UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

STUDYING THE RELATIONSHIP OF BIOTIC AND ABIOTIC PROCESSES IN THE OCEAN BIOME FROM THE POINT OF VIEW OF GENOMIC REGULATION USING MATHEMATICAL MODELING AND MACHINE LEARNING TECHNIQUES.

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS.

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO.

NICOLÁS ANDRÉS TORO LLANCO

PROFESOR GUÍA:
ALEJANDRO MAASS SEPÚLVEDA
PROFESOR CO-GUÍA:
SEBASTIÁN DONOSO FUENTES

MIEMBROS DE LA COMISIÓN:
MAURICIO GONZÁLEZ CANALES
JAIME SAN MARTÍN ARISTEGUI

Este trabajo ha sido parcialmente financiado por:
CMM ANID BASAL FB210005, INSTITUTO MILENIO Nº ICN2021 044 Y
PROYECTO EXPLORACIÓN ANID Nº 1322000.

SANTIAGO DE CHILE
2023

RESUMEN DE TESIS PARA OPTAR AL GRADO
DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA,
MENCIÓN MATEMÁTICAS APLICADAS
Y MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL MATEMÁTICO
POR: NICOLÁS ANDRÉS TORO LLANCO
FECHA: 2023
PROF. GUÍA: ALEJANDRO MAASS SEPÚLVEDA

ESTUDIANDO LA RELACIÓN DE LOS PROCESOS BIÓTICOS Y ABIÓTICOS EN EL BIOMA OCEÁNICO DESDE EL PUNTO DE VISTA DE LA REGULACIÓN GENÓMICA UTILIZANDO MODELACIÓN MATEMÁTICA Y TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

El océano, que abarca el 70% de la superficie de la Tierra, presenta muchos ecosistemas, cada uno caracterizado por hábitats, temperaturas y nutrientes distintos. En el núcleo de estos ecosistemas se encuentra el microbioma oceánico, dominado por entidades planctónicas como las bacterias, que juegan un papel fundamental en los ciclos biogeoquímicos, influyen en los patrones climáticos globales y contribuyen al ciclo del carbono de la Tierra. Con un enfoque en la regulación genómica dentro de estas comunidades bacterianas, este trabajo usa el conjunto de datos generado por la expedición TARA Oceans para estudiar las relaciones de los factores de transcripción con las variables ambientales marinas.

En esta tesis, nuestro objetivo específico es analizar cómo las abundancias de los motivos de unión asociados a una familia de 88 factores de transcripción, presentes en las regiones intergénicas de un metagenoma bacteriano, pueden capturar las condiciones ambientales. Para ello, utilizamos los metagenomas bacterianos reconstruidos de las expediciones TARA Oceans y construimos una matriz de abundancia, donde las filas representan una muestra (o un metagenoma bacteriano), las columnas están asociadas a un factor de transcripción (de los 88 utilizados), y en cada posición de la celda almacenamos la abundancia de los motivos de unión asociados al factor de transcripción en las regiones intergénicas de la muestra. El objetivo principal de este trabajo es descubrir si esta información biótica está relacionada de alguna manera con los datos ambientales, en particular, si podemos predecir características del ambiente a partir de la información regulatoria de los metagenomas bacterianos encapsulada en esta matriz.

Analizamos nuestro conjunto de datos de variables ambientales y biológicas primeramente desde un punto de vista descriptivo. Investigamos la estructura de estas variables, revelando agrupaciones de factores de transcripción independientes de su funcionalidad e identificando interacciones biótico-abióticas clave influenciadas por la geografía y la profundidad del agua marina. Luego exploramos la estructura de nuestras matrices biológicas utilizando reducción de dimensionalidad y construimos modelos predictivos. Estos modelos diferencian muestras de aguas oceánicas polares y no polares, regiones oceánicas y profundidades de capas de agua. Desarrollando un concepto de robustez para las predicciones, enfatizamos, por ejemplo, los roles de `FabR` y `BirA` en la diferenciación de la polaridad y capas oceánicas respectivamente.

Estos hallazgos subrayan el papel de los factores de transcripción como sensores ambientales relevantes, afirmando nuestra hipótesis inicial. Además, refuerzan la noción de que un número limitado de componentes puede producir predicciones significativas, en contraste con el énfasis en genes o virus como objeto principal de estudio.

RESUMEN DE TESIS PARA OPTAR AL GRADO
DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA,
MENCIÓN MATEMÁTICAS APLICADAS
Y MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL MATEMÁTICO
POR: NICOLÁS ANDRÉS TORO LLANCO
FECHA: 2023
PROF. GUÍA: ALEJANDRO MAASS SEPÚLVEDA

STUDYING THE RELATIONSHIP OF BIOTIC AND ABIOTIC PROCESSES IN THE OCEAN BIOME FROM THE POINT OF VIEW OF GENOMIC REGULATION USING MATHEMATICAL MODELING AND MACHINE LEARNING TECHNIQUES.

The ocean, encompassing 70% of Earth's surface, presents an intricate tapestry of ecosystems, each characterized by distinct habitats, temperatures, nutrients, etc. At the core of these ecosystems is the ocean microbiome, dominated by planktonic entities such as bacteria, which play a paramount role in biogeochemical cycles, influence global climate patterns, and contribute to Earth's carbon cycle. With an emphasis on the genomic regulation within these bacterial communities, this research exploits the immense dataset generated by the TARA Oceans expedition to investigate the relations of transcription factors, the molecular switches of genomic regulation, with marine environmental variables.

In this thesis, we specifically aim to analyze how the abundances of the binding motifs associated to a family of 88 transcription factors, appearing in the intergenic regions of a bacterial metagenome, are able to capture the environmental conditions. For that, we used the bacterial metagenomes reconstructed from TARA Ocean expeditons and we build an abundance matrix, where rows represent a sample (or a bacterial metagenome), columns are associated to a transcription factor (among the 88 used), and at each cell position we store the abundance of the binding motifs associated to the transcription factor in the intergenic regions of the sample. The main objetive of this work is to unravel whether this biotic information is related in some way with environmental data, in parituclar, if we can predict characteristics of the enviroment from the regulatory information of bacterial metagenomes encapsulated in this matrix.

We comprehensively analyzed our dataset of environmental and biological variables, referencing literature for the environmental aspects and visualizing distributions for the biological ones. We probed the structure of these variables, revealing clusters of transcription factors independent of their functionality and identifying key biotic-abiotic interactions influenced by geography and seawater depth. We then explored our biological matrices' geometry using dimensionality reduction and built predictive models. These models differentiate samples from polar-non polar, ocean regions and layer depth seawaters. Developing a robustness concept for the predictions, we emphasize, for instance, the roles of `FabR` and `BirA` in polarity and layer differentiation respectively.

These results accentuate transcription factors as key environmental indicators, demonstrating that a few select components can provide significant predictions. This challenges the conventional focus on genes or viruses as primary study objects.

A mi familia

Acknowledgments

La culminación de este trabajo representa el esfuerzo colectivo de generaciones pasadas, que con sacrificio permitieron que hoy pueda finalizar esta tesis y cumplir el anhelo que ellos no tuvieron la oportunidad de lograr.

Por un lado, quiero agradecer a mis abuelos maternos, quienes desde su llegada a Santiago a los 14 años, lograron sortear la pobreza extrema y sacar adelante a la familia. Mi abuelo Mariano, nacido en Nueva Imperial, llegó a Santiago y desempeñó diversos trabajos, desde hacer aseo y ser pastelero hasta laborar en fábricas textiles y, finalmente, establecerse como feriante. Por su parte, mi abuela, la *mama rosa*, originaria de San Nicolás, arribó a Santiago y trabajó como nana, posteriormente como dueña de casa y, al final, también se desempeñó como feriante. Sus historias siempre vivirán en mi y les estaré eternamente agradecido por todo, especialmente por acogerme y preocuparse de mi bienestar en mis años universitarios.

De igual manera, me gustaría agradecer a mis padres, Sergio y Carmen, quienes me han apoyado en todo y forman parte importante de lo que soy. Ambos, también con sus historias de sacrificio, colaboraron para que yo pudiera salir adelante y tener una vida mejor. Igualmente, les estaré por siempre agradecido.

Quiero también extender mis agradecimientos a mis amigos: Milena, Marcelo, Eduardo, Nicolás, Matias, Fernanda, Enrique, Maria José, Juan Carlos, Joaquín, Daniel, Juan y Catalina por hacer mas amenos todos estos años. Además, un especial agradecimiento a Constanza, quien ha sido un apoyo fundamental en este proceso y acompaña mis días con amor y cariño.

En el ámbito académico, quisiera dar las gracias en primer lugar a Alejandro por su apoyo constante durante la realización de este trabajo y por darme la oportunidad de trabajar con él, introduciéndome al mundo de la matemática aplicada. En segundo lugar, a Sebastián por su comprensión como docente en los momentos difíciles de pandemia.

Quiero finalmente dar las gracias al laboratorio Mathomics perteneciente al Centro de Modelamiento Matemático de la Universidad de Chile y en particular a Ricardo Palma e Iñaki Hojas por proveer la información genómica y las matrices de abundancias usadas en este trabajo.

Esta tesis fue parcialmente financiada por CMM ANID BASAL FB210005, Centro para la Regulación del Genoma (CRG), Instituto Milenio Proyecto Nº ICN2021 044, y el Proyecto Exploración 'Deciphering the regulatory architecture of microbial communities' de ANID número 1322000.

Table of content

List of Tables

List of Figures

Chapter 1

Introduction

Spanning approximately 70% of Earth's surface and housing 97% of all water, the ocean biome is an awe-inspiring, complex ecosystem. It is far from being a monolithic entity; instead, it comprises a vast range of habitats, from the sunlit, surface-dwelling epipelagic zone to the deep, lightless abyss of the hadalpelagic zones. These marine realms exhibit remarkable variation in temperature, salinity, pressure, and nutrient availability, creating unique ecosystems that host an extraordinary array of organisms [1].

Among these life forms, plankton - composed of zooplankton, protists, bacteria, archaea and viruses - hold a dominant position in the so-called ocean microbiome. These mostly microscopic entities serve as the foundation of marine food webs [2] and are pivotal to global primary production, playing a key role in Earth's carbon cycle and helping maintain the planet's climate by capturing a large fraction of atmospheric carbon dioxide and releasing oxygen via photosynthesis [3].

Bacteria, a crucial component of the planktonic community, play a significant role in various biogeochemical cycles, contributing to the provision of vital ecosystem services. Their pivotal role extends to impacting global climate patterns, largely through their involvement in the carbon cycle. The enormous genomic diversity within these bacterial communities [4] offers a unique opportunity to explore genetic adaptations to different environments, potentially leading to biotechnological advancements. Understanding how these bacteria respond to, and adapt within, their highly variable and dynamic environment is important due to the aforementioned reasons. The objective of this work is to address these issues through the lens of genomic regulation.

Genomic regulation refers to the complex network of processes that control gene expression within an organism. It governs how genes in the DNA are transcribed into RNA and translated into proteins, effectively controlling the function and adaptability of an organism. Central to this regulation are transcription factors, proteins that bind to specific DNA sequences (transcription factors binding motifs) to help in the regulation of the transcription of DNA to mRNA [5]. These transcription factors act as molecular switches, initiating or halting the transcription process in response to environmental or intracellular signals; in this sense or abstractly, they can be thought as environmental sensors of the ocean biome [6].

Historically, major oceanic expeditions like the late 19th century Challenger Expedition and the early 21st century Global Ocean Sampling Expedition have paved the way for our understanding of the ocean biome and its inhabitants [7, 8]. The TARA Oceans expedition, launched in 2008, expanded this understanding further by bringing back around 35,000 samples based on high throughput DNA sequencing and advanced microscopy. Notably, these

samples span the entirety of the ocean, allowing for distinctions based on the pelagic zone depth, the specific oceanic region of origin, and whether the region is polar or non-polar, among other characteristics [1].

Although this project has generated a vast array of genomic data, the scale and complexity of this information make it challenging to derive meaningful insights. The relative scarcity and crucial role of transcription factors as environmental sensors provide a manageable and practical focus for our study.

Building on the rich literature derived from TARA Oceans expeditions and capitalizing on the role of the transcription factors as environmental sensors, the present work seeks to address a key question: "How are the abundances of binding motifs of transcription factors in oceanic bacterial metagenomes related to environmental factors in the ocean biome?". The formulation of this question is not random, reflecting observations that transcription factors adjust their regulatory targets under varying environmental scenarios [9]. Such adaptability suggests that cellular responses are frequently an outcome of multiple transcription factors working together to modulate gene expression in response to specific environmental challenges [6]. The significance of addressing this question lies in its departure from contemporary approaches. Instead of the conventional descriptive focus on genes or viruses, as prevalent in current literature, this investigation emphasizes the genomic regulation of the organisms providing insights into the factors that determine the composition of microbial community transcriptomes. Moreover, the relative scarcity of transcription factors presents a unique analytical advantage, enabling a clearer insight when these factors emerge as predictors of abiotic variables. To answer this, we examine the binding motifs associated with a family of 88 transcription factors found in the intergenic regions of bacterial metagenomes. Our primary dataset comprises bacterial metagenomes from the TARA Oceans expeditions. We have structured this data into an abundance matrix: each row signifies a bacterial metagenome sample; each column corresponds to one of the 88 transcription factors; and each cell records the abundance of binding motifs linked to its respective transcription factor in the sample's intergenic regions.

Utilizing this dataset, we discerned structural relationships within our data, which facilitated the clustering of transcription factors based on correlation-type distances. Additionally, we constructed bipartite graph-based correlation networks to capture biotic-abiotic interactions. We further explored the dataset's predictive potential by training machine learning models, specifically those based on extreme gradient boosting, assessing their capability to accurately predict environmental variables. Building upon this, we shifted our focus to discern which biological variables drive these predictive outcomes. To achieve this, we devised a robust feature selection model based on a Monte Carlo cross-validation method.

In light of the diverse origins of the samples derived from the TARA Oceans expeditions, an intriguing query arose: Do categorizations, such as by pelagic zone, polar region, or specific oceanic area, influence our findings? More specifically, do these categorizations alter the biological 'drivers' with predictive potential?. Furthermore, while various statistical methods, namely causality methods, Neural Networks, and Graph Neural Networks, could be employed given the novelty of this dataset, the 'workflow' we've crafted in this study — encompassing an exploratory data analysis, a comprehensive spearman-correlation analysis of the variables to discern their structural relations, followed by the construction of predictive models using boosting algorithms — is specifically tailored to address the primary question of this thesis. Not only does it aim to answer the fundamental query, but it also strives to pinpoint those transcription factors that excel in predicting environmental fluctuations.

This endeavor is rooted in an interdisciplinary approach that bridges genomic regulation, environmental science, mathematical modeling, and data science.

The primary contributions of this thesis include: 1) A detailed exploration of biotic signals across the ocean using the abundance of binding motifs related to specific transcription factors in bacterial metagenomes. We presented their distributions and showed that approximately two-thirds of the transcription factors under study follow a normal distribution; furthermore through the application of Andrew's curves to the biological dataset we were able to differentiate various environmental target variables. This thorough exploratory analysis, elaborated in Chapter 3, not only provides deeper insights for the current study but also sets a foundation for future research in the area, 2) In contrast to prevailing literature that indicates latitudinal diversity of genes or species, our study uniquely focused on the functionality associated with each transcription factor, as sourced from RegPrecise. Strikingly, we discovered an absence of functional latitudinal diversity, challenging established understandings and contributing a novel perspective to the field. This is detailed in the latter part of Chapter 3, 3) A groundbreaking identification of clusters of transcription factors based on their binding abundance. This novel approach illuminated strong biotic and biotic-abiotic relationships, providing a more nuanced understanding of the intricate interactions in marine ecosystems. This significant contribution further enhances our comprehension of how environmental factors influence bacterial transcriptional regulation and it is elaborated in Chapter 4, 4) A demonstration of the predictive prowess of the binding abundance matrix for polar/non polar, ocean pelagic zones and ocean regions target variables, emphasizing the immense potential of this newfound data resource. Additionally, we developed a novel mathematical modeling approach to identify key transcription factors that play pivotal roles in predicting the aforementioned environmental target variables. This involves utilizing a variation of the Monte Carlo Cross-Validation algorithm to rank biological features. The ranking is built upon feature and permutation importance metrics. All these methodologies and innovations are detailed further in Chapter 5 and 5) An integrative analysis of existing literature to contextualize and interpret the study's results. This involved drawing connections between independent studies on specific transcription factors, seamlessly blending the biological perspective with oceanic applications.

The implications of this research for policy-making can be relevant. A deeper grasp of genomic regulation in marine bacteria can guide policy choices related to marine conservation, adapting to climate change, and managing biodiversity. In particular, if our studies shed light on important oceanic functions like the carbon pump.

