j, suggesting a competitive interaction. That is, mutualistic interactions tend to support or co-occur with another species, while competitive interactions tend to reduce the abundance of another species.

Because the interaction is direction-specific (i.e., $\rho_{\cdot j}^{\top}\beta_{\cdot m} \neq \rho_{\cdot m}^{\top}\beta_{\cdot j}$), the matrix is symmetrized by averaging the two directed terms.

Simply put, competitive interactions (negative) tend to negatively influence the abundance of another species and mutualistic interactions (positive) imply co-occurrence. In the plots, the upper triangle shows aggregated negative (competitive) interactions and the lower triangle shows aggregated positive (mutualistic) interactions Mishra et al. (2024).

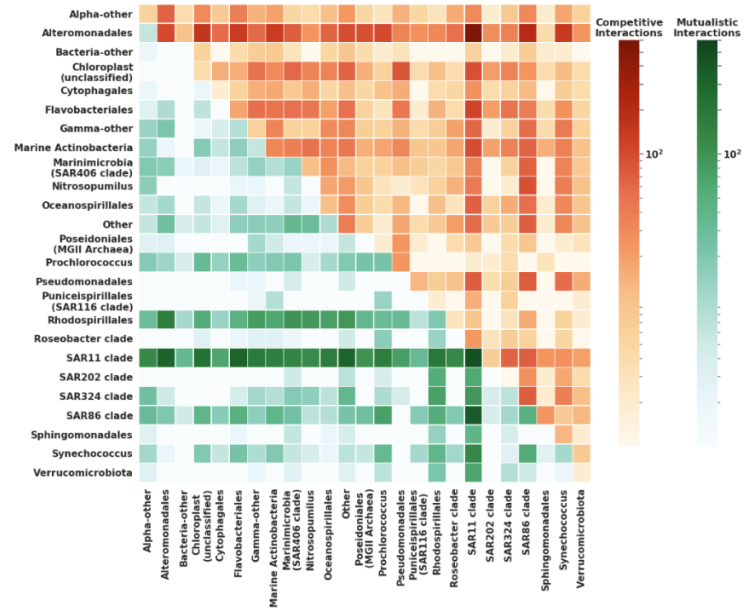

Figure 4.10: **Direct Coupling** - Taxonomic Interactions

Alteromonadales has negative interactions with almost all other ERCs, with the strongest negative interaction observed with the SAR11 clade. In contrast, SAR11, SAR86, and Rhodospirillales form positive interactions with the majority of the ERCs. SAR86 clade, in particular, exhibits broadly mutualistic patterns, especially with the SAR11 clade, suggesting potential ecological cooperation or shared environmental niches.

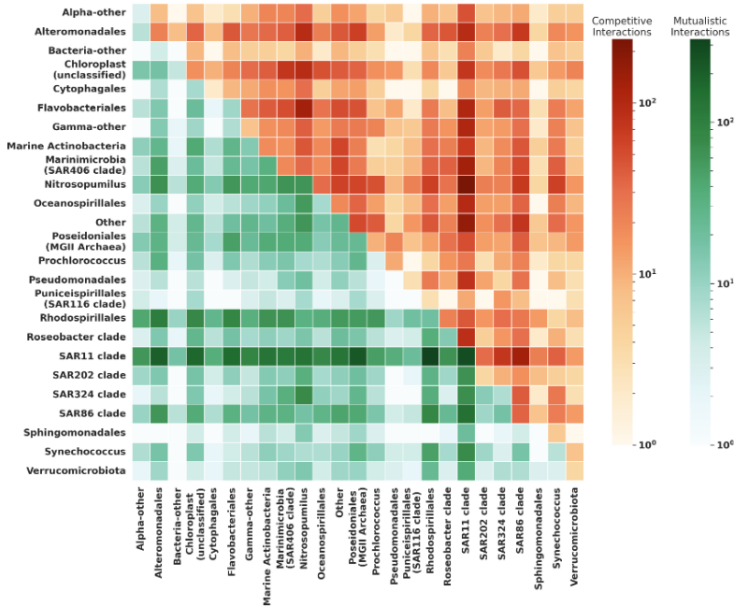

Figure 4.11: **No Direct Coupling** - Taxonomic Interactions

Compared to the direct coupling model, the interactions under the no direct coupling model exhibits sharper and more prominent interaction patterns. The positive and negative interactions appear more saturated and distinct, making the underlying structure of taxon–taxon effects more evident.

The SAR11 clade exhibits a strong mutualistic interaction with Rhodospirillales, as indicated by the deep green cell in the interaction matrix, suggesting that the presence of Rhodospirillales strongly supports SAR11 abundance. In contrast, the interaction between SAR11 and Alteromonadales is strongly negative, reflecting a competitive relationship where Alteromonadales likely suppresses SAR11 abundance. Notably, the interaction is asymmetric: while SAR11 benefits significantly from Rhodospirillales, the reverse effect is only moderately positive.

Latent Representation, β

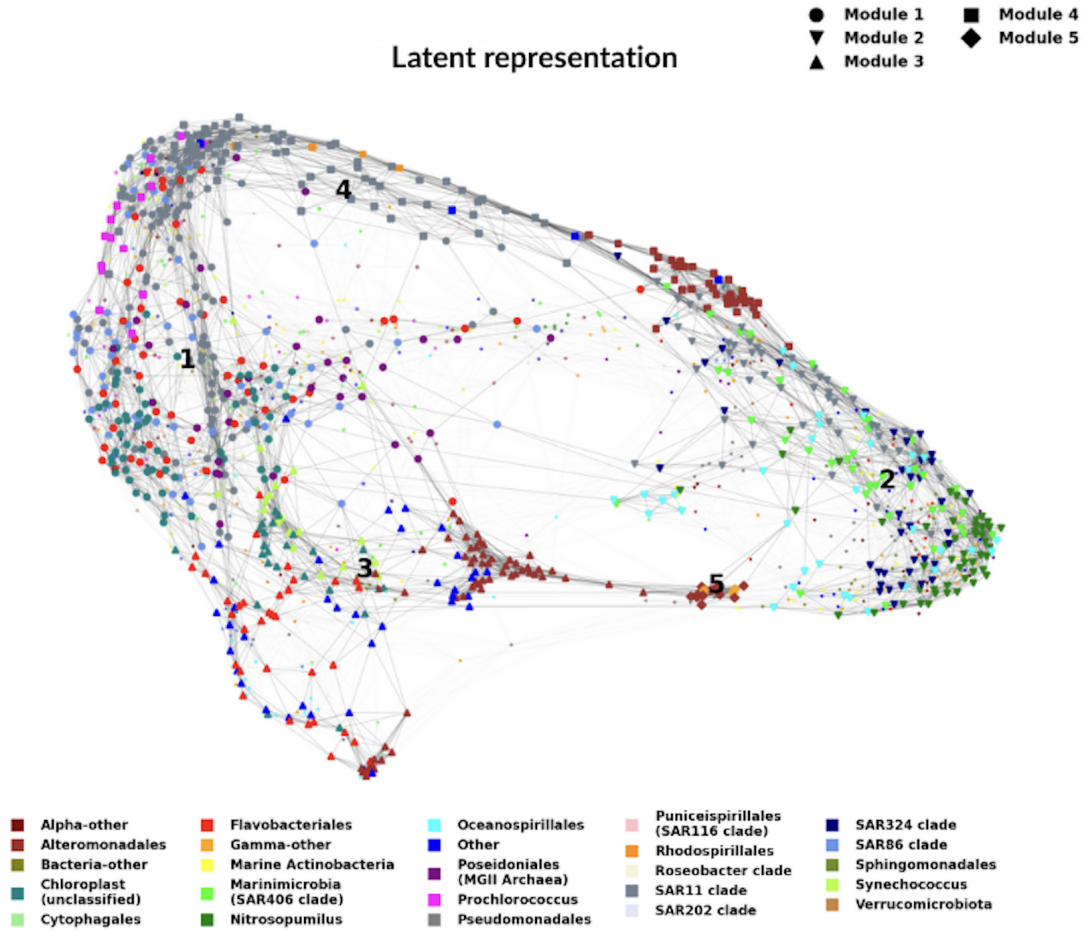

Figure 4.12: Low-dimensional embedding of the latent representation β using a k-nearest neighbor ($k_{\text{nn}} = 10$). Modularity analysis revealed 5 distinct graph modules.

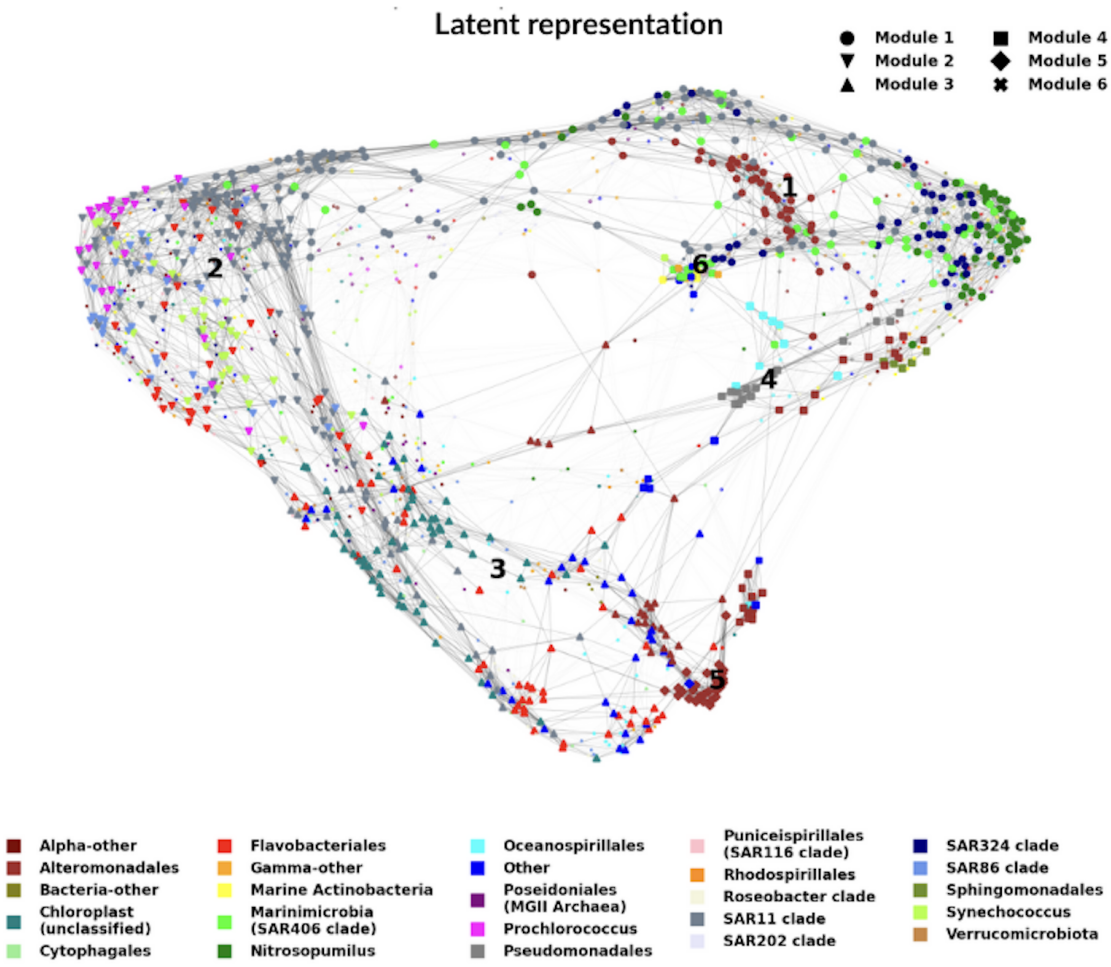

Figure 4.13: Low-dimensional embedding of the latent representation β using a k-nearest neighbor, ($k_{\text{nn}} = 10$). Modularity analysis revealed 6 distinct graph modules.

4.3 New Data

Model comparison and ablation Study

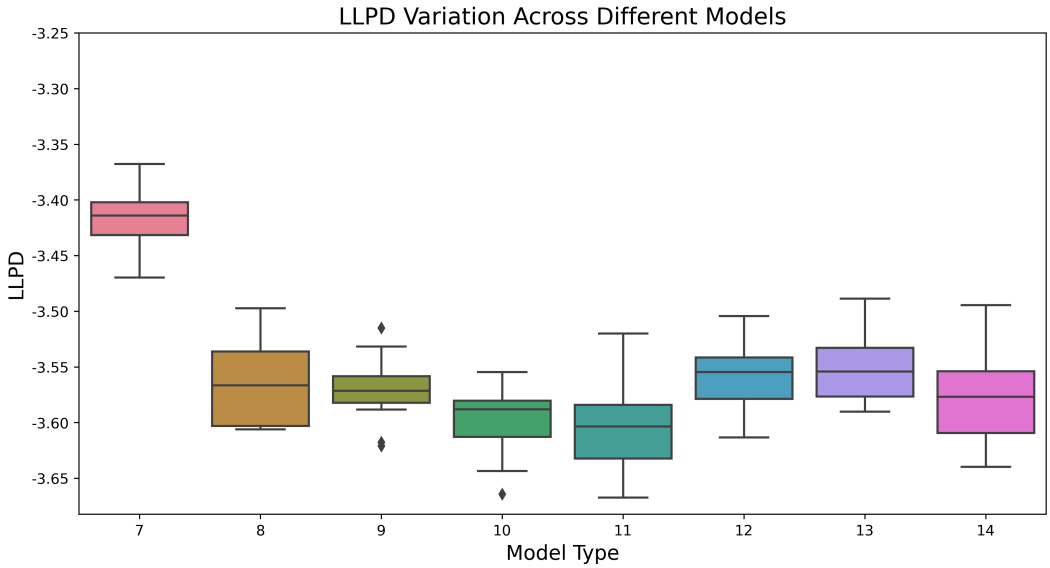

Figure 4.14: LLPD Distribution of the models. The Models (7-14) are explained in the table below

Model Type	Excluded Component
7	All covariates in the latent space (no direct coupling model)
8	No component excluded
9	Species-Species Interaction
10	Environmental
11	Spatial - Province (Biome)
12	Spatial - Depth Layer
13	Seasonal
14	Satellite

Table 6: Model Types and their excluded components

In the case of the 'new data', we also had the satellite covariates. Therefore, we evaluated eight model types. Model 8 corresponds to the full VI-MIDAS model with all covariates - the direct coupling model where no covariates are excluded. Model 7 is the no direct coupling model. Models 9 through 14 are the component-excluded models, each omitting a specific covariate(s). For example - Model 14 excludes the 'Satellite' component. Table 6 lists what each of the Models 0-6 are.