In essence, the goal of this research is to augment our comprehensive understanding of the multifaceted nature and resilience of the oceanic ecosystem, especially when confronted with environmental shifts. This could, in turn, shed light on and shape policy and conservation initiatives aimed at preserving our oceans.

Chapter 2

Background and Literature Review

The ocean is the foundation for the global health of our planet, yet we know surprisingly little about it. From 2009 to 2013, the research sailing ship Tara, owned by the Tara Ocean Foundation, sailed the oceans of the world to collect samples of microscopic plankton.

This expedition brought back around 35,000 samples based on high throughput DNA sequencing and advanced microscopy. Much has already been done, including 1) description of the ocean microbiome containing around 40 million genes, an atlas of 116,000,000 genes from eukaryotes and characterization of close to 200,000 different types of viruses [1], 2) identification of two subnetworks of gene functions that are significantly associated with carbon export [10], 3) replicating the Latitude Gradient Diversity hypothesis, which has been proven on land, in seawaters [11], among other results; but there are still questions that remain open.

In this chapter, we present a concise overview of pivotal studies concerning the ocean microbiome, detailing its structure, function, and impact, as well as its intricate relationship with the surrounding environment. This literature review serves as a state-of-the-art contextualization, offering a comprehensive understanding of the current scientific landscape. Through this examination, we aim to spotlight the multifaceted and captivating dynamics of the marine ecosystem, emphasizing its essential role in supporting oceanic life. Armed with this foundational knowledge, we will launch into our own investigation, building upon this established framework.

2.1. Marine Microbiome Diversity and Function

Microbes are the invisible giants of the marine ecosystem. Despite their minuscule size, these organisms play a fundamental role in marine environments, contributing significantly to the planet's biogeochemical cycles. Understanding the structure and function of the marine microbiome is therefore key to unravelling the mysteries of the ocean functioning and its global impact.

One of the most comprehensive studies to date on the global ocean microbiome was carried out through the Tara Oceans expedition, where the team analyzed 7.2 terabases of metagenomic data from 243 samples across 68 locations from epipelagic and mesopelagic waters worldwide, in which samples where primarily structured by depth at each location. A more detailed understanding of the sampling method can be found in the paper by Sunagawa et al. [1].

The study, which generated a microbial reference gene catalog with more than 40 million

nonredundant sequences, identified viruses, prokaryotes, and picoeukaryotes as the primary constituents of the ocean's microbial life. A critical finding from it is that the microbial community's composition in epipelagic waters is primarily driven by temperature, as opposed to other environmental factors or geography [12]. Moreover, recent studies also showed that the diversity of all major groups of plankton is highest around the equator and decreases around the poles [11]. The existence of such latitudinal diversity gradients is well established on land and was first described by Alexander von Humboldt on XIX century. This suggests that global warming could significantly alter the biodiversity and functional dynamics of marine microbial communities [12, 11].

An additional notable finding from the study is the shared functional core between the oceanic and human gut microbiomes, despite the vast physicochemical differences between these two ecosystems [12]. This underscores the pervasive influence of microbiomes across diverse habitats and suggests possible parallels in their responses to environmental changes.

Understanding the structure and function of the marine microbiome is crucial for predictive models of the ocean and its potential responses to climate change. Moreover, the generated TARA Oceans dataset has become a significant resource for investigating the biodiversity and functional roles of marine microbes [12]. For instance, in a pivotal study, Paul Fremont [13] utilized machine learning to analyze plankton biogeography through genomics. He identified unique oceanic genomic provinces, shaped by major currents and excluding only the Arctic Ocean. Alarmingly, with a high greenhouse gas emission scenario, these provinces face an extensive reconfiguration by the 21st century's end, impacting over 50% of the oceans studied and potentially causing a 4% drop in export production. By assembling numerous plankton genomes, Fremont enhanced our understanding of the provinces' genomic structure and the key species defining them, offering insights into how plankton communities might evolve under climate change.

2.2. Gene Expression in the Ocean Microbiome

Investigating gene expression within the ocean microbiome is integral to our understanding of marine microbial community adaptation and functionality.

In this sense, and as a complement of the study of plankton diversity [11], a study measured the activity of microbial communities by analyzing gene transcripts in combination with a newly established catalog of 47 million microbial genes [14]. These analyses allowed us to study not only what ocean microbes are capable of doing (MetaG), but also what they actually do (MetaT).

Indeed, the evidence suggests that gene expression within the ocean microbiome can exhibit significant geographical and depth-related variations [14]. One example for this is that microbial communities in warmer waters are more diverse and benefit from a large pool of genes [11, 14], which can be switched on or off to help the microbes adaptation. In polar waters, however, the variety of species and genes is much smaller. These communities are more hardwired to their environment. They might struggle to adapt their activity by changing gene expression in response to ocean warming, leading to the conclusion that alterations in community activity in response to ocean warming in polar regions might be more predominantly driven by changes in the composition of organisms (community turnover), rather than shifts in gene regulatory mechanism (gene expression) [14].

In a more detailed examination, it has been revealed that the variance observed in micro-

bial community transcriptomes is a product of two primary factors: community turnover[1] and gene expression changes[2] (See Figure 2.1). Notably, the influence of these factors is temperature-dependent, with a pivotal transition at 15°C. Below this thermal threshold, community turnover plays a more significant role in how communities vary. Conversely, at temperatures above 15°C, adaptations are predominantly driven by alterations in gene expression. This temperature demarcation aligns with the earlier observation that in the relatively stable and cold polar waters, microbial communities, which have a reduced diversity of species and genes, may rely more on changes in community composition to respond to climatic shifts. In contrast, the more diverse microbial communities in warmer waters have the capacity to adapt through more dynamic gene regulation.

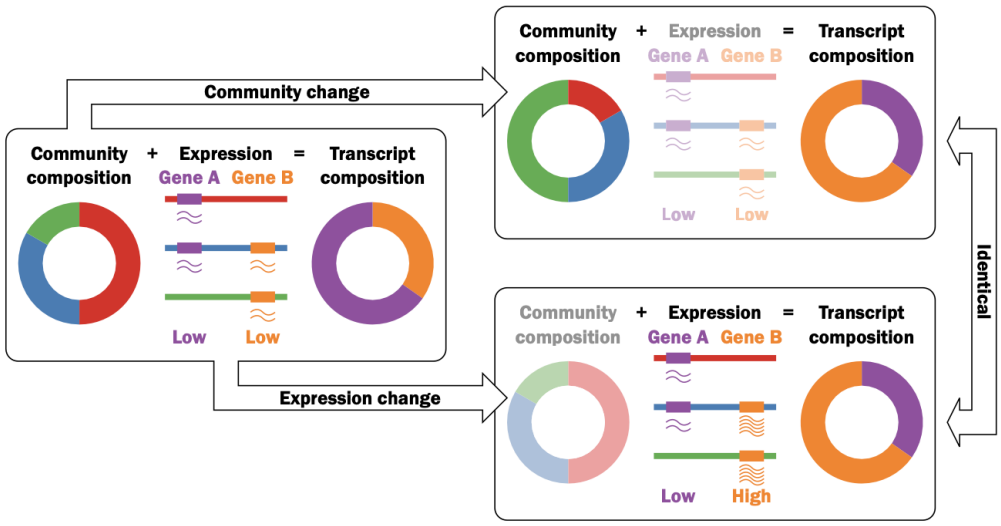


Figure 2.1: Cartoon exemplifying how an initial community with a given expression profile may result insimilar transcript abundance profiles through two different mechanism: 1) changes in the community composition and 2) changes in gene expression [14]

The warmer and colder waters thus appear as two ecosystems with distinct adaptive mechanisms for their microbial populations. The paramount influence of temperature raises obvious questions related to climate change. How will these communities be affected and what could be the consequences? Using the IPCC climate models, the results indicate that higher oceanic temperatures would lead to a tropicalization of temperate and polar regions with an increased diversity of planktonic species [14].

[1] Community turnover refers to the changes in species composition within an ecosystem over time due to environmental factors, disturbances, or human activities. It is a critical aspect of ecological dynamics, affecting biodiversity and ecosystem functionig.

[2] Gene expression change refers to the variation in the rate or manner at which genes are transcribed and translated into proteins, typically in response to environmental cues or cellular conditions, affecting cellular function and organismal adaptation.

2.3. Carbon Export in Marine Ecosystems

The *biological carbon pump* is a critical process in the world's oceans [15], whereby CO2 is transformed into organic carbon through photosynthesis, exported through sinking particles, and eventually sequestered in the deep ocean. The intensity of this pump has significant implications for the global carbon cycle and is closely tied to the composition of plankton communities [10].

A study based on environmental and metagenomic data from the Tara Oceans expedition has shed light on the ecosystem structure driving this process in the oligotrophic ocean, characterized by low nutrient concentrations. Notably, specific plankton communities at the surface and the *deep chlorophyll maximum*[3] were found to correlate strongly with carbon export at 150m depth [10].

Interestingly, some of the taxa associated with carbon export included Radiolaria and alveolate parasites, in addition to the expected role of phytoplankton like Synechococcus and their phages [10]. These findings underscore the complex nature of the biological carbon pump and its ties to the broader marine ecosystem.

Furthermore, this research suggests that the relative abundance of a few bacterial and viral genes can predict a significant portion of the variability in carbon export in these regions [10]. As climate change is predicted to alter phytoplankton size and diversity, further understanding of these relationships is crucial for predicting the future behavior of the oceanic carbon sink [10].

The influence of marine viruses, often overlooked due to sampling limitations, emerged as potentially significant in this study. The data suggest that viral lysis, usually considered as reducing the intensity of the biological carbon pump, might actually enhance carbon export through the production of colloidal particles and *aggregate formation*[4] [10]. This revelation implies that we need a more nuanced understanding of viruses' roles in carbon export.

Overall, the process of carbon export in marine ecosystems is far more complex and nuanced than previously understood. Future research should aim to illuminate the specific mechanisms by which different microbial communities and their interactions influence the carbon cycle.

2.4. Role and Influence of Viruses in the Ocean

Viruses in the marine environment represent a vast reservoir of genetic diversity, with their abundance far surpassing that of any other life form in the sea [16, 17]. Indeed, estimates suggest that there are as many as 10^{30} viruses in the ocean. These viruses infect hosts from various marine species, from microorganisms to large marine animals like shrimp and whales [17]. Such infections are a major source of mortality and cause various diseases, thereby playing a significant role in shaping the composition of marine communities [17].

Notably, marine viruses do not merely destroy their hosts but also serve as vehicles for gene flow in the ocean. Each viral infection carries the potential to introduce new genetic information into an organism or progeny virus, driving the evolution of both host and viral assemblages [17]. By this means, they contribute to the genetic diversity of marine life.

In addition to genetic diversity, marine viruses play a critical role in global biogeochemical

[3] It is the region below the surface of water with the maximum concentration of chlorophyll

[4] The process where particles clump into larger masses, playing a vital role in carbon sequestration and providing food and habitats for deep-sea life

cycles [16, 17]. They influence these cycles by regulating the population sizes and metabolic outputs of their microbial hosts. Moreover, they have been shown to be a major force behind the biological carbon pump, contributing to carbon export in the marine ecosystem [10].

The influence of marine viruses on ocean biogeochemistry was underscored by the findings from the Tara Oceans expedition. Quantitative dsDNA viral-fraction metagenomic (viromic) datasets from the expedition indicated that viral communities in the upper ocean were locally structured by environmental conditions affecting host community structure [16]. These observations support the seed-bank hypothesis, which explains how oceanic viral communities maintain high local diversity despite their limited global diversity [16].

Overall, the role and influence of viruses in the ocean are of crucial ecological and evolutionary significance. As we refine our methods and deepen our understanding, we are likely to uncover even more about the ways in which these tiny entities drive the functioning of Earth's largest ecosystem.

2.5. Research Directions

Building upon the pivotal studies discussed previously, we have established a rich understanding of marine biodiversity, microbiome composition, gene expression and its relation with biogeochemical cycles in the oceans. However, we realize there are still many dimensions left unexplored. A key focus for future research, and a primary objective of our own work, lies in the area of genomic regulation of marine microbiomes.

While Salazar et al. [14] made significant progress in studying transcript abundance, the investigation did not delve deep into the gene regulatory factors shaping transcriptomic composition. This gap becomes especially pertinent when considering that changes in gene expression predominantly define microbial community transcriptomes in non-polar waters. To bridge this gap, our focus shifts to understanding the underlying regulatory mechanisms, notably through the analysis of transcription factor binding motif abundances.

Transcription factors have a central role in regulating gene expression, influencing microbial adaptability through their binding abundance and variety. Given the variations in microbial behaviors across diverse geographical and environmental gradients, it prompts essential questions: Are specific transcription factors more common in warmer waters, aiding in the observed gene diversity? How do they function in colder, polar regions? Answering these can provide clarity on marine microbial adaptability and their anticipated reactions to upcoming climatic changes.

Additionally, with temperature's evident impact on microbial biodiversity [11] and gene expression [14], it's crucial to assess how climate change, especially ocean warming, might alter transcription factor binding abundance abundance in marine microbiomes. Exploring these potential shifts can help predict how microbial adaptability might evolve, thereby affecting the overall health of our oceans.

Chapter 3

Data description

The upcoming chapter is dedicated to outlining the comprehensive range of environmental (abiotic) and biological (biotic) data used to support our research efforts. We initiate this examination by presenting an overview of the environmental data that are integral elements of our metadata. This context-rich framework aids in understanding the backdrop against which our study unfolds.

Following this, we concentrate on the crux of our investigation, namely, the biological signals represented by transcription factor binding motifs abundance (TFBM) associated to a family of 88 transcriptin factors (TF). This exploration involves not just the decoding of these biological indicators but also entails a discussion on the multifaceted methods employed to measure TFBM abundances. We further highlight these diverse methodologies by drawing comparisons with our preliminary findings, offering us an opportunity to gauge their relative precision and effectiveness.

Through this chapter, our objective is to establish a well-defined foundation upon which subsequent analyses are built, thereby ensuring a comprehensive understanding of our data sets.

3.1. Environmental data

The following section is dedicated to detailing the wealth of environmental data harvested by the Tara Oceans expedition, which underpins our investigation. This diverse array of parameters has been logically classified into distinct categories to enable a systematic and meaningful exploration in our study and subsequent analysis. This will be useful for fostering insights into the marine microbiome in the context of its surrounding milieu.

In our study, the environmental data[5] collected by the Tara Oceans expedition can be divided into the following categories:

Geographical and Physical Data: includes parameters such as Depth, Depth Mixed Layer, Residence time, Temperature, Density, Nitracline depth, Depth Min O_2, Depth Max O_2, Latitude&Longitude, Layer, Polarity (From a geographical standpoint, we define polar oceans as those located at an absolute latitude greater than 60°. Conversely, non-polar oceans are characterized as those situated at an absolute latitude less than 60°.).

Nutrient Availability Data: encompasses NO_2, NO_2+NO_3, NO_3, PO_4, Ammonium at

[5] The one used in this work can be downloaded from https://drive.google.com/file/d/1nyCJcbptRM0UT9uI8gxsE_SPzSmGlW5F/view?usp=sharing

5m, Iron at 5m, Si and N:P ratio (ratio of Nitrogen and Phosphorus. Notably, we used NO_3 and PO_4 to calculate this).

Chemical Composition Data: comprises Oxygen, Salinity, Carbon total, Alkalinity total, CO_3 and HCO_3.

Flux Data: involves Mean Flux at 150m (the Carbon Export), Flux Attenuation and NPP (Net Primary Production)

Biological Data: is composed of Chlorophyll-a, Fluorescence and PAR.PC (Photosynthetically Active Radiation)

Ocean Dynamics Data: refers to measures of Lyapunov exponents, Brunt–Väisälä frequency, Okubo–Weiss parameter and Gradient Surface temperature (SST).

Following this brief overview, we will dive deeper into each of these categories. Detailed descriptions of these variables will be presented in the upcoming section to have a clear enviromental knowledge in the subsequent chapters..

3.1.1. General description

- **Layer:** It is a label dependent on the depth at which a sample is taken. This is divided into Surface (SRF), Deep Chlorophyll Maximum (DCM) and Mesopelagic zone (MES). Surface samples are found at a depth of 5m to 9m and sum up to 83 samples. Deep Chlorophyll Maximum is the region below the surface of water with the maximum concentration of chlorophyll and its samples are found at a depth of 17m to 188m, they add up to 51 samples. Mesopelagic zone begins at the depth where only 1% of incident light reaches and ends where there is no light; the depths of this zone are between approximately 200m to 1000m, but our samples range between 250m to 1000m[6]; there are a total of 39 samples of these category. Here is a table with the compressed information:

Table 3.1: Summary of Sample Layer, Sample Count, and Depth Range

Layer	Sample count	Depth range (m)
SRF	83	5 - 9
DCM	51	17 - 188
MES	39	250 - 1000

Additionally, the combination of samples present in the Surface (SRF) and the DCM zone is referred to as the Epipelagic (EPI) zone, resulting in a separation by pelagic zone when Mesopelagic (MES) zone is considered. Table 3.2 synthesizes the previous information.