Figure 4.14 presents the out-of-sample LLPD distribution for the eight models. Similar to the original data, Model 7 - the 'no direct coupling' model - achieves the highest LLPD, indicating the best performance.

Among the 'direct coupling' models (Models 8-14), Models 12 and 13 performed the best and have almost the same out-of-sample LLPD values. This is justified since microbial

community structure is known to vary significantly with light availability, stratification, and nutrient fluxes, all of which change across depths and seasons. However, in order to keep the study comparable in terms of model performance, we chose Model 8 as the best since it is the full model where no component is excluded.

The out-of-sample LLPD is lowest for Model 11, which excludes the Spatial - Province (Biome) component, suggesting that this covariate contributes less substantially to explaining abundance patterns in the new dataset. Notably, Model 9, which excludes the Species-Species Interaction component, performs only marginally worse than the full model (Model 8), indicating that direct environmental and spatiotemporal factors may play a more dominant role than interspecies associations in shaping microbial community structure.

Species Abundance

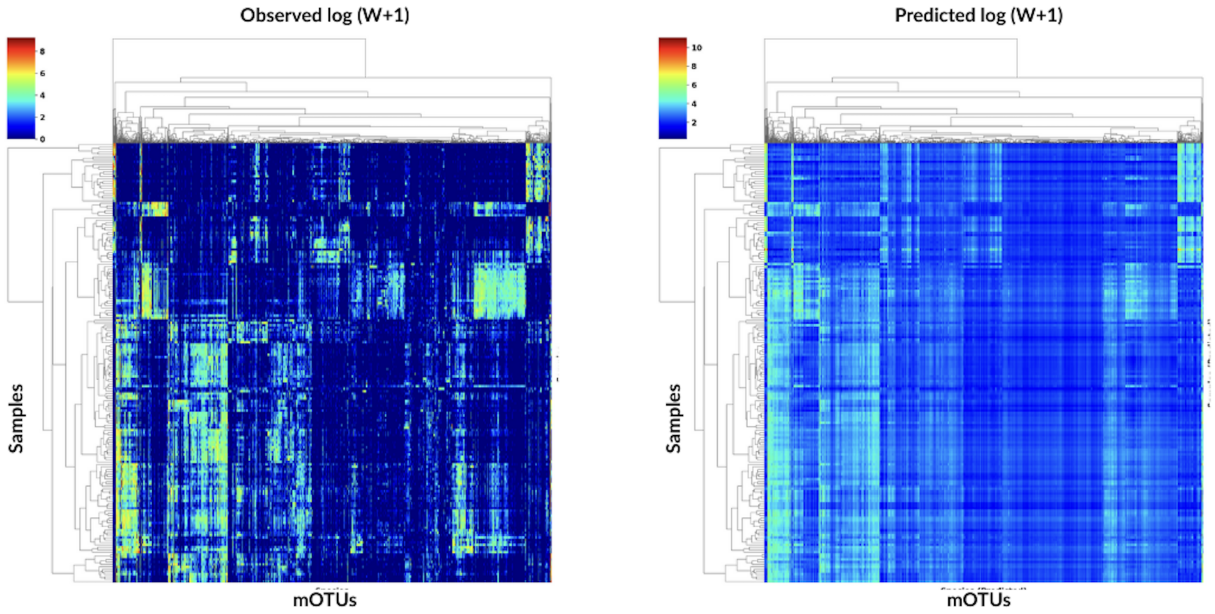

Figure 4.15: **Direct Coupling** - Abundance profile log (W+1) of 1077 mOTUs for n = 180 samples

Figure 4.15 compares the observed (left) and predicted (right) species abundance (log(W+1)) of 1077 mOTUs for 180 samples under the direct coupling model. Dark blue indicates absence of mOTUs, while warmer colors reflect increasing abundance.

The model partially predicts the abundance structure, capturing broader abundance patterns. However, it fails to predict very high abundance and, in some cases, does not predict abundance at all. It tends to overestimate presence in areas where the true abundance is near zero, as seen in the light green and blue regions of the predicted heatmap. Additionally, it predicts presence in regions where the observed abundance is zero.

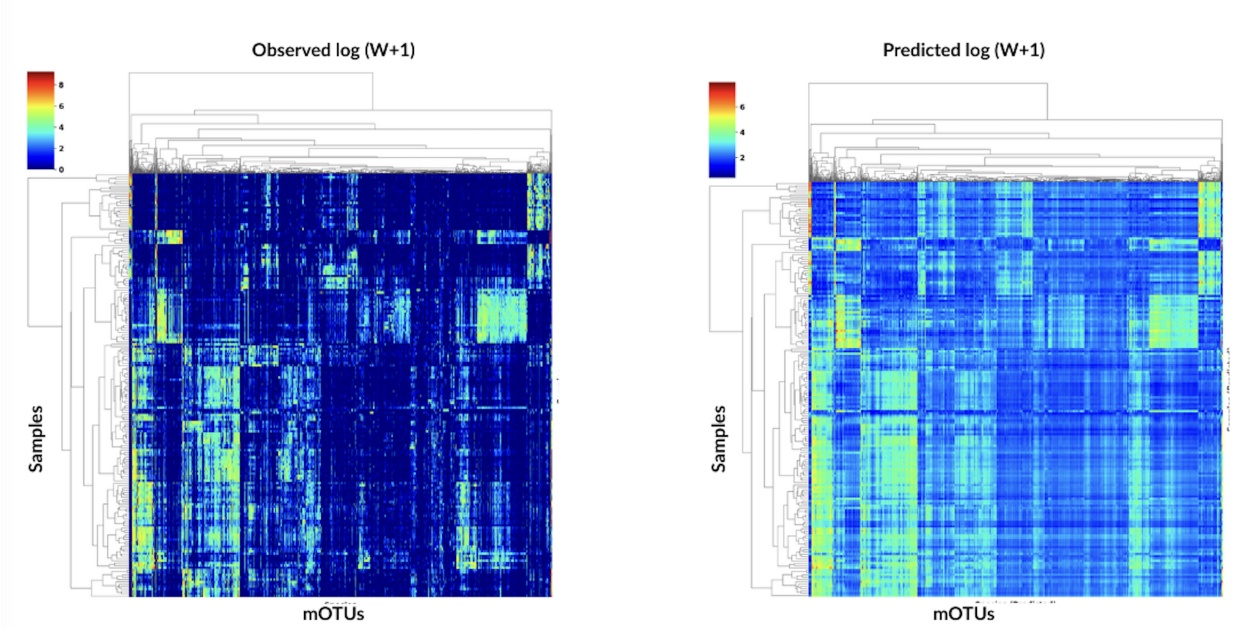

Figure 4.16: **No Direct Coupling** - Abundance profile log (W+1) of 1077 mOTUs for n = 180 samples

Figure 4.15 compares the observed (left) and predicted (right) species abundance (log(W+1)) of 1378 mOTUs for 139 samples under the no direct coupling model.

The model is able to broadly predict regions of high abundance, demonstrating its ability to capture the overall structure of mOTU abundance. The predicted heatmap captures the primary pattern of the high abundance more effectively than the direct coupling model. The brightest (yellow) regions in the observed heatmap, which indicate peak mOTU abundance, appear more pronounced in the predicted heatmap, suggesting that the model is able to predict very high abundance. However, it tends to overestimate presence in areas where the true abundance is near zero, as seen in the light green and blue regions of the predicted heatmap.

Purely from the heatmaps of the two models, the no direct coupling model appears to perform better, as it more accurately predicts high and very high abundance (yellow areas) compared to the direct coupling model. The structure of the prediction more closely resembles the observed mOTU abundance, showing a better fit to the abundance patterns across samples. While both models overpredict low-abundance regions (light blue/green where the observed is dark blue), this does not outweigh the better performance of the no direct coupling model in capturing high abundance.

Q-Q plot of Abundance profile of species

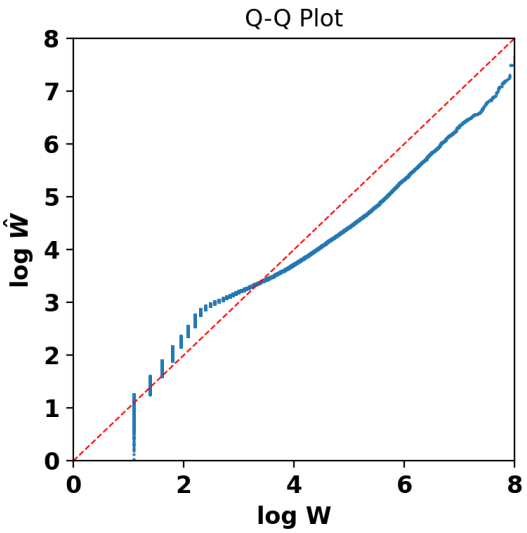

Figure 4.17: **Direct Coupling**

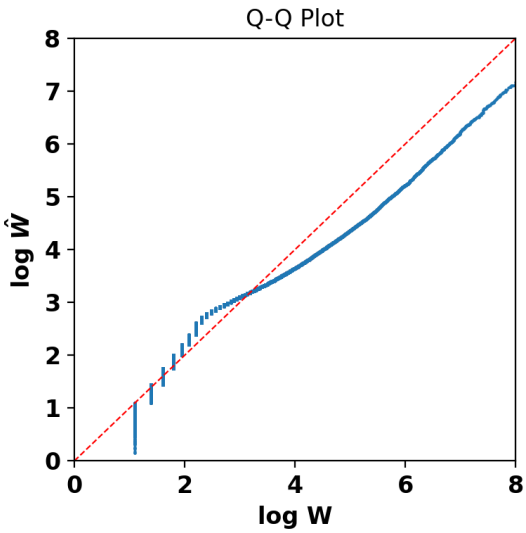

Figure 4.18: **No Direct Coupling**

Figure 4.17 and Figure 4.18 are the Q-Q plots comparing the log-transformed observed $\log W$ and predicted $\log \hat{W}$ mOTU abundances under the direct and no direct coupling models, respectively. In both models, predicted values closely track the observed distribution up to moderate abundance levels. Both models show deviation from the red line (observed) in the upper tail, indicating that predicted values are systematically lower than the observed values at high abundance levels. The direct coupling model shows slightly closer alignment with the observed data at higher abundance levels.

Environmental Covariates and ERCs

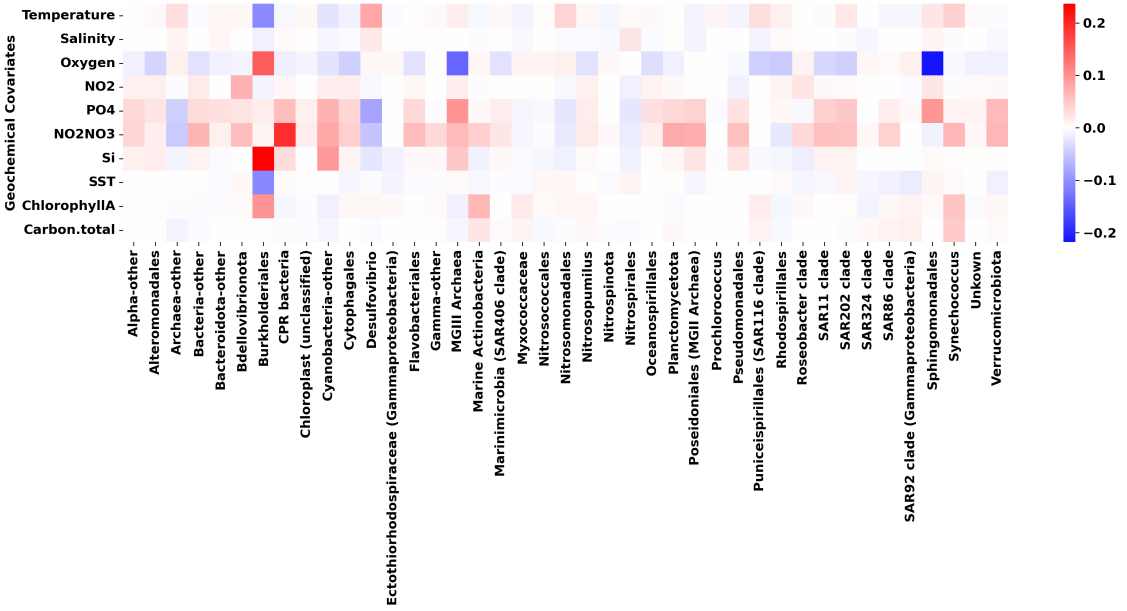

Figure 4.19: **Direct Coupling**- estimated effect of Environmental covariates on all ERCs

Figure 4.19 provides a summary of the estimated average effect side of the influence of the environmental covariates on the ERCs for the direct coupling model. The heatmap displays the VI-MIDAS model component, γ, with red indicating positive influence on abundance of the ERCs and blue indicating negative influence. Oxygen and NO_2NO_3 appear as strong drivers of abundance for several ERCs.

Oxygen is the primary negative driver for the abundance of Sphingomonadales and MClll Archaea and positive driver of abundance for Burkholderiales. It has a near zero influence on the abundance of Nitrospinota, Prochlorococcus and Poseidoniales.

NO_2NO_3 and Si are the primary negative factor influencing the abundance of CPR bacteria and Burkholderiales respectively. An interesting obersevation is that the abundance of Burkholderiales is strongly negatively influenced by Si, Oxygen and moderately negative influence by ChlorophyllA and positively influenced by Temperature and SST. NO_2NO_3 has near zero influence on the abundance of Prochlorococcus and SAR116 clade.