Table 3.2: Summary of Sample Pelagic Zone, Sample Count, and Depth Range

Pelagic Zone	Sample Count	Depth range (m)
Epipelagic	134	5 - 200
Mesopelagic	39	250 - 1000

[6] except for sample *TSC155* which is at a depth of 177 meters and is considered as a mesopelagic sample

- **Polar:** This feature has to do with the geographical location where the samples were taken. Samples labeled as 'polar' are samples obtained both on the thawed coasts of the Arctic polar circle and near Antarctica (Latitude further south than -60°). While 'non-polar' samples are samples taken in the open ocean or near some continent. The former ones add up to 42 and the last ones 131. Table 3.3 summarizes the above.

Table 3.3: Summary of Sample Polarity and Sample Count

Polarity	Sample count
Polar	42
Non Polar	131

- **Latitude&Longitude:** These represent the geographical coordinates where the sample was taken.

- **Depth.nominal:** This represents the approximate depth at which the sample was taken.

- **Ocean.region:** It is a categorical variable that refers to the various oceanic regions around the world, such as the North Atlantic Ocean, South Pacific Ocean, Arctic Ocean, and others. It is used to classify samples or measurements according to their geographic location within the global ocean system. Table 3.4 shows the quantity of samples per Ocean.

Table 3.4: Number of samples from each ocean region

Ocean Region	Number of Samples
[**AO**] Arctic Ocean	35
[**SPO**] South Pacific Ocean	31
[**IO**] Indian Ocean	27
[**NAO**] North Atlantic Ocean	23
[**SAO**] South Atlantic Ocean	19
[**NPO**] North Pacific Ocean	16
[**MS**] Mediterranean Sea	12
[**RS**] Red Sea	6
[**SO**] Southern Ocean	4

The samples are shown in the Global Map shown in Figure 3.1

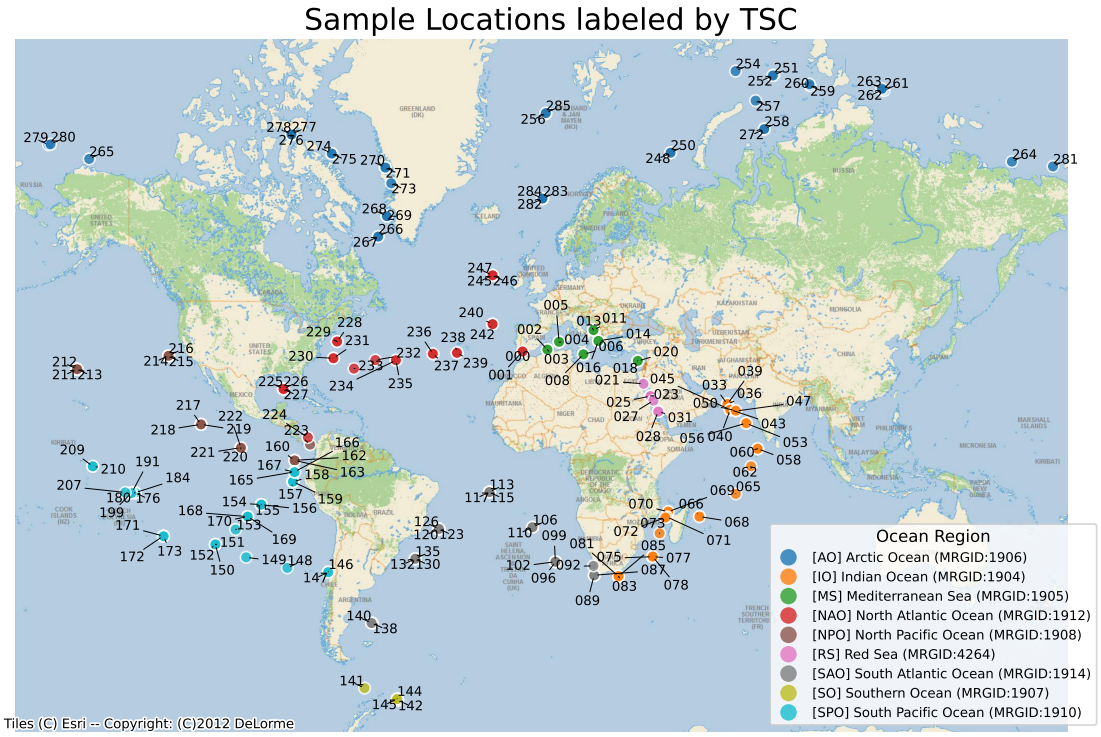


Figure 3.1: This dataset comprises samples collected from various locations across the oceans. Importantly, these samples are labeled using a Total Station Count (TSC) system, which treats each depth as a unique sample, separate from other depths at the same station.

Clearly, most of the samples are concentrated in the north. Additionally, the Gulf Stream is well-sampled helping this study for further conclusions on climate change concerns [18, 19].

- **Bioprovinces:** Bioprovinces arise from the delineation of environmental niches for bacteria size fraction organisms (0.22 - 3 μm). These niches were then extrapolated to inform global ocean biogeography, indicating the most probable province characterized by signature genomes. The environmental niches were determined and verified using machine learning models as described by Fremont [13]. Predictors used for this were sea surface temperature (SST), salinity, dissolved silica, nitrate, phosphate and iron.

  The bioprovinces (or provinces) can be visualized in Figure 3.2:

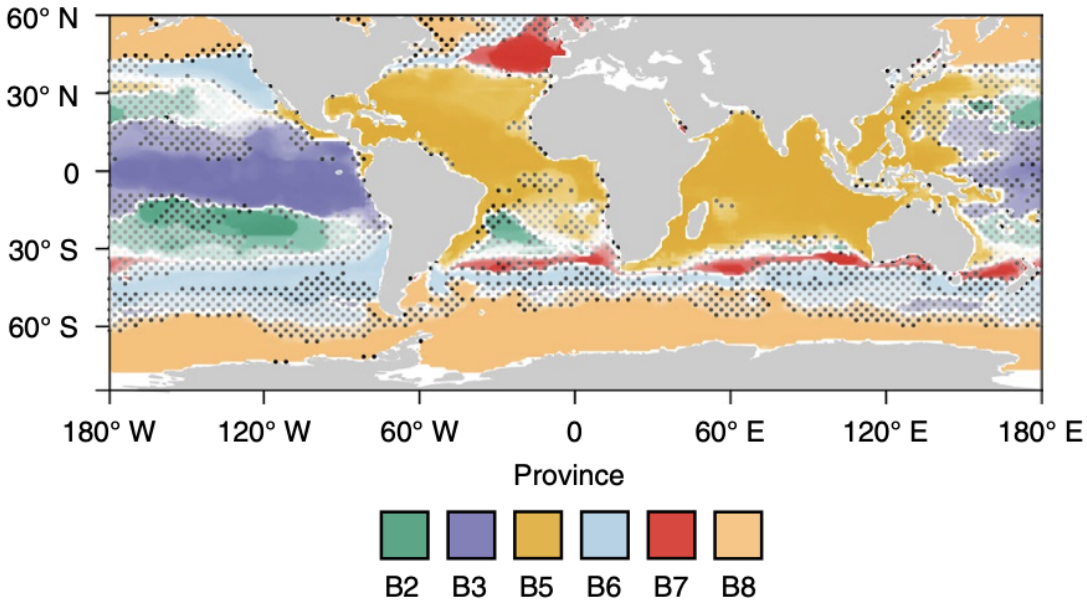


Figure 3.2: Bioprovines characterized by signature genomes. Global geographical patterns for 0.22-3 μm plankton size fractions in present day

While the author aims to understand how these genomic provinces adapt under climate change, we utilize these bioprovinces to bridge niche theory, which informs the derivation of these provinces, with genomic regulation via transcription factors.

- **Temperature:** This represents the mean of the temperature [°C] of the water at the time the sample was taken. Figure 3.3 shows a separation of temperature into quartiles across the samples.

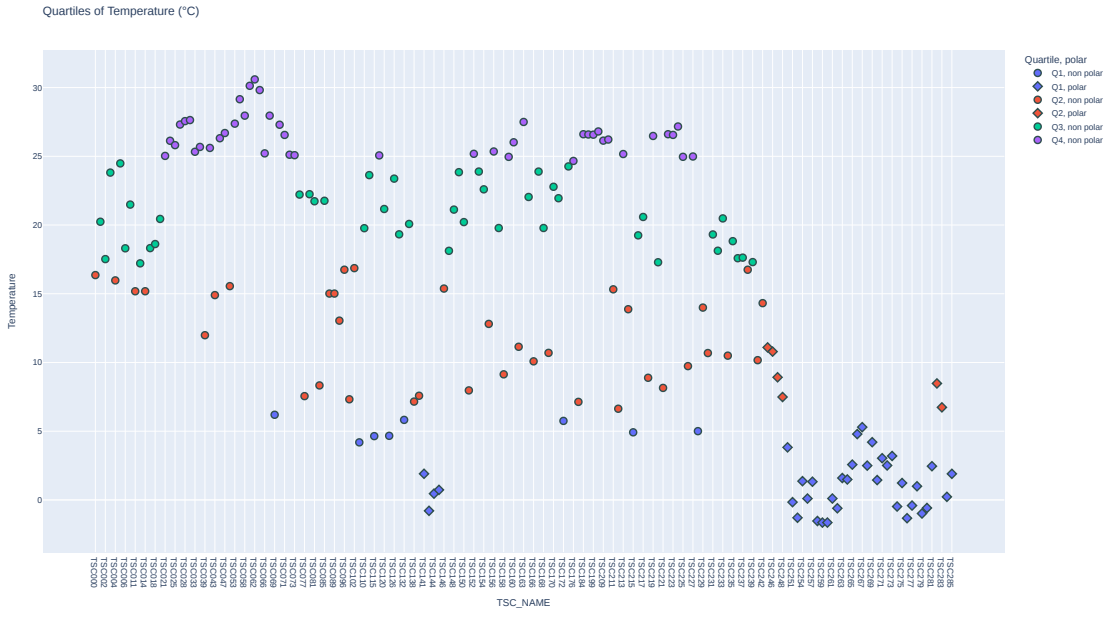


Figure 3.3: Distribution of Temperature Quartiles (°C), distinguished by polar and non-polar samples.

There is a noticeable drop in temperature in the polar zones, which is the primary attribute to consider for these areas. Distinct temperature bands are defined by the quartiles. Notably, we observe divisions at above 25°C, below 5°C, and the intermediary range. These well-defined divisions offer insights into the diversity of marine ecosystems and potential evolutions in response to possible future ocean temperature rises [20].

- **Oxygen:** This represents the mean of the oxygen [µmol/kg] of the water at the time the sample was taken. Figure 3.4 shows a separation of oxygen into quartiles across the samples.

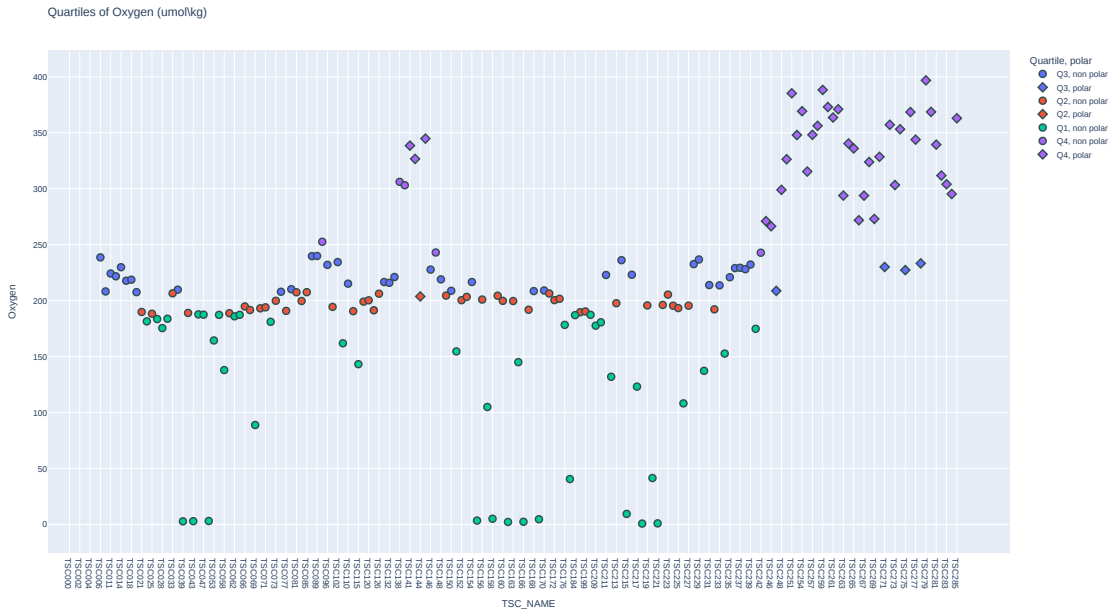


Figure 3.4: Distribution of Oxygen [µmol/kg], distinguished by polar and non-polar samples.

In contrast to the observed temperature trends, there is a notable rise in oxygen concentration within polar regions. Outside these areas, the oxygen content tends to stabilize, typically falling within a 150 to 250 [µmol/kg] range. It's worth noting that a minor fraction, roughly 5%, of the samples register extremely low or virtually zero oxygen concentrations.

- **ChlorophyllA:** ChlorophyllA represents the chlorophyll-a [mg/m^3], which is a specific form of chlorophyll used in oxygenic photosynthesis. This photosynthetic pigment is essential for photosynthesis in eukaryotes, cyanobacteria and prochlorophytes because of its role as primary electron donor in the electron transport chain. Figure 3.5 shows a separation of chlorophyll-a into quartiles across the samples.

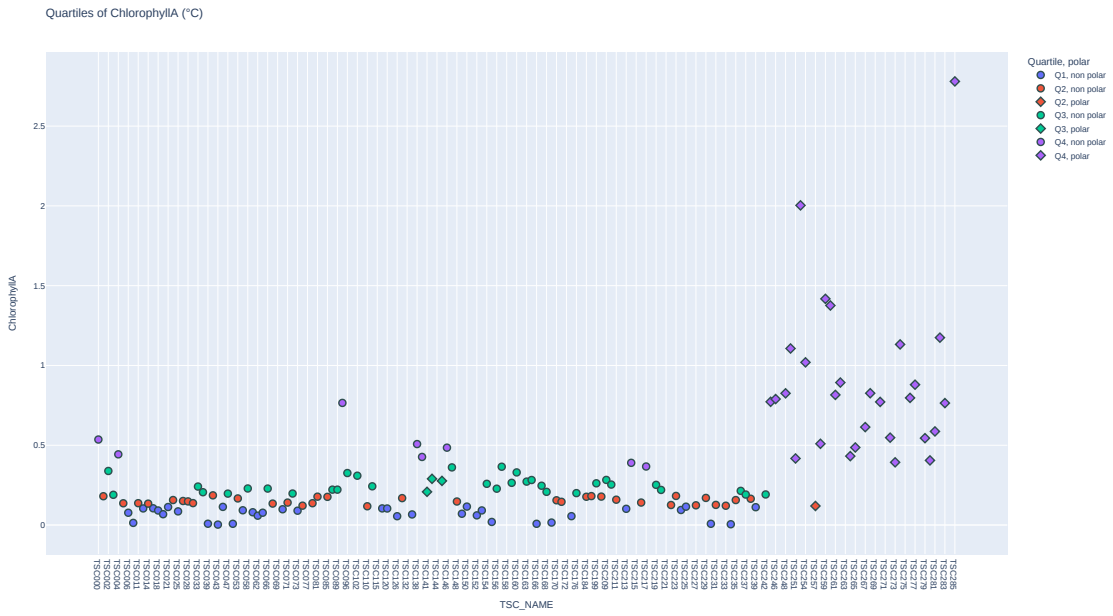


Figure 3.5: Distribution of chlorophyll-a [mg/m^3], distinguished by polar and non-polar samples.

In polar areas, both chlorophyll-a and oxygen levels increase. Chlorophyll-a plays a crucial role in photosynthesis, a process that generates oxygen. These cold, nutrient-dense polar waters nurture phytoplankton blooms that rely on chlorophyll-a. As a result, regions with elevated photosynthetic activity likely have higher oxygen concentrations, underscoring the tight interplay of marine ecosystem elements. Outside of polar regions, chlorophyll-a levels typically stay low, ranging between 0 and 0.5 [mg/m^3].

- **Salinity:** This represents the salinity of the water at the time the sample was taken, measured in [g/L]. Figure 3.6 shows a plot of this across the samples taking four quartiles.