The the heatmap of estimated effect of Environmental covaiates on ERCs where the ERCs are the same as those in the original data can be found in Appendix A.

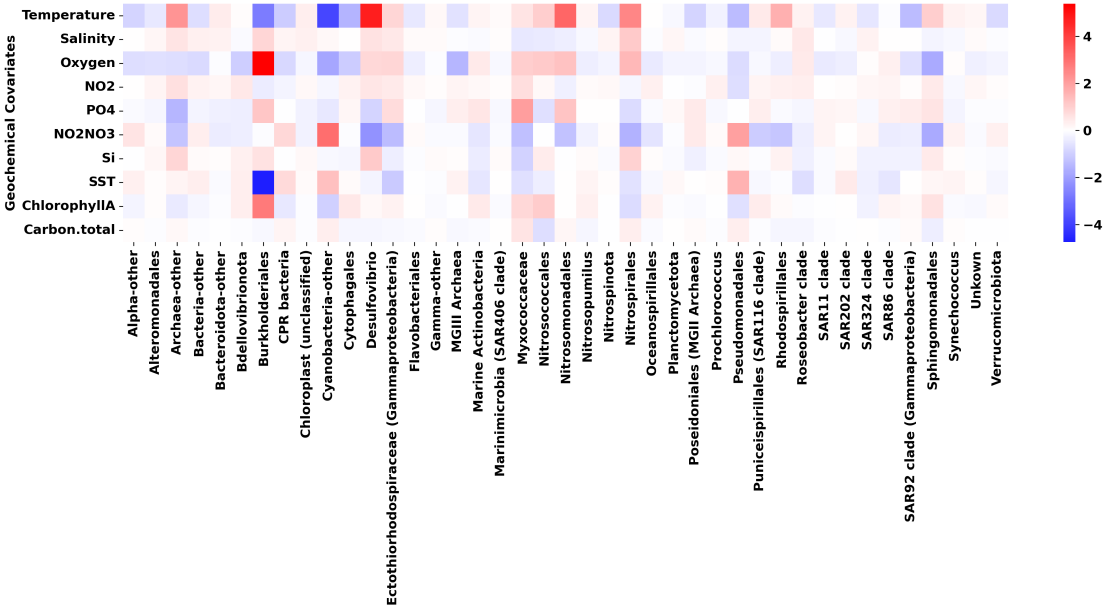

Figure 4.20: **No Direct Coupling**- estimated effect of Environmental covariates on all ERCs

Figure 4.9 provides a summary of the estimated average effect side of the influence of the environmental covariates on the ERCs for the no direct coupling model. Here, Temperature and Oxygen appear as strong drivers of abundance for several ERCs.

Temperature is the primary positive factor influencing the abundance of Cyanobacteria-other and negatively influences Desulfovibrio. Similar to the direct coupling model, Burkholderiales' abundance is positviely influenced by SST and Temperature and negatively influenced by Oxygen and ChlorophyllA.

The heatmap of estimated effect of Environmental covaiates on ERCs where the ERCs are the same as those in the original data can be found in Appendix A.

Positive and Negative Taxonomic Interactions

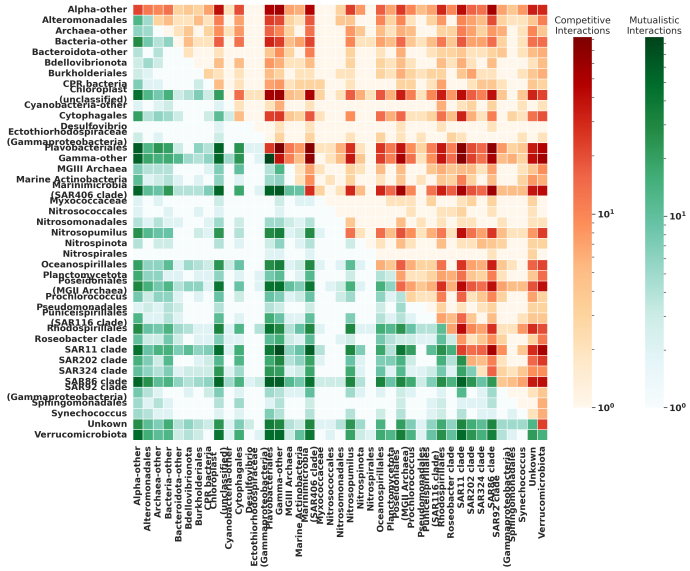

Figure 4.21: **Direct Coupling** - Taxonomic Interactions

The SAR11 clade, SAR86 clade and Rhodospirillales exhibits mutualistic interactions with majority of the ERCs. Plavobacteriales, Gamma-other and Marinimicrobia have competitive interactions with majority of the ERCs.

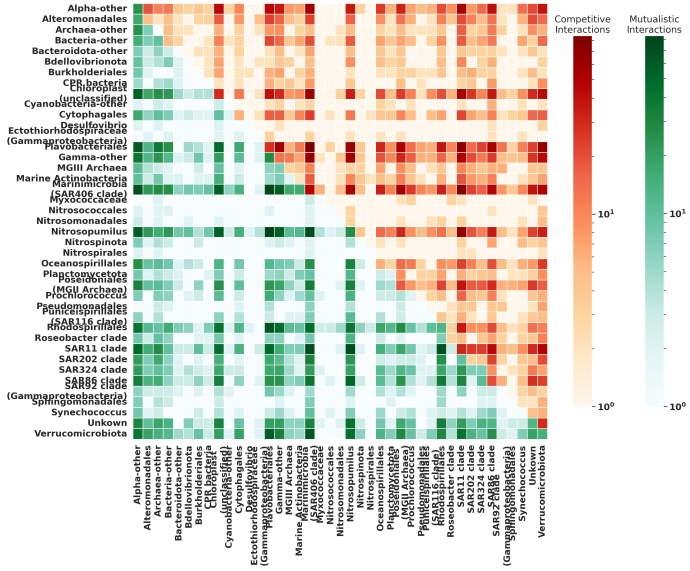

Figure 4.22: **No Direct Coupling** - Taxonomic Interactions

We see a similar interaction pattern for the ERCs that were highlighted in the direct coupling model as well. Both models show quite similar interactions - comparable patterns of

competitive and mutualistic interactions and the strengths of these relationships also does not different drastically between the two models.

Latent Representation, β

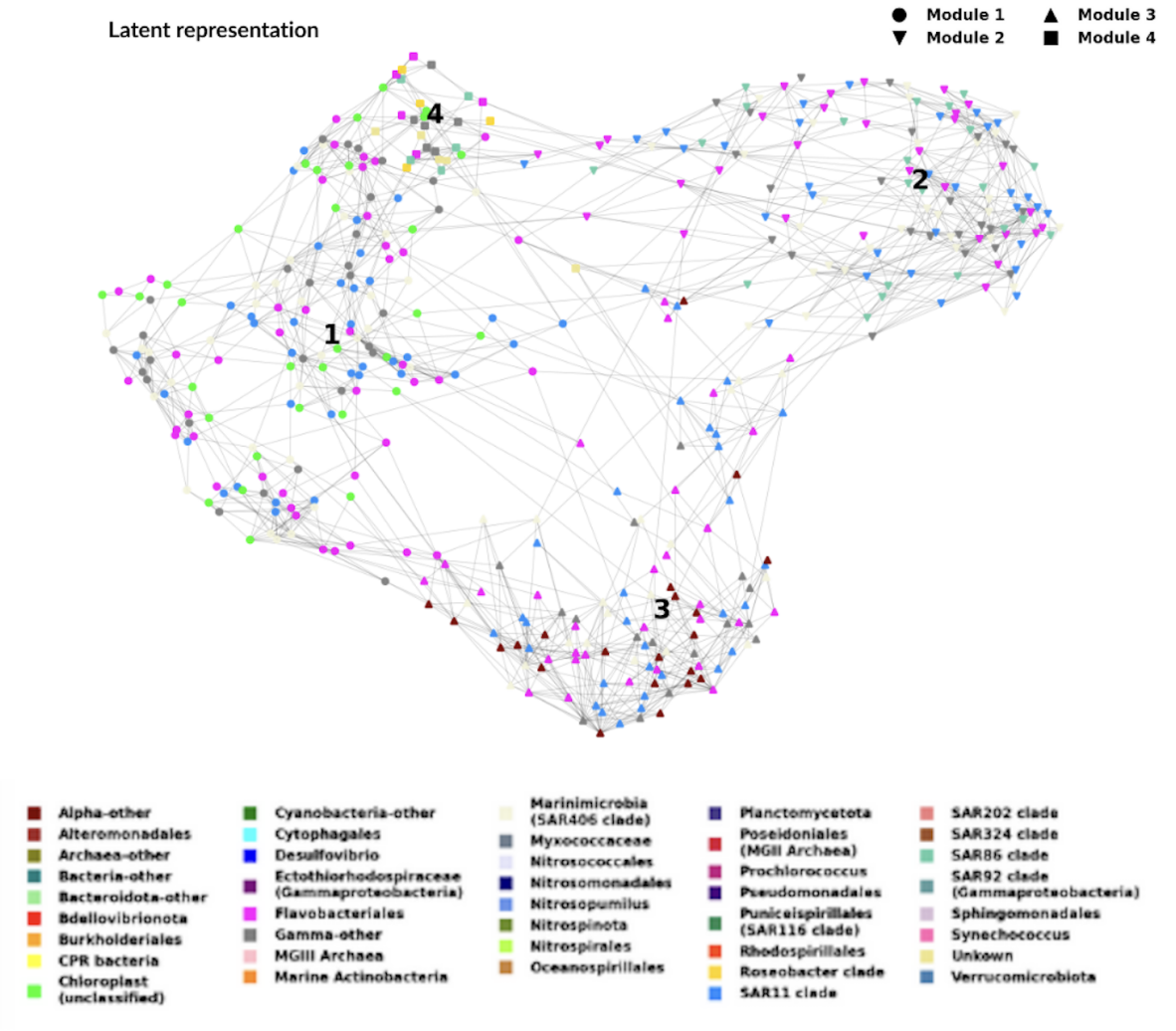

Figure 4.23: Low-dimensional embedding of the latent representation β using a k-nearest neighbor ($k_{\text{nn}} = 10$). Modularity analysis revealed 4 distinct graph modules.

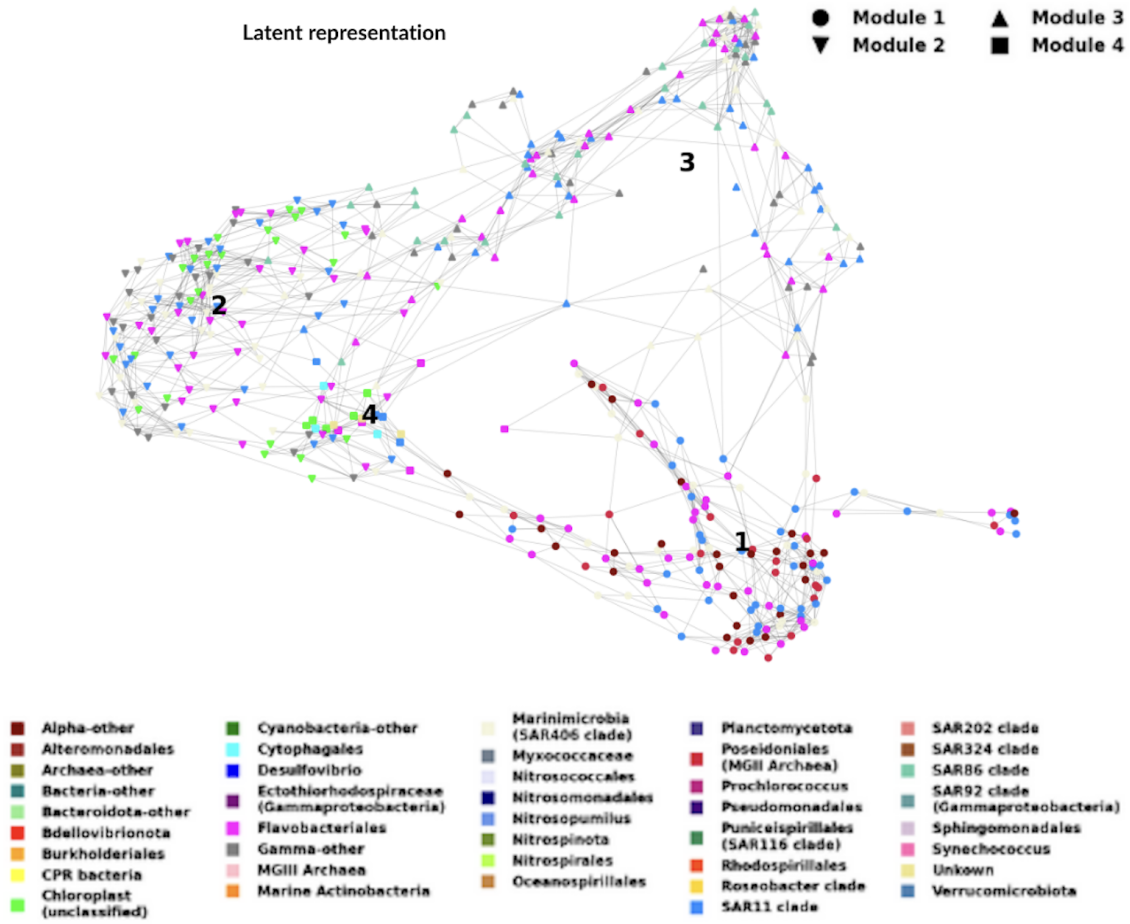

Figure 4.24: Low-dimensional embedding of the latent representation β using a k-nearest neighbor, ($k_{\text{nn}} = 10$). Modularity analysis revealed 4 distinct graph modules.