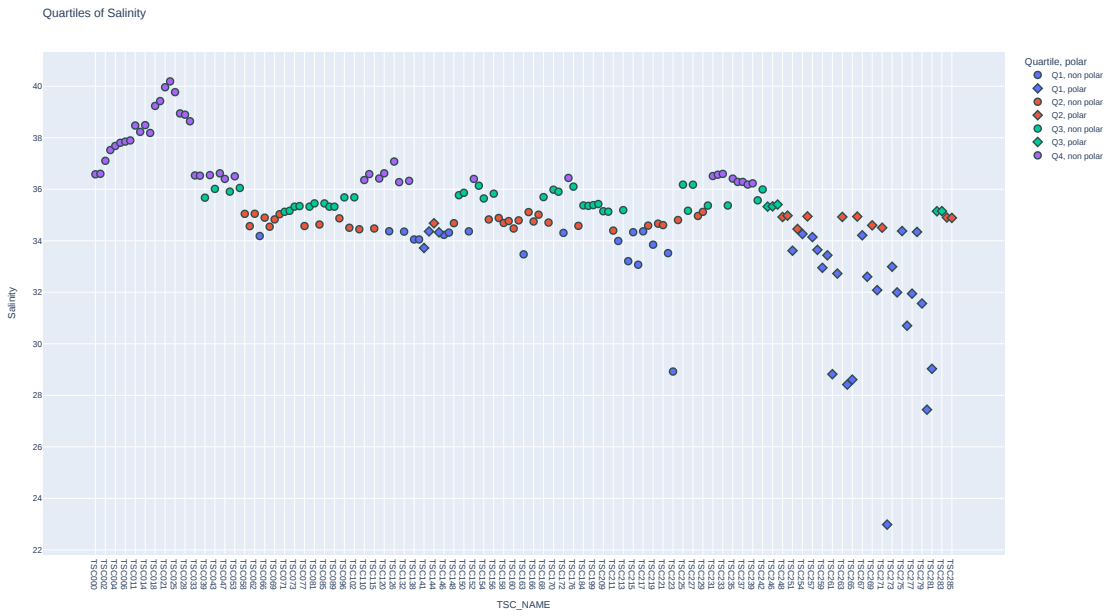


Figure 3.6: Distribution of Salinity [g/L], distinguished by polar and non-polar samples.

Salinity in polar regions exhibits a decline, which is intricately linked to the temperature behavior. As we approach the poles and temperatures drop, various factors reduce seawater salinity: 1) Melting ice introduces freshwater, which lacks salt, into the adjacent seawater, 2) Cooler temperatures slow down evaporation, leading to a lesser concentration of salinity, and 3) Enhanced precipitation further dilutes the ocean's salt concentration. In tropical regions, salinity is mostly uniform, with notable exceptions in samples from the Mediterranean and Red Sea. These regions register salinity levels exceeding the usual 34 - 36 [g/L] range, primarily due to elevated evaporation rates combined with restricted freshwater inflows..

- **NO_2:** This represents the concentration of nitrite [µmol/L] in the water at the time the sample was taken. Figure 3.7 shows a separation of the concentration into quartiles across the samples.

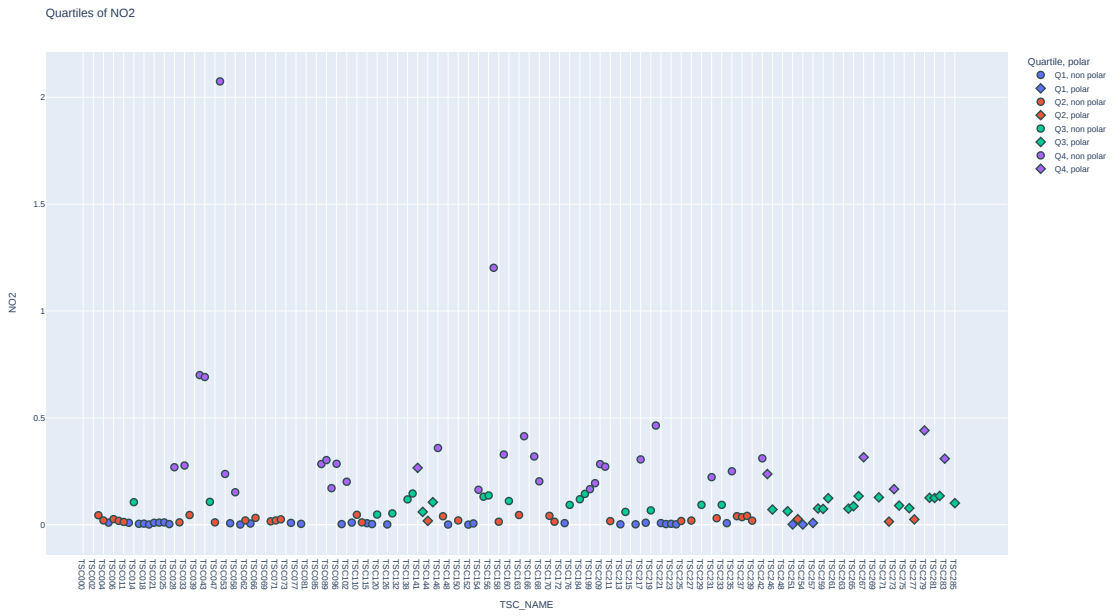


Figure 3.7: Distribution of NO_2 [µmol/L], distinguished by polar and non-polar samples.

Across the collected samples, nitrite concentrations are generally homogeneous. However, notable exceptions have been observed in a specific zone. Specifically, the three highest nitrite concentrations are associated with Indian Ocean samples, while the fourth highest comes from the Pacific Ocean. The real cause of these peaks is not clear and could be anything from a measurement error to an 'upwelling' area that raised nitrite concentrations in that location.

- **NO_3:** NO_3 is the chemical formula for nitrate [µmol/L], which is a type of inorganic nitrogen compound. Nitrates are a form of nitrogen that is easily taken up by plants and other organisms, and are an essential nutrient for the growth of many types of plants and phytoplankton. In the ocean, nitrate is derived primarily from the breakdown of nitrogen compounds, such as nitrite, ammonium, and organic nitrogen compounds, by various forms of nitrogen-fixing bacteria. These bacteria play an important role in the ocean's biogeochemical cycling, which impacts ocean ecosystem and the climate. NO_3 can also be used as a parameter to measure the nutrients and productivity of the ocean and marine ecosystem. Since Nitrate is an important nutrient for phytoplankton, higher nitrate concentration in seawater means higher productivity in phytoplankton which in turn support the marine food web. Figure 3.8 shows a separation of the concentration into quartiles across the samples.

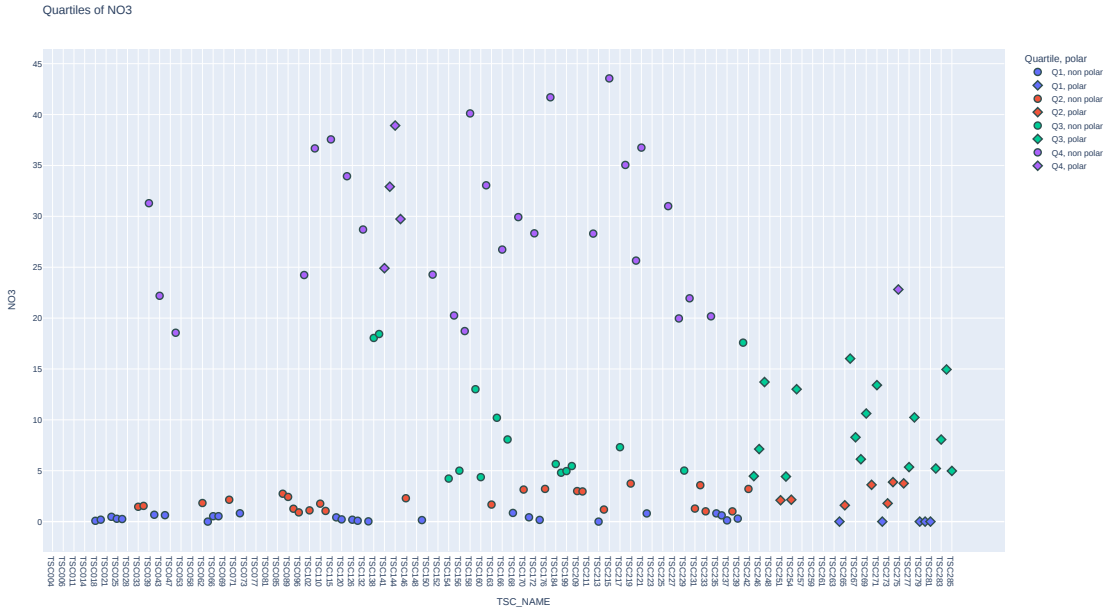


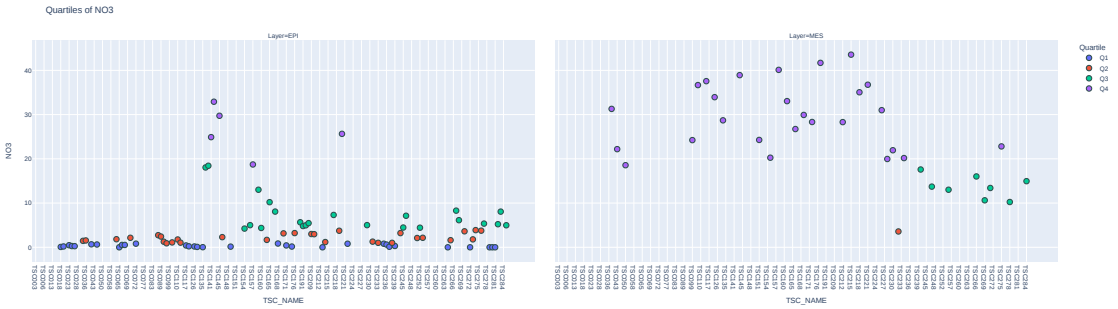
Figure 3.8: Distribution of NO_3 [µmol/L], distinguished by polar and non-polar samples.

Nitrate concentrations in the global ocean vary significantly. This variation is primarily influenced by depth changes across pelagic zones, illustrated in Figure 3.9 [21], and by differences between specific oceans [22]. Notably, Phytoplankton and other photosynthesizing organisms utilize nitrate in the surface waters as a primary nutrient source. As these organisms take up nitrate for growth in the euphotic zone [7], nitrate concentrations decrease in the surface waters.

We can also see that the surface waters of the Pacific Ocean exhibit elevated nitrate levels. This zones are known for being very productives. Furthermore, the mixing of deep, nutrient-rich waters with nutrient-scarce surface layers results in a vertical nitrate flux, influencing carbon export. Thus, nitrate acts as a pivotal marker for understanding the ocean's capacity to regulate atmospheric carbon dioxide levels.

[7] the sunlit surface layer where photosynthesis occurs

(a) SRF/DCM/MES Layer separation.



(b) EPI/MES Pelagic Zone separation.

Figure 3.9: Distribution of NO_3 quartiles stratified by ocean depth layers.

- **NO2NO3:** This represents the concentration of nitrite NO_2^- [µmol/L] and nitrate NO_3^- [µmol/L] in the water at the time the sample was taken. Figure 3.10 shows a separation of the concentration into quartiles across the samples.

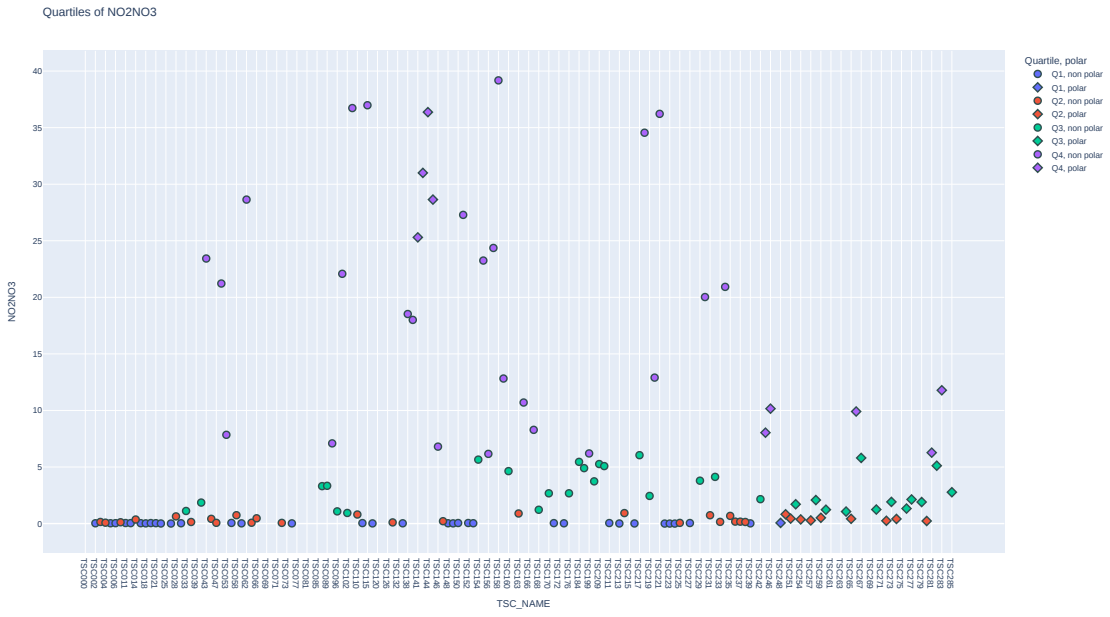
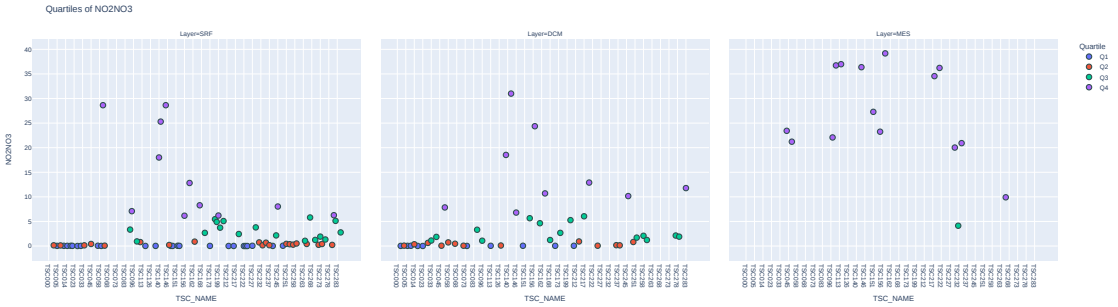


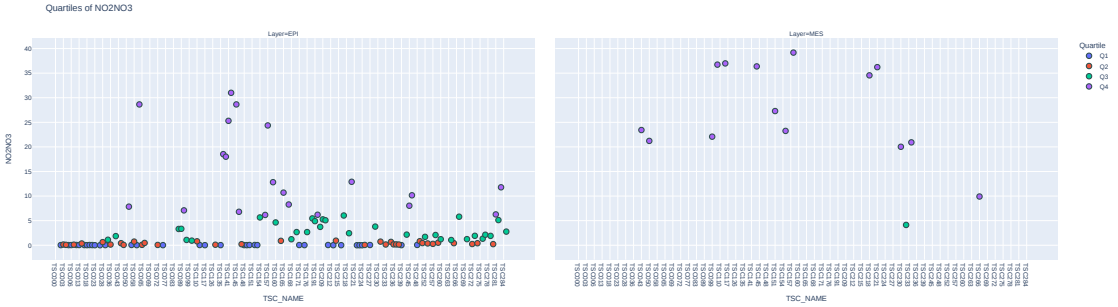
Figure 3.10: Distribution of NO_2NO_3 [µmol/L], distinguished by polar and non-polar samples.

The concentration of Nitrite+Nitrate varies across the ocean, largely due to differences

between layer specific zones and primarly affected by nitrate concentration. This is illustrated in Figure 3.11



(a) SRF/DCM/MES Layer separation.



(b) EPI/MES Pelagic Zone separation.

Figure 3.11: Distribution of NO_2NO_3 quartiles stratified by ocean depth layers.

- **PO_4:** This represents the concentration of phosphate [µmol/L] in the water at the time the sample was taken. Figure 3.12 shows a separation of this concentration into quartiles across the samples.