5 Discussion

In this study, we applied the VI-MIDAS framework using two modeling variants—direct coupling and no direct coupling—and analyzed two datasets. We established proof of concept by reproducing the VI-MIDAS direct coupling model from Mishra et al. (2024), and conducted our own hyperparameter tuning to identify the optimal settings and, consequently, the best-performing model. We then implemented the no direct coupling model on the original dataset, which represents the key novelty and extension we introduce to the VI-MIDAS modeling framework. Hyperparameter tuning was also conducted for this variant. The no direct coupling model outperformed the direct coupling model based on out-of-sample LLPD values.

We repeated this exercise on the new dataset; however, due to limited computational resources, we used the best hyperparameter settings obtained from the original dataset for the no direct coupling and direct coupling models on the new data. The next step to work on VI-MIDAS could involve systematic hyperparameter tuning on the new data to potentially improve model fit, as we have shown that hyperparameter tuning significantly improves model performance.

We used the Negative Binomial distribution within the VI-MIDAS generative modeling framework. Future exploration of alternative distributions such as the Dirichlet-Multinomial, could improve robustness and better accommodate sparse and over-dispersed data.

Exploratory data analysis revealed that the additional samples in the new dataset are distinct from those in the original data. EDA also showed that Spatial - Province (Biome) and Spatial - Depth Layer play key roles in shaping mOTU abundance. This finding was reinforced when excluding Spatial - Province (Biome) from the direct coupling model (new data) led to a significant drop in performance. We also examined the influence of environmental covariates in the context of oceanographic variation.

Our results highlight the respective strengths and limitations of both modeling variants. While the direct coupling model may be preferable for interpretability—particularly in estimating the effects of environmental covariates—the no direct coupling model excels in predicting overall microbial abundance. Both models yield consistent and similar insights into species-species interactions.

Thus, VI-MIDAS is a flexible modeling framework that allows for the incorporation of additional covariates as they become available—whether environmental, spatiotemporal, satellite-derived, or associated with additional samples—offering promising opportunities for future research.

6 Contribution

Iris contributed to the modeling aspect of the project by configuring the referenced repository from Mishra et al. (2024), running the model, and generating the results plot. She ensured the modeling workflow was reproducible and created comprehensive documentation to support future use by other researchers.

Eesha worked on data preprocessing, exploratory data analysis (EDA), oceanography-related theory, and parallel execution of the models. She also wrote this entire project report and contributed code.

All the code and output delivered by both is accessible through the link in Electronic Appendix B. Both team members coordinated throughout the project to ensure timely delivery of milestones and collaborated effectively to support each other's work.

A Appendix

Direct Coupling (new data) - Environmental Covariates on ERCs

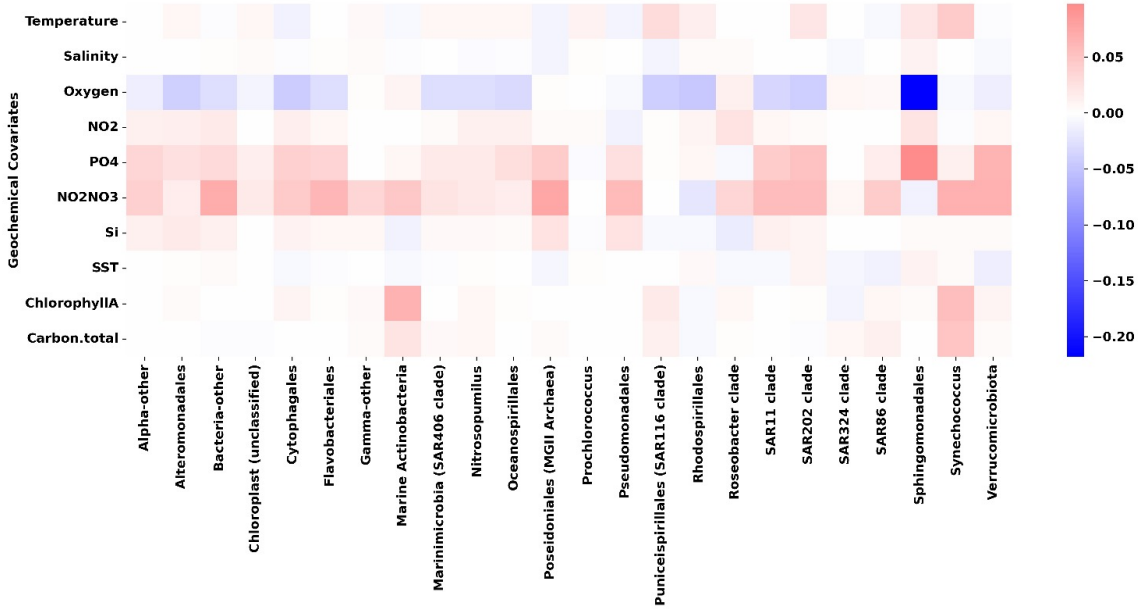

Figure A.1: Estimated effect of Environmental covariates on the same ERCs as the original data

No Direct Coupling (new data) - Environmental Covariates on ERCs

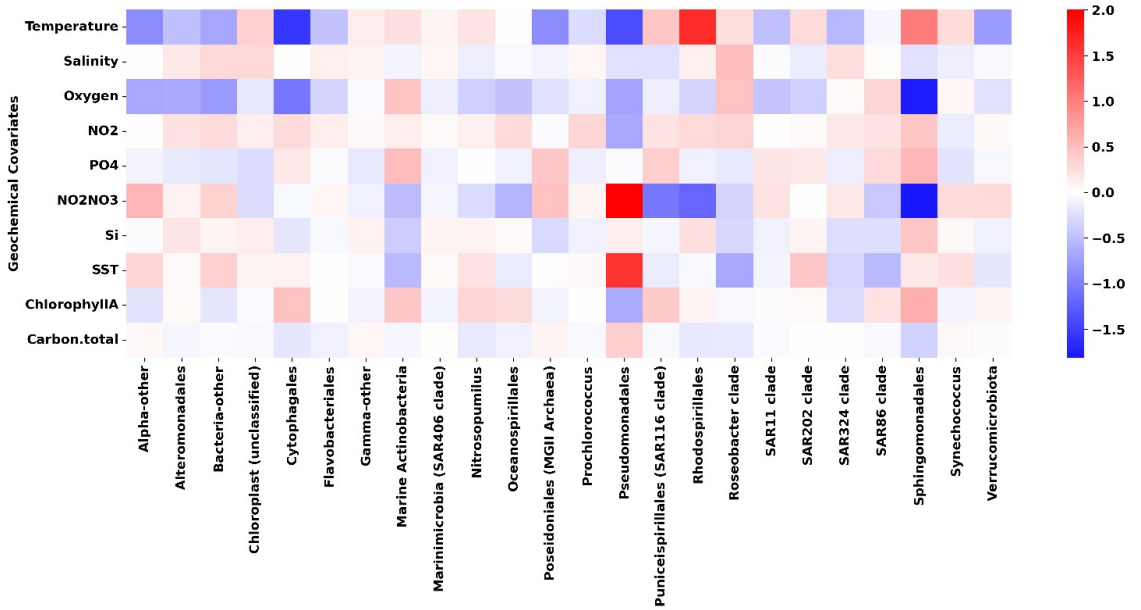

Figure A.2: Estimated effect of Environmental covariates on the same ERCs as the original data

Definitions:

The total **library size** refers to the sum of all mOTU counts in a single sample. It represents the total number of reads (or counts) for that sample. Relative abundance is calculated by dividing each mOTU count by this total library size.