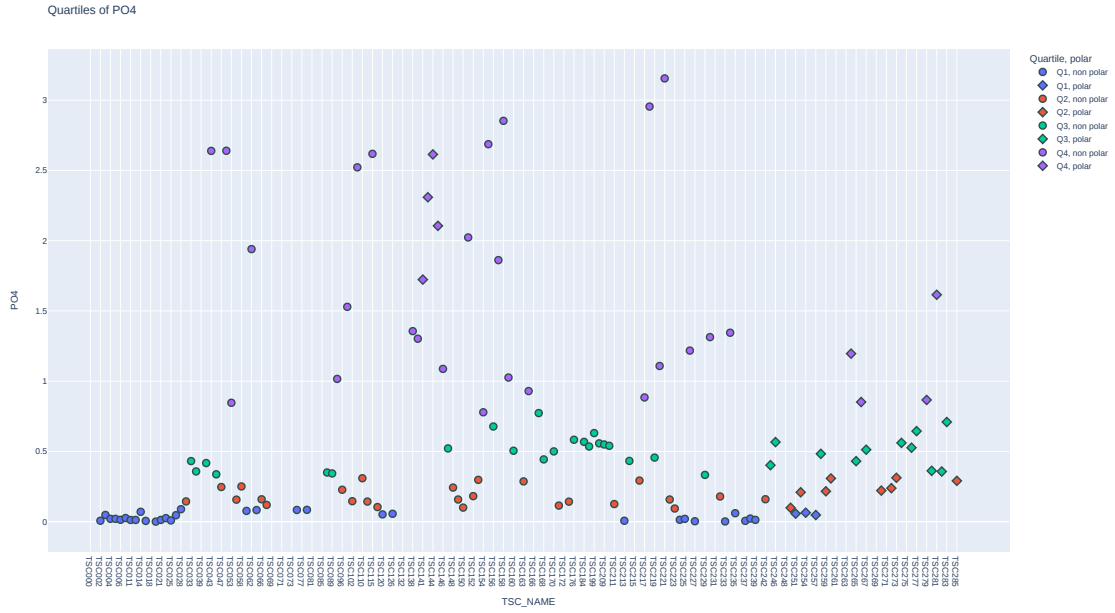
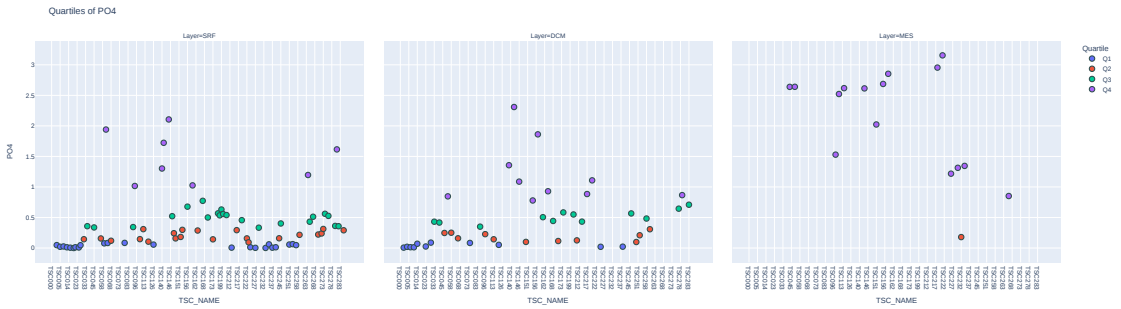
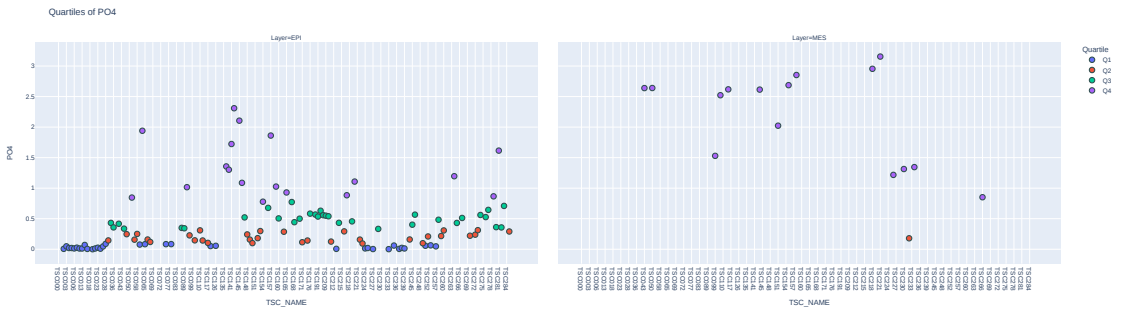


Figure 3.12: Distribution of PO_4 [µmol/L], distinguished by polar and non-polar samples.

Certain samples exhibit elevated phosphate concentrations. This heightened level is influenced by the collection depth, illustrated in Figure 3.13, and the ocean from which they originate [22]. Specifically, the surface waters of the Pacific Ocean have unusually high phosphate concentrations. Furthermore, deeper waters generally contain increased nutrient content [21].



(a) SRF/DCM/MES Layer separation.



(b) EPI/MES Pelagic Zone separation.

Figure 3.13: Distribution of PO_4 Quartiles Stratified by Ocean Depth Layers.

- **Ammonium.5m:** This represents the concentration of ammonium [µmol/L] in the water at a depth of 5 meters at the time the sample was taken. It plays several important roles, both biologically and chemically. Firstly, it is a crucial nitrogen source for phytoplankton, the foundation of the marine food chain, as they preferentially absorb it due to its lower assimilation energy. This process not only fuels their growth but also recycles nitrogen within the marine system. Secondly, ammonium features prominently in the marine nitrogen cycle, being produced when bacteria break down organic matter. This can later be used by marine plants or transformed into nitrate through nitrification. Its presence also interacts with the marine chemical environment; the balance between ammonium and its counterpart, ammonia (a toxic compound), is influenced by the pH of seawater. High ammonium concentrations in surface water can hint at significant organic decay, possibly pointing to marine ecosystem imbalances like eutrophication. Furthermore, the levels of ammonium can also affect how marine organisms uptake other nutrients, like nitrate. Monitoring ammonium is therefore essential for understanding ocean health and biogeochemical interactions.

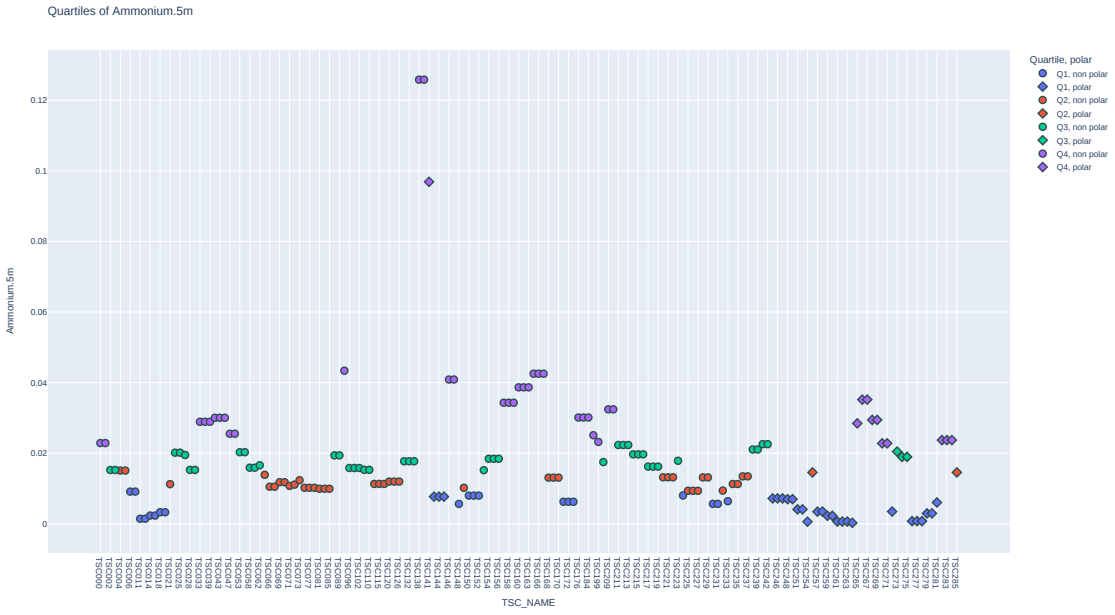


Figure 3.14: Distribution of Ammonium at 5m [µmol/L], distinguished by polar and non-polar samples.

In the figure, distinct peaks of ammonium concentration are evident in samples from the South Pacific and the Southern Ocean. Outside of these peaks, concentrations predominantly fall between 0 and 0.05 µmol/L. The lowest concentrations are observed in the far northern latitudes, aligning with polar samples. However, an exception is seen in samples near Greenland, which exhibit slightly elevated concentrations, just above 0.04 µmol/L.

- **Iron.5m:** This represents the concentration of Iron (Fe) [µg/L] in the water at a depth of 5 meters at the time the sample was taken. This element plays an important role in the ocean's biogeochemistry and ecosystem, in fact, most life forms are heavily dependent on iron and phytoplankton, with their Fe-rich photosynthetic apparatus, have significantly

higher Fe demands as opposed to their heterotrophic counterparts, and evidence of this is that intracellular Fe content of these phototrophic microorganisms is 4–6 orders of magnitude greater than the Fe concentrations in their surroundings. Unfortunately, for phytoplankton, readily bioavailable iron in aquatic environments is vanishingly scarce [23]. Iron is also an important element in the formation of marine sediments, such as iron-rich clay minerals and iron oxides. These sediments can help to remove carbon from the ocean and to store it in the seafloor for long periods of time. Figure 3.15 shows a separation of the measurements into quartiles across the samples.

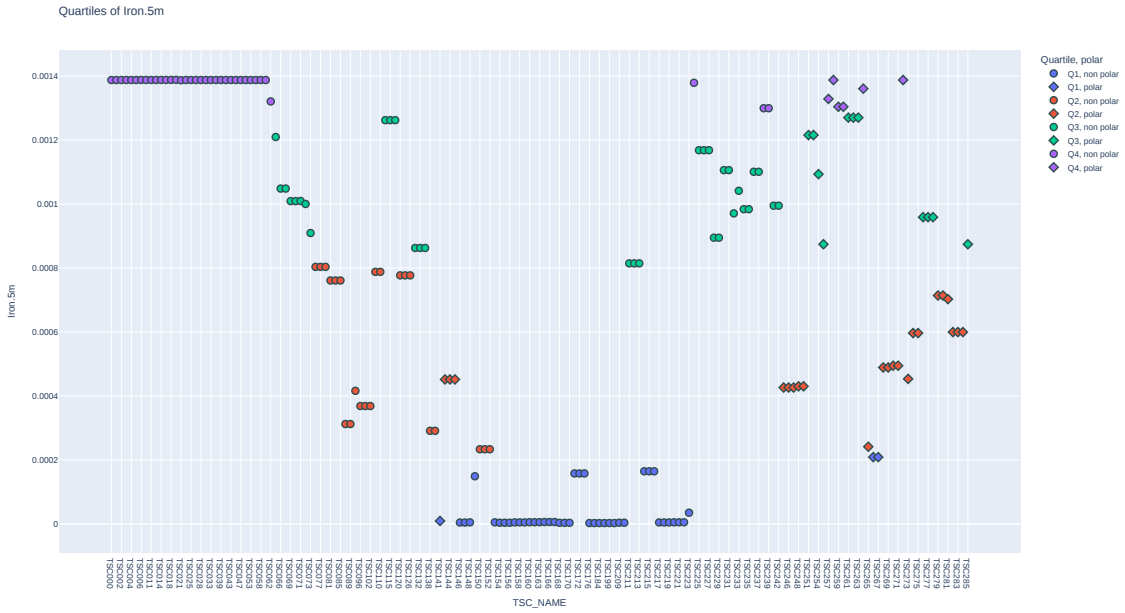


Figure 3.15: Distribution of Iron at 5m [µg/L], distinguished by polar and non-polar samples.

Certain areas, particularly the surface waters of the equatorial Pacific, exhibit deficient iron levels. This region is notably recognized for its low iron concentrations, a factor that has significant implications for marine ecosystems and their dynamics [22].

- **Alkalinity.total:** Total Alkalinity quantifies the water's capacity to counteract acids, effectively preventing a decrease in pH. It is conventionally gauged by concentrations of bicarbonate, carbonate, and hydroxide ions. Essentially, it is a testament to the water's resistance against pH drops. In marine ecosystems, alkalinity plays a pivotal role in stabilizing pH and providing a buffer against escalating acidity. This is crucial for carbonate chemistry, which establishes the saturation state of calcium carbonate minerals—foundational for coral reefs, shells, and various marine structures. Monitoring alkalinity levels is imperative, especially given the rise in oceanic CO2 absorption stemming from heightened atmospheric CO2 concentrations. Figure 3.16 shows a separation of the amount of Alkalinity into quartiles across the samples.

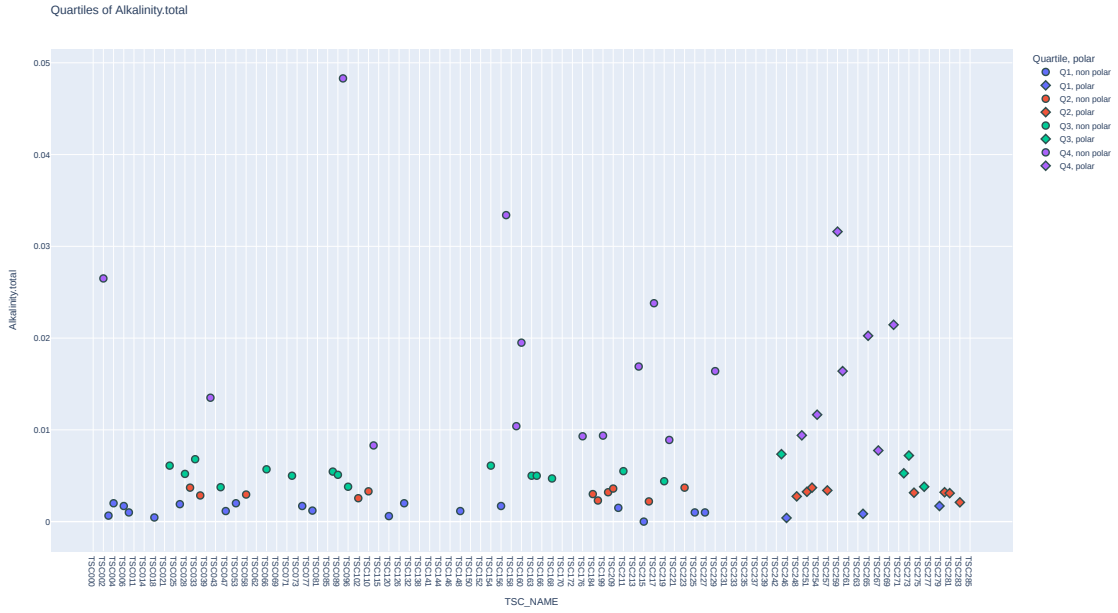


Figure 3.16: Distribution of Alkalinity, distinguished by polar and non-polar samples.

On average, alkalinity levels are below 0.01 eq/L. However, exceptions are observed in some samples from the Pacific and Atlantic. Notably, no alkalinity measurements are available for mesopelagic samples.

- **CO_3:** CO_3^{2-} is the chemical formula for carbonate, which is a type of inorganic carbon compound that is present in seawater. The concentration of these ions in seawater can vary depending on a number of factors, such as temperature, salinity, and the amount of CO_2 dissolved in the water. The concentration of CO_3^{2-} in seawater, along with and HCO_3^- ions, is closely related to the pH and alkalinity of the water, and changes in their concentrations can affect the water's acidity. Carbonates also play a critical role in marine ecology. Carbonate ions are the building blocks for calcium carbonate ($CaCO_3$) which is the main component of shells, corals and other marine organism's skeletons and exoskeletons. In addition, changes in the amount of carbonate ions in seawater can have a significant impact on the Earth's climate. Rising atmospheric CO_2 concentrations can lead to increased amounts of CO_2 dissolving in seawater, which can decrease the pH and saturation state of calcium carbonate minerals. This can have major implications for marine organisms and ecosystems, such as coral reefs and other calcifiers, and on global climate regulation. Figure 3.17 shows a separation of the amount of CO_3 into quartiles across the samples.

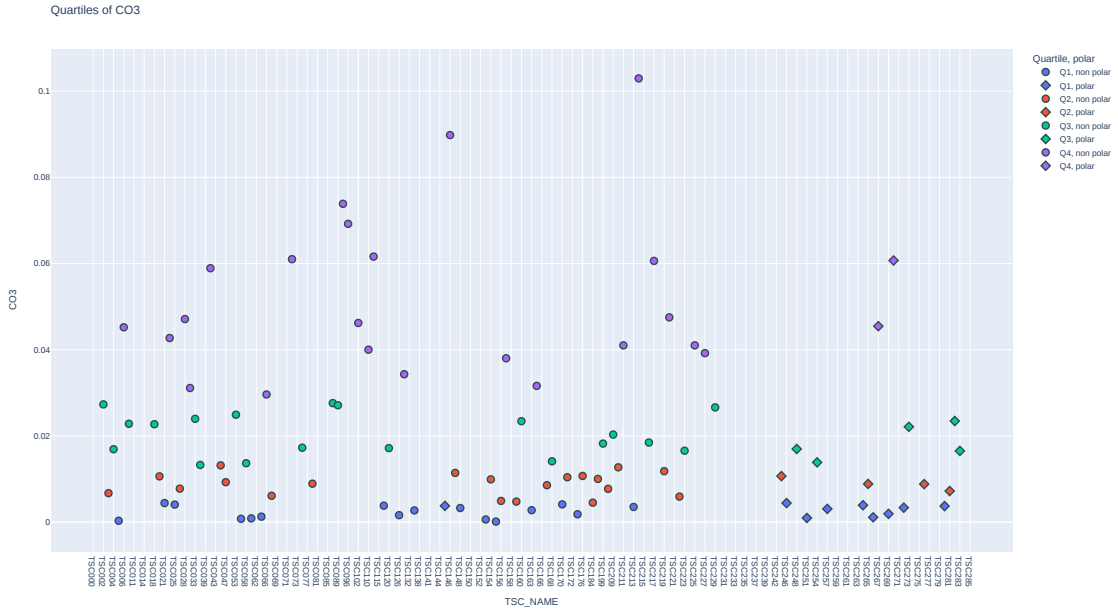
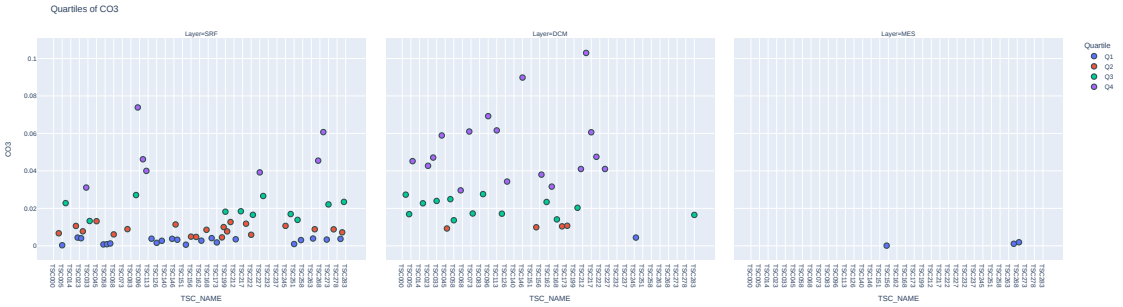
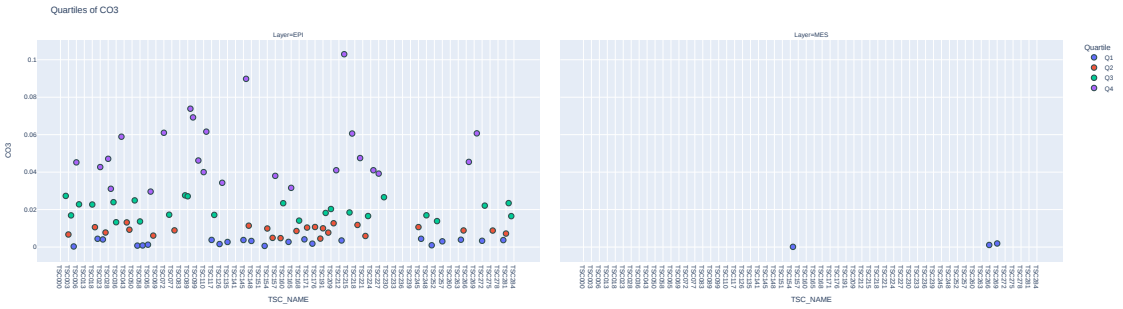


Figure 3.17: Distribution of CO_3, distinguished by polar and non-polar samples.

We can separate the samples by 'Layer' and 'Pelagic Zone', obtaining the graph shown in Figure 3.18



(a) SRF/DCM/MES Layer separation



(b) EPI/MES Pelagic Zone separation

Figure 3.18: Distribution of $CO_3$3 Quartiles Stratified by Ocean Depth Layers

Given the few samples collected for mesopelagic samples, we can not conclude about the impact of depth into the carbonate concentration. However,

- **HCO_3:** HCO_3 is the chemical formula for bicarbonate, which is a type of inorganic carbon compound that is present in seawater. Bicarbonate ions (HCO_3^-) are a byproduct of the combination of carbon dioxide (CO_2) and water (H_2O). In seawater, bicarbonate ions are one of the main forms of dissolved inorganic carbon along with carbonate ions (CO_3^{2-}), and they play an important role in the ocean's carbon cycle. Bicarbonate ions help to buffer changes in acidity (pH) in seawater, which helps to maintain a relatively constant pH. They do this by reacting with protons (H^+) to form carbonic acid (H_2CO_3), which can then dissociate into CO_2 and H_2O. This reaction helps to neutralize acids and prevent large changes in pH. Bicarbonate ions also play a key role in the marine carbonate system, because they are involved in many chemical reactions that affect the pH and alkalinity of seawater and influence the saturation state of calcium carbonate minerals, as we mentioned in the previous subsection. Figure 3.19 shows a separation of the amount of HCO_3 into quartiles across the samples.

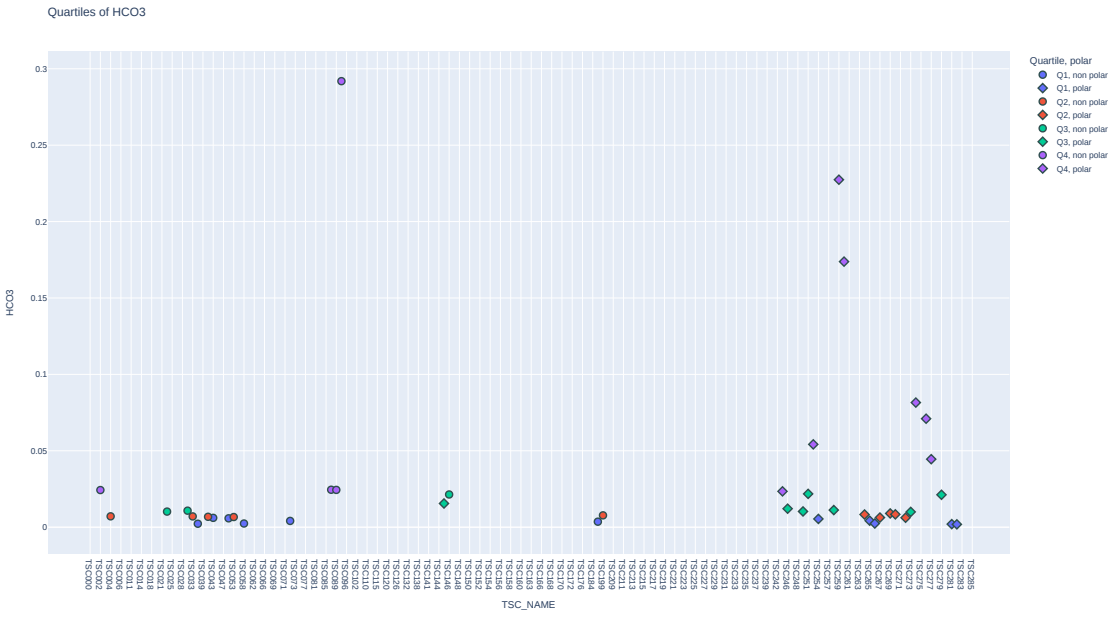


Figure 3.19: Distribution of HCO_3, distinguished by polar and non-polar samples.

- **N:P ratio:** The N:P ratio refers to the Nitrate-to-Phosphate ratio, which compares the concentration of nitrate (NO_3^-) to the concentration of phosphate (PO_4^{3-}) in seawater. The ratio is used as an indicator of the availability of different nutrients for phytoplankton growth in the ocean. Phytoplankton, the base of the marine food web, requires both nitrate and phosphate to grow. However, they typically require more nitrate than phosphate. In well-fertilized waters, the ratio of nitrate to phosphate is usually around 16:1, indicating that there is enough nitrate to support phytoplankton growth, but not enough phosphate to limit it. However, in nutrient-poor waters, the ratio may be lower, indicating that phosphate is limiting phytoplankton growth. Similarly, in waters that are highly fertilized by human activities, such as sewage discharge or agricultural runoff, the ratio may be higher, indicating that nitrate is the limiting nutrient. Therefore, the Nitrate-to-Phosphate ratio can be used as a valuable tool to understand the ocean ecosystem. Figure 3.20 shows a separation of this ratio into quartiles across the samples.

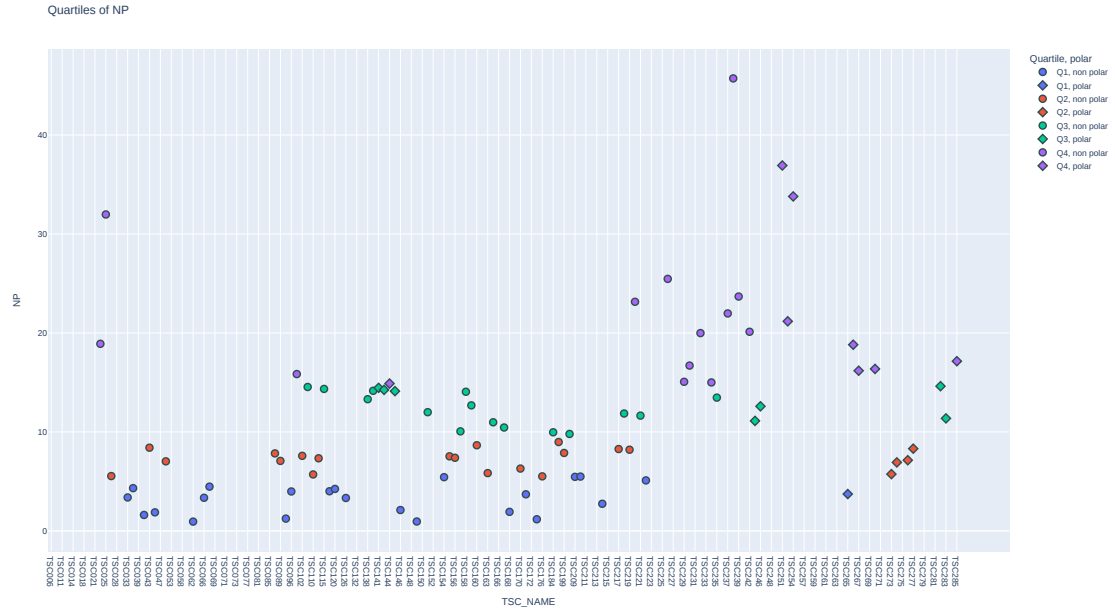
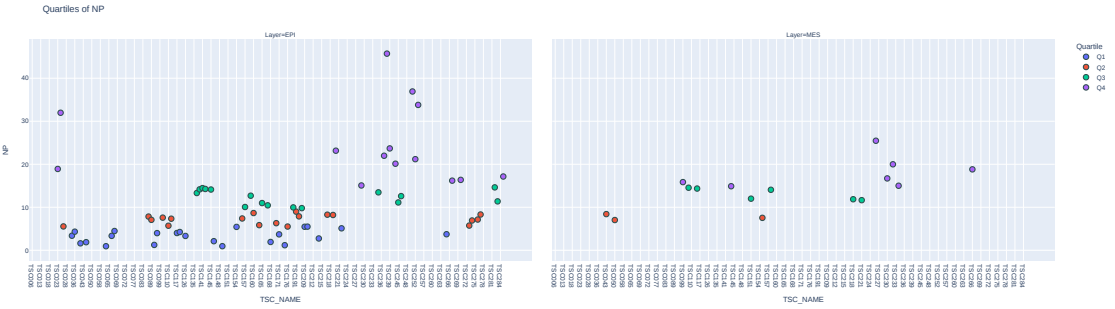


Figure 3.20: Distribution of N:P ratio, distinguished by polar and non-polar samples.

(a) SRF/DCM/MES Layer separation



(b) EPI/MES Pelagic Zone separation

Figure 3.21: Distribution of N:P ratio Quartiles Stratified by Ocean Depth Layers

In light of Figures 3.20 and 3.21, which do not immediately convey discernible insights, we will visualize the distribution of the values on a geographical map for clarity.

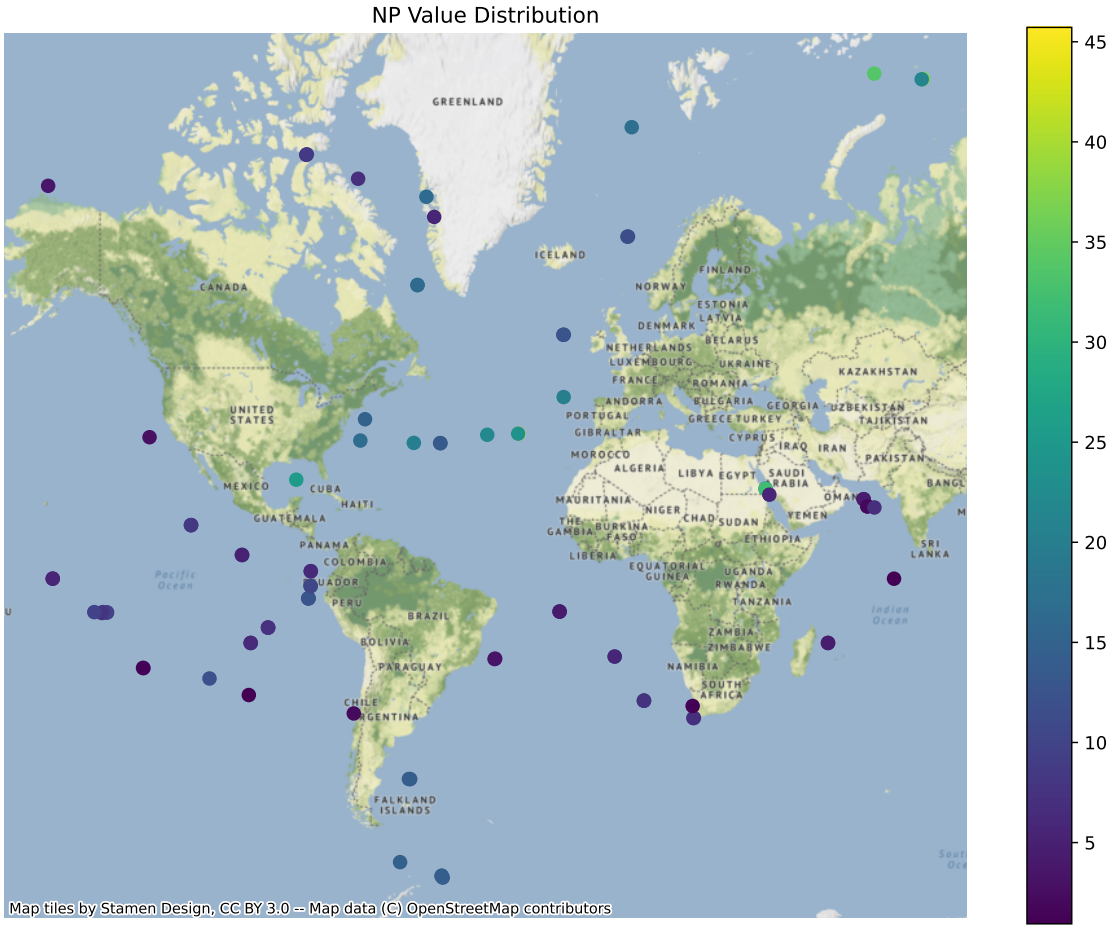


Figure 3.22: Global distribution of samples colored by their respective 'N:P' ratio values. The map showcases the spatial variability of the 'N:P' values using a gradient from the a colormap. Regions with no data points have been omitted for clarity.

We observe that samples from the North Atlantic Ocean exhibit the highest N:P ratios globally. This might be duo to phosphate depletion instead of high amount of nytrogen [24]. We also observe that in the Red Sea, two neighboring samples display stark differences in their N:P ratios; the reason for this is not clear. Additionally, it is noteworthy that the observed N:P ratios are not consistently aligned with the Redfield ratio of 16:1, as commonly cited in literature [25]. Lastly, from our observations, we cannot discern any definitive correlation between the N:P ratios and factors like polarity or depth layer.

- **Depth.Mixed.Layer:** This represents the depth of the mixed layer [m] at the time the sample was taken.
- **Lyapunov:** This represents the Lyapunov exponent [1/day] at the time the sample was taken. In oceanography, it can be used to quantify the degree of chaotic behavior in oceanic systems, such as the mixing of water masses or the dispersion of particles. This value can provide information about the degree of mixing or dispersion of the water masses in the area.
- **Si:** It stands for the amount of Silicon dioxide (SiO_2) measured in [µmol/L], also known as Silica, present in a water sample. It is an important nutrient for phytoplankton since

they use it to build their cell walls (skeletons). In fact, phytoplankton are responsible for the majority of the silica uptake in the ocean, and their uptake of silicon from seawater can significantly impact the concentration of dissolved silicon in the ocean.