Relative abundance is the proportion of each mOTU's count compared to the total count in a sample. It's calculated by dividing each mOTU's count by the sample's total count (i.e., row sum). This normalizes data across samples, making them comparable regardless of sequencing depth.

Table 7: PERMANOVA results testing the effect of polar status (polar vs non polar) and depth layer on microbial community composition. Based on Bray-Curtis dissimilarity with 999 permutations.

Term	**Df**	**Sum of Squares**	**R^2**	**F**	**p-value**
Model	4	22.362	0.36387	25.025	0.001
Residual	175	39.093	0.63613		
Total	179	61.455	1.00000		

Table 8: PERMANOVA results testing the effect of polar status (polar vs non-polar) and biome on microbial community composition. Based on Bray-Curtis dissimilarity with 999 permutations.

Term	**Df**	**Sum of Squares**	**R^2**	**F**	**p-value**
Model	4	14.880	0.24213	13.977	0.001
Residual	175	46.575	0.75787		
Total	179	61.455	1.00000		

VI-MIDAS - direct coupling

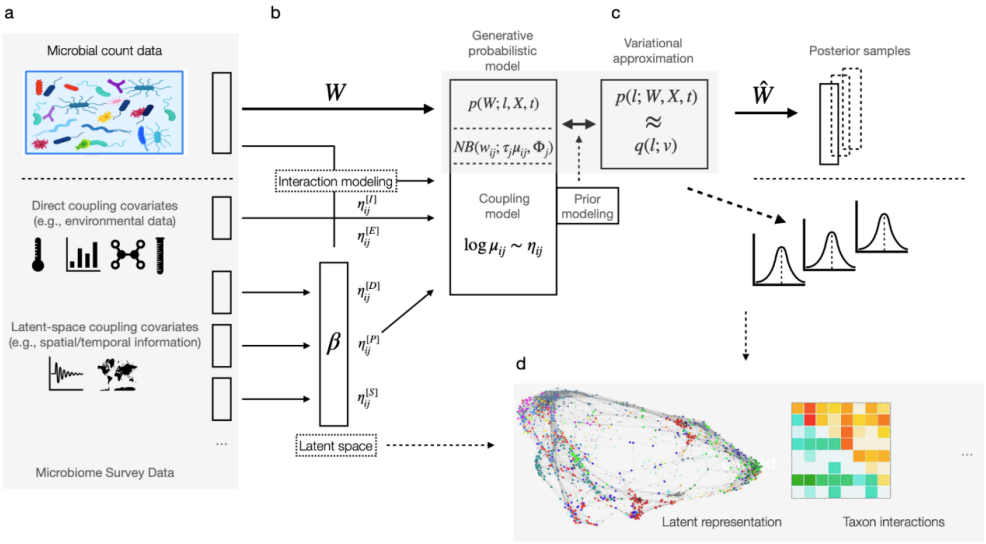

Figure A.3: Direct Coupling model - schematic diagram adapted from (Mishra et al., 2024)

VI-MIDAS - no direct coupling

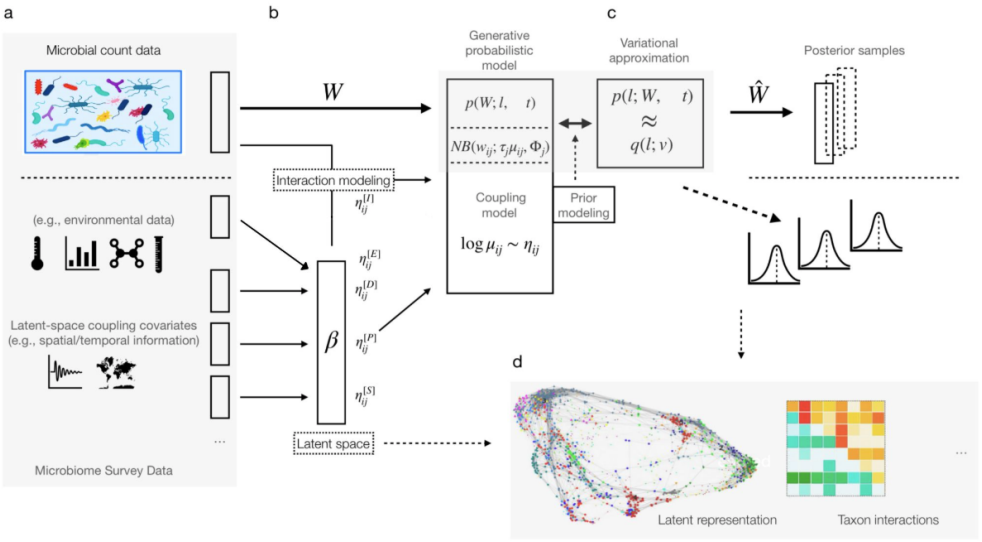

Figure A.4: No Direct Coupling model schematic diagram adapted from (Mishra et al., 2024)

B Electronic appendix

The data sets, code for the models, preprocessing, eda and analysis is available here : https://github.com/irisdaniaj/vi-midas

The repository is public.

References

Birdeau, L. (2023). Starting with non-metric multidimensional scaling (nmds), *University of Virginia Library: StatLab Articles* .
URL: *https://library.virginia.edu/data/articles/starting-non-metric-multidimensional-scaling-nmds*

Blei, D. (2017). *Variational Inference: Foundations and Innovations*, Video presentation, Simons Institute for the Theory of Computing.
URL: *https://simons.berkeley.edu/talks/variational-inference-foundations-innovations*

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational Inference : A Review for Statisticians, *Journal of the American Statistical Association* **112**(518): 859–877.
URL: *http://dx.doi.org/10.1080/01621459.2017.1285773*

Cai, W. J. and Ouyang, Z. (2022). Melting sea ice is acidifying the Arctic Ocean, *World Wildlife Fund : The Circle* .
URL: *https://www.arcticwwf.org/the-circle/stories/melting-sea-ice-is-acidifying-the-arctic-ocean/*

Ebner, J. (2018). *Permutational Multivariate Analysis of Variance (PERMANOVA) in R.*
URL: *https://tinyurl.com/msxmwr6z*

Gelman, A., Hwang, J. and Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models, arXiv preprint.
URL: *https://arxiv.org/abs/1307.5928*

Mishra, A., McNichol, J., Fuhrman, J., Blei, D. and Mueller, C. L. (2024). *Variational inference for microbiome survey data with application to global ocean data*, biorXiv preprint.
URL: *https://doi.org/10.1101/2024.03.18.585474*

NOAA (2019). *Arctic Report Card: Update for 2019.*
URL: *https://tinyurl.com/2448srvb*

NOAA (2023). *Sea Water.*
URL: *https://www.noaa.gov/jetstream/ocean/sea-water*

Webb, P. (2023). *Introduction to Oceanography*, Roger Williams University Open Publishing.
URL: *https://rwu.pressbooks.pub/webboceanography/*

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Munich, 06.05.2025

Eesha Samir Chitnis and Iris Dania Jimenez