We have chosen not to present a plot for Si. While Si may appear to be significant, our analysis will prioritize other environmental features.

- **Nitracline:** This represents the depth [m] of the nitracline at the time the sample was taken. The nitracline is a region in the ocean where there is a significant decrease in the concentration of nitrate with increasing depth. Nitrate is a key nutrient for marine phytoplankton growth, and its concentration is highest in the sunlit surface waters where it is produced by organic matter oxidation and phytoplankton uptake. As the surface waters become depleted of nitrate by phytoplankton growth, the nutrient is transported to deeper waters by sinking particles and ocean currents. The position and depth of the nitracline can vary depending on various factors, and it is an important feature of the oceanic ecosystem as it marks the transition between the surface waters, where primary production occurs, and the deeper waters, where nutrients are conserved and recycled.

  We have chosen not to present a plot for Nitracline. While Nitracline may appear to be significant and its description is important for drawing conclusions, our analysis will prioritize other environmental features.

- **Brunt.Väisälä:** It is a measure of the stability of a fluid to vertical displacements such as those caused by convection. More precisely it is the frequency at which a vertically displaced parcel will oscillate within a statically stable environment. So, this measurement represents the Brunt-Väisälä frequency at the time the sample was taken. It is commonly used in oceanography to study the mixing of water masses, the transport of heat and other materials in the ocean, and the distribution of plankton and other organisms.

- **Okubo-Weiss:** This is a measurement of the Okubo–Weiss parameter, which is a measure of the relative importance of deformation and rotation at a given point. This is widely applicable in fluid properties particularly in identifying and describing oceanic eddies. It is commonly used in oceanography to study the dispersion of particles, nutrients and pollutants in the ocean. It can also be used to study the spread of plankton and fish larvae, the movement of oil spills, the transport of pollutants and other materials in the ocean, since it can detect ocean eddies.

- **Residence time:** It measures the amount of time it takes for a substance to be completely replaced by new water in a particular area of the ocean. This can be affected by factors such as ocean currents and mixing processes, and can vary depending on the location and depth of the water. In general, residence time in the open ocean is much longer than in coastal areas or estuaries.

- **PAR.PC:** It refers to the amount of Photosynthetically Available Radiation (PAR) measured in [Einstein$\cdot$m$^{-2}\cdot$day^{-1}]. It is a measurement of the amount of light each cell receives on average per day. Phytoplankton are the primary producers that harness PAR to fuel photosynthesis. The availability of PAR thus influences the primary productivity of oceans and freshwater bodies, which in turn affects the entire aquatic food web.

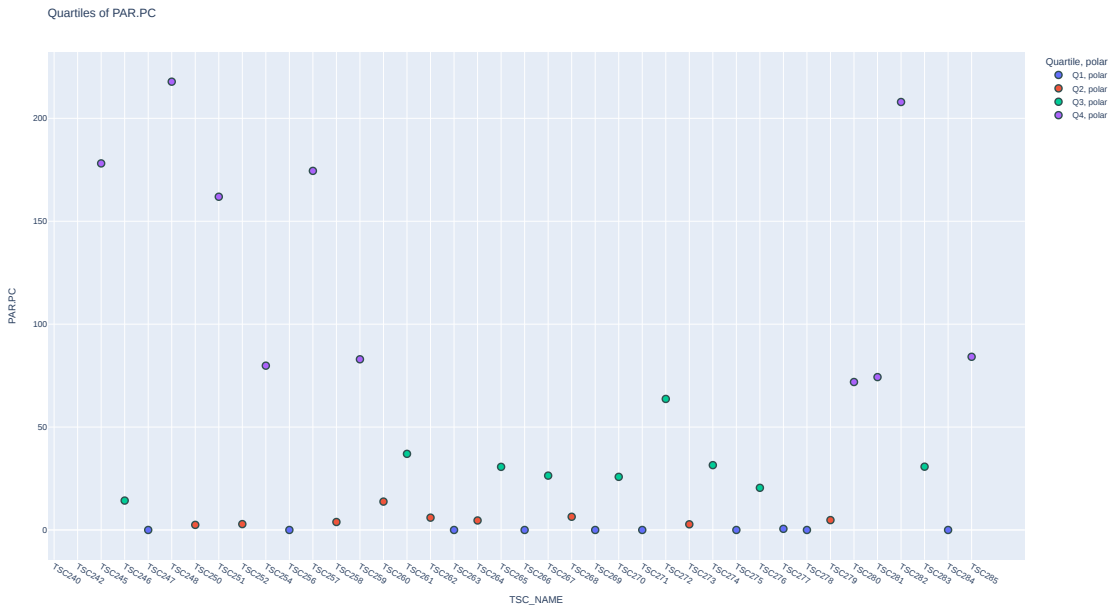


Figure 3.23: Quartiles of PAR.PC with a Polarity symbol.

We only have data for the Arctic samples, which can be better deduced in Figure 3.24:

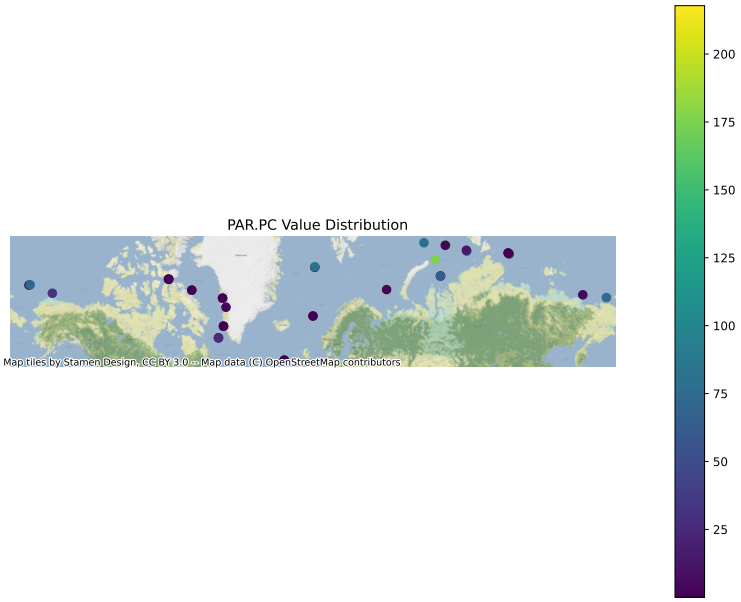


Figure 3.24: Global distribution of samples colored by their respective 'PAR.PC' values. The map showcases the spatial variability of the 'PAR.PC' values using a gradient from the a colormap.

Figure 3.24 illustrates that samples from coastal zones tend to display the lowest levels of PAR.PC. From an ecological perspective, examining its correlation with other biologically-related environmental variables is of significant interest. In particular, understanding its relationship with phytoplankton blooms, which are often associated with levels of chlorophyll-a, becomes especially intriguing.

- **Gradient.Surf.Temp(SST):** It measures the Sea Surface Temperature Gradient [°C/100km]. It is particularly important as it is one of the driver for defining environmental niches with aim of delimiting ocean biogeographies [13]
- **Fluorescence:** Fluorescence refers to the phenomenon where a substance absorbs light at one wavelength and then emits light at a longer wavelength. In the context of seawater, fluorescence is often used to measure the concentration of dissolved organic matter (DOM) and phytoplankton pigments. Measurements of fluorescence can be used to estimate phytoplankton biomass and primary productivity, which is the rate at which phytoplankton convert light energy into organic matter. Figure 3.25 shows a separation of the amount of Fluorescence into quartiles across the samples. Sample 'TSC272' was dropped for a better visualization.

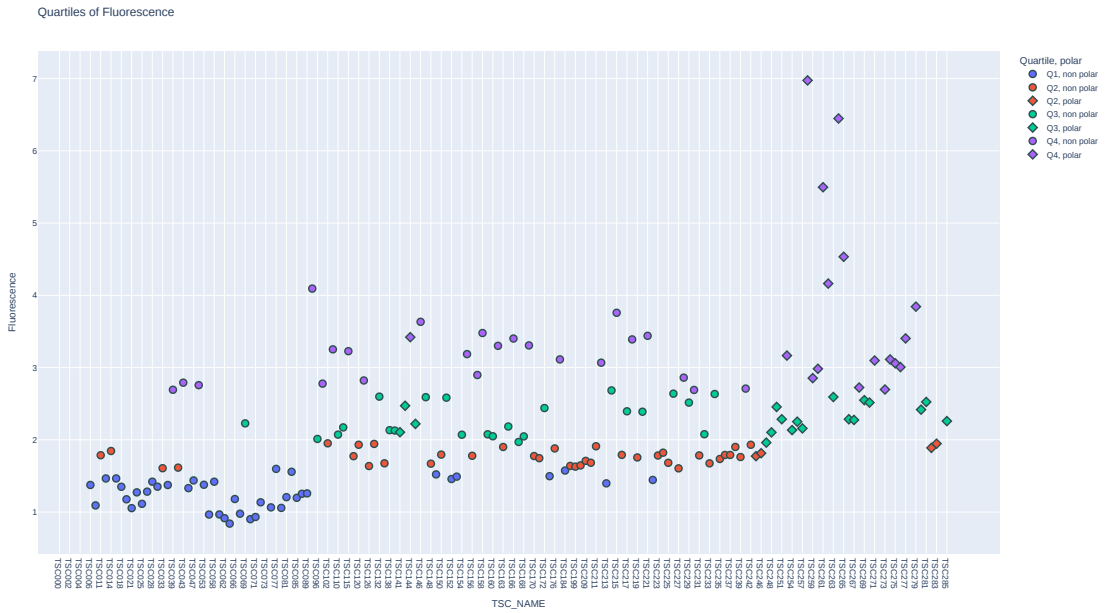


Figure 3.25: Distribution of Fluorescence, distinguished by polar and non-polar samples.

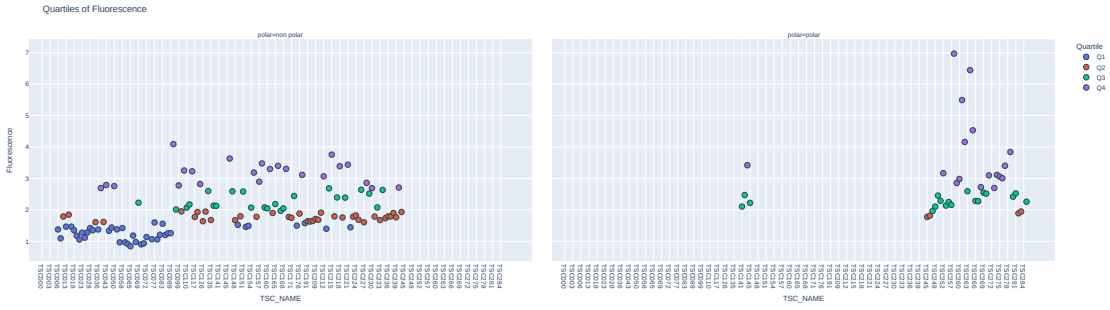


Figure 3.26: Quartiles of Fluorescence, Polar Faceted.

From Figure 3.25, there is a noticeable gradient as one approaches the polar oceans. Furthermore, a significant portion of the first quartile is observed in samples from the Mediterranean Sea, Red Sea, and the African coast. Figure 3.26 confirms elevated fluorescence levels in polar samples. This increase can be attributed to two primary factors:

1) The availability of iron, which promotes phytoplankton growth, and 2) Extended daylight in polar regions, where the sun remains above the horizon for prolonged periods, allowing phytoplankton to sustain photosynthesis for longer durations.

- **Density:** It is a measure of the Sigma-T (σ_T), which is a quantity used in oceanography to measure the density of seawater at a given temperature. It is defined by the equaiton $\sigma_T = \rho(S,T) - 1000[\text{kg/m}^3]$, where $\rho(S,T)$ is the density of a sample of seawater at temperature T and salinity S, measured in $[\text{kg/m}^3]$, at standard atmospheric pressure.

- **Mean Flux at 150m:** This variable (**Carbon Export**) refers to the process by which carbon, initially in the form of CO_2, is transformed into organic carbon through photosynthesis by marine organisms such as phytoplankton. This organic carbon then becomes part of sinking particles that are eventually transported to the deep ocean, where the carbon is sequestered or stored. The process is a key component of the biological carbon pump, which helps regulate the global carbon cycle and contributes to mitigating the impacts of climate change. It is measured in $[\text{mg C} \cdot \text{m}^{-2} \cdot \text{day}^{-1}]$. Figure 3.27 shows a separation of the amount of Carbon Export Flux into quartiles across the samples.

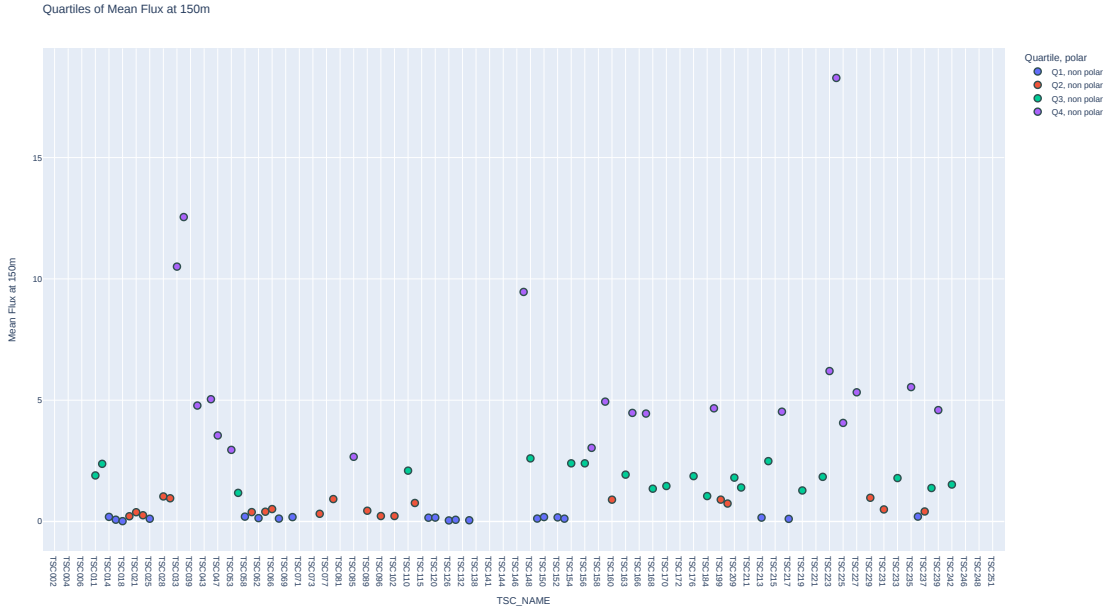


Figure 3.27: Distribution of Carbon Export.

We note that for polar samples, there is no available data on carbon export. We also note that there are no samples for the mesopelagic zones, since it is the carbon pump we are tryind to measure.

The Figure 3.28 encapsules better the data:

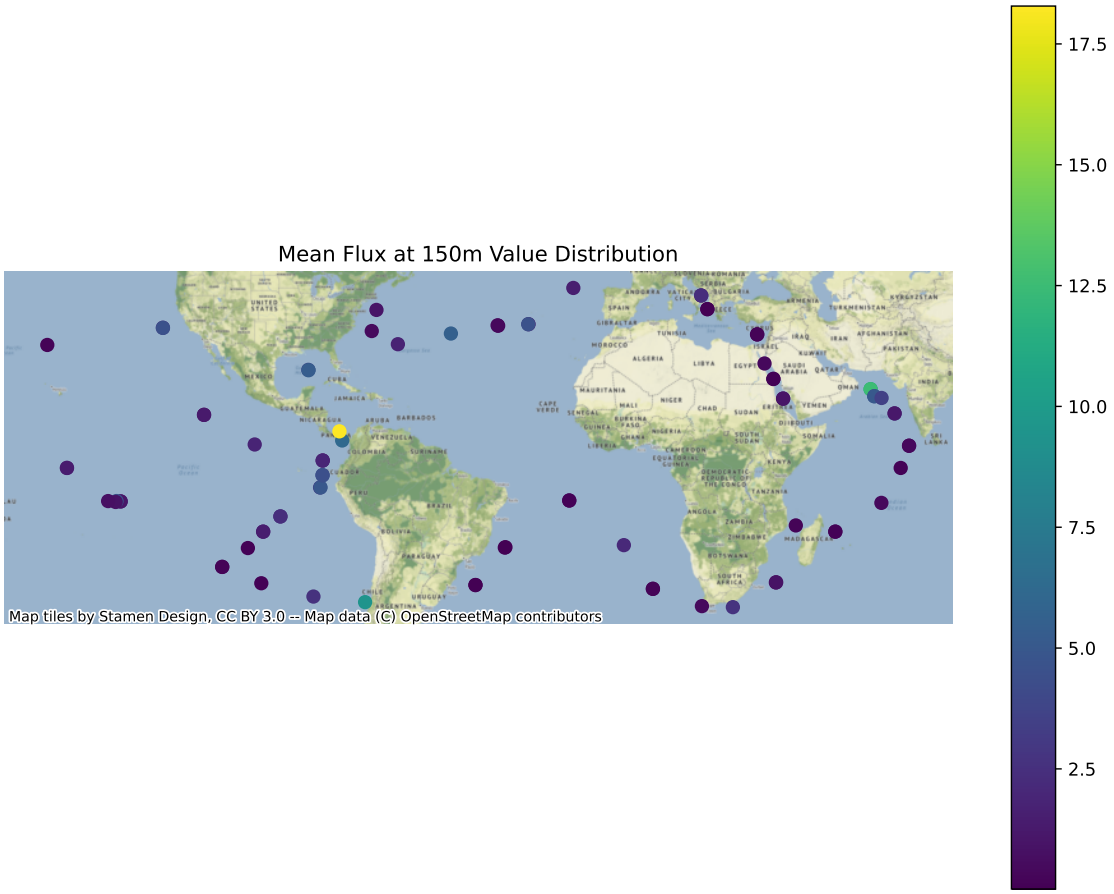


Figure 3.28: Global distribution of samples colored by their respective 'Carbon Export Flux' values. The map showcases the spatial variability of the 'Carbon Export' values using a gradient from the a colormap.

We can observe that out to sea, the carbon pump functions better than in coastal regions, except for some samples present in the Indian Ocean.

- **Flux Attenuation:** It refers to the reduction in the amount of particulate organic carbon (POC) sinking from the surface to the deeper parts of the ocean. This process is part of the ocean's biological carbon pump, which is crucial for carbon sequestration and climate regulation. The process starts with photosynthesis by phytoplankton in the upper layers of the ocean, converting carbon dioxide into organic matter. When these phytoplankton die or are consumed, the waste products sink, exporting carbon from the surface to deeper waters. However, much of this carbon is consumed by deep-sea organisms in a process called remineralization, causing the attenuation, or decrease, of the carbon flux with depth. The efficiency of this biological carbon pump can be influenced by various factors and is a key area of research in understanding the ocean's role in the global carbon cycle. Figure 3.29 shows a separation of the amount of Flux Attenuation into quartiles across the samples.

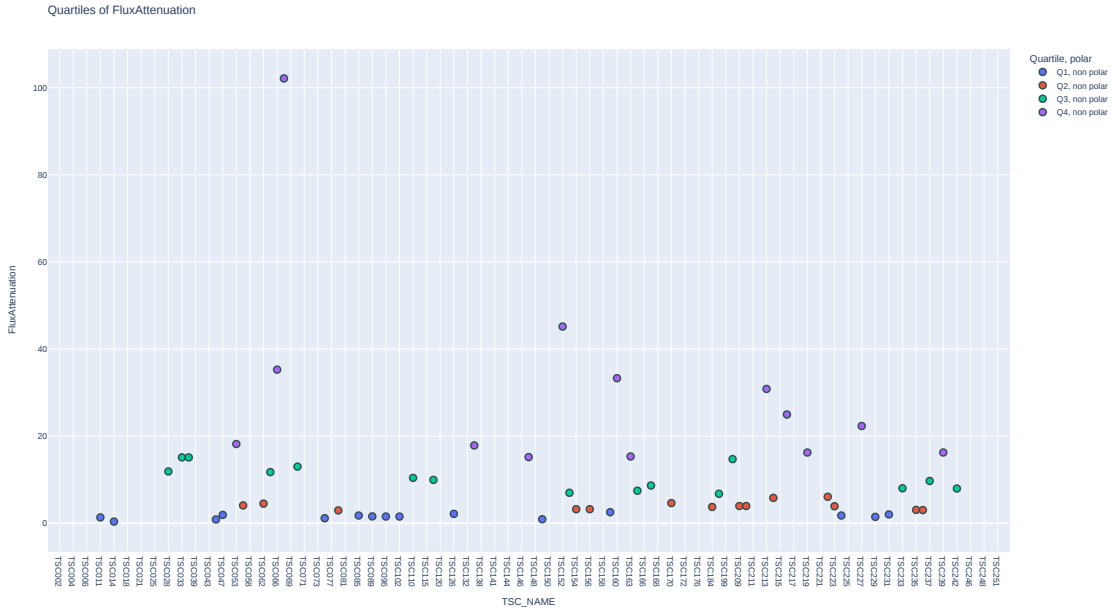


Figure 3.29: Distribution of Flux Attenuation.

As we can see, the measurements were conducted only in non-polar samples. From Figure 3.29 and Figure 3.27, a clear correlation between Flux Attenuation and Carbon export across global seawaters is evident. This correlation is grounded in the fundamental marine processes where the biological pump drives carbon export from surface waters. As this exported organic carbon descends through the water column, it undergoes remineralization, leading to flux attenuation. Essentially, the amount of carbon initially exported dictates the potential carbon available for attenuation in the deeper layers.

3.1.2. Summary

We have discerned significant differences in the distribution of environmental variables based on the geographical layout and pelagic layer of the samples. Of these, temperature is particularly distinguishing, clearly differentiating between the Arctic and non-Arctic oceans. Other variables like salinity, chlorophyll-a, and oxygen also exhibit this differentiation and often correlate with temperature. Furthermore, it is hypothesized that temperature might encapsulate the impacts of these other variables in the biology, offering a more focused avenue for future research into biotic/abiotic relationships. This assertion is supported by two primary observations: a clear correlation between temperature and the aforementioned variables, and the influence of temperature on both biological [13] and environmental [26] characteristics.

To provide a comprehensive visualization of the data described, we present a faceted grid histogram plot in Figure 3.30, where the histograms for these environmental features are colored by ocean region. This color differentiation allows us to discern differences across geographical locations and offers insights into potential methods for further separating them.

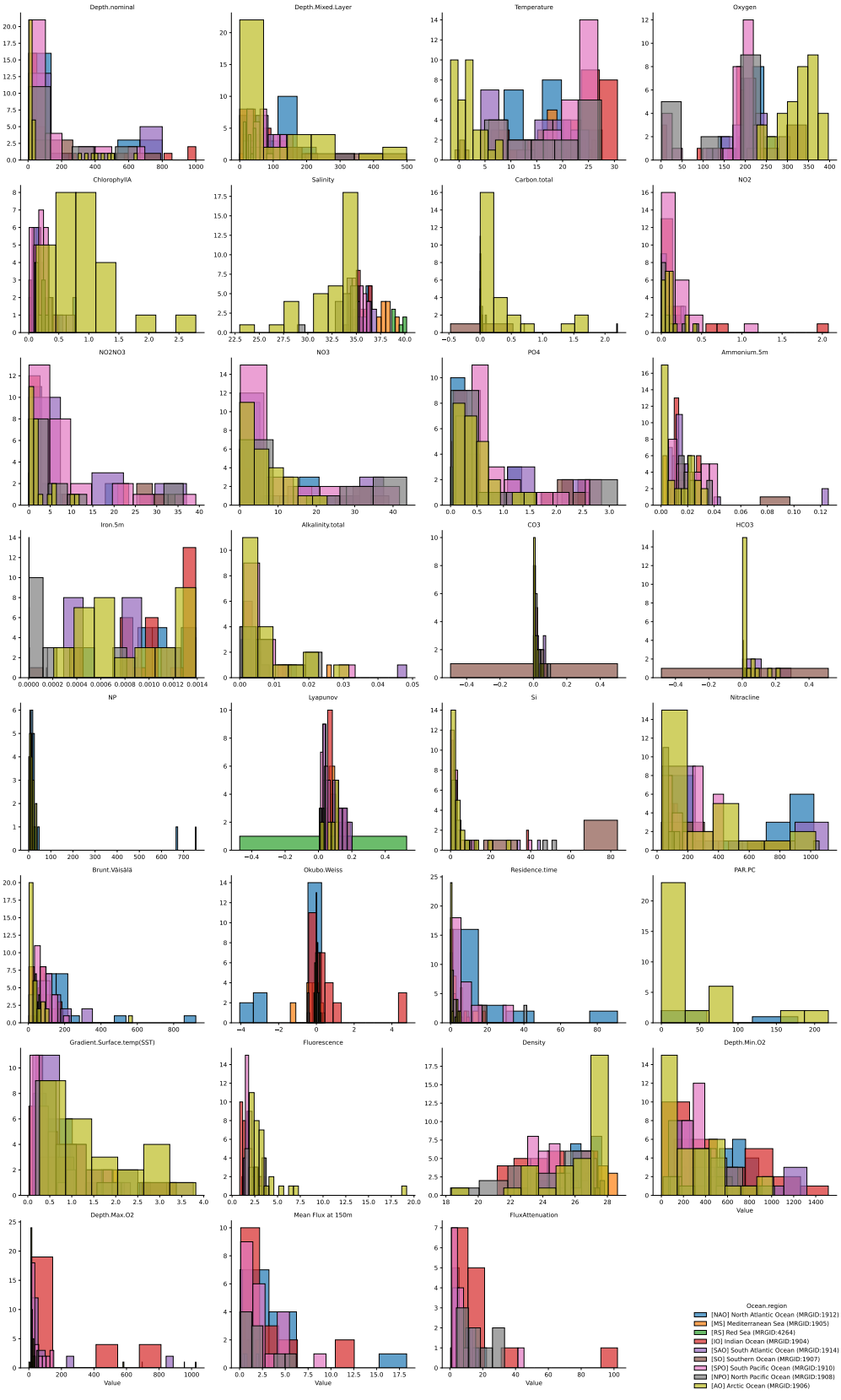


Figure 3.30: Distribution of environmental features across various ocean regions. Each histogram represents the distribution of values for a specific environmental feature, color-coded by the ocean region.

As we expected, the variations in the distributions of temperature, oxygen, chlorophyll-a and salinity between the Arctic Ocean and other marine regions are evident. Specifically, the Arctic and Southern Oceans exhibit lower temperatures, with the Arctic also demonstrating elevated oxygen levels. The Arctic Ocean is distinguished by its higher chlorophyll-a concentrations and slightly reduced salinity. Furthermore, nitrate and nitrite concentrations are notably lower in the Arctic. It is worth noting that data regarding carbon export (denoted as mean flux at 150m) and flux attenuation is absent for polar oceans.

Also, most of the Arctic Ocean samples have a density between 27 and 28. This is different from other marine areas, where the density is spread out between 22 and 27. This is highly influenced by temperature.

3.2. Biological data

In this section, we delve into the biological aspects of our study, specifically focusing on transcription factors (TFs). Transcription factors are proteins that control gene expression by binding to distinct DNA sequences known as transcription factor binding motifs (TFBMs). These motifs serve as specific "docking sites" where TFs exert their regulatory functions. The precise interaction between a TF and its binding motif is crucial for ensuring correct gene regulation. Even though there is an ideal consensus sequence for each binding motif, variations do occur. Resources like the RegPrecise database document these motifs and their associated TFs, with a record of up to 88 distinct TFs. In essence, the interplay between TFs and TFBMs is pivotal in determining when and where genes are expressed or inhibited in a cell, functioning similarly to environmental sensors.

The data under study is derived from prokaryotic bacteria samples collected by the TARA Oceans expedition, which then passed through an ensemble process. For each transcription factor, we tally how frequently it binds to a transcription factor binding site, using data from the RegPrecise database. Specifically, we examine binding occurrences within intergenic zones of a bacterial metagenomic sample from TARA Ocean expedition (163 in total). The final abundance value for a given transcription factor is determined by the abundance of the contigs where this binding event takes place.[8].

The abundance outcome is influenced by three factors: 1) the specific intergenic zones linked to specific genes, which are crucial for obtaining refined results, 2) the length of the Potentially Regulatory Region (PRR). This length determines how far from the start of the CDS we identify binding sites, subsequently affecting which transcription factors might attach, and 3) the threshold probability used to decide whether a transcription factor can feasibly attach to a binding site. We will refer to the abundance outcome from the previously described process as 'binding abundance' or simply 'transcription factor abundance'.

We refer to Figure 3.31 to further elaborate on the points previously discussed.

[8] The datasets used for this work can be found in the following link https://drive.google.com/drive/folders/1rvkxcZTFECgZhTjo4M247_7tFky-8uvH?usp=sharing

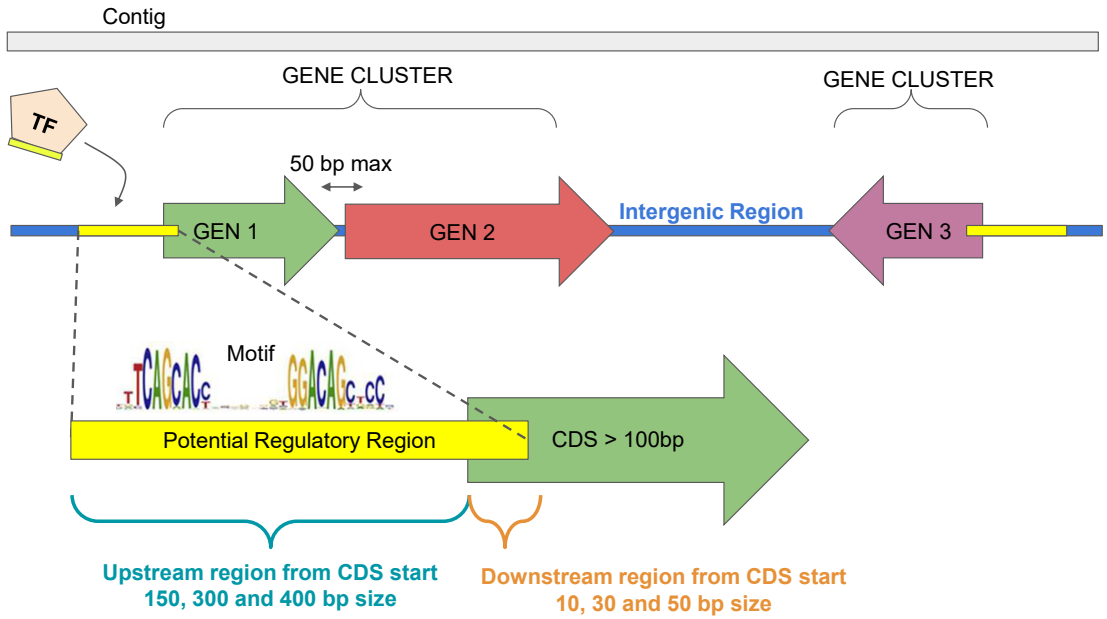


Figure 3.31: Illustration of the Potential Regulatory Region inside a Contig constructed for a bacetrial metagenome and the different upstream and downstream regions from the CDS start that can be studied. Credits: Ricardo Palma from Mathomics Laboratory at CMM.

From Figure 3.31 we can point out that the abundance tally of transcription factors can be honed further by focusing on specific intergenic zones associated with particular genes. For example, we have the option of considering all intergenic zones or narrowing our focus to those linked to functionally annotated genes or to a particular/relevant class of genes. This sets the stage for introducing classes of TFs-Abundance matrices, which we will delve into in subsequent sections.

A universally agreed-upon length for the PRR does not exist. Therefore, we will explore various Upstream and Downstream regions from the CDS start for our analysis, as depicted in Figure 3.31. This affects the final outcome of the binding abundance for each transcription factor.

For the third point is mandatory to explain some preliminaries. Position Weight Matrices (PWMs) are a powerful tool to represent the likelihood of each nucleotide occurring at specific positions within DNA motifs that transcription factors (TFs) bind to. Derived from sets of experimentally determined binding sites, PWMs provide a quantitative score indicating how well a particular DNA sequence matches the known binding sites for a TF. To predict if a TF binds to a given sequence, this score is compared to a threshold, which is often based on a probability value. For instance, setting a threshold of $> 1 - 10^{-06}$ implies sequences with a score correlating to a probability greater than this value are deemed binding sites. Adjusting this threshold allows flexibility in predictions—stricter thresholds yield higher confidence but fewer predictions, while looser ones predict more sites but include potential false positives. In essence, PWMs, coupled with adjustable thresholds, offer a nuanced approach to predicting TF binding sites with precision. In this context, varying the threshold (often referred to as 'cutoffs') results in distinct binding abundance outcomes. We will use $1 - 10^{-x}$, 10^{-x} and x to indicate the cutoffs indistinctly.

3.2.1. Binding abundance matrices

As we delve into the exploration of transcription factors abundances, it becomes essential to elaborate on the methodologies employed for quantifying these factors. Given the plethora of available counting techniques, we have devised a system to generate distinctive matrices corresponding to each method. This section will provide a concise description of the matrices constructed for each counting approach, thereby laying the groundwork for our investigation into the relationship between transcription factors abundances and environmental variables.

The abundances associated with transcription factors (TFs) in each metagenome are calculated using the abundance of the contigs of these metagenomes as follows: 1) the abundance of each contig is the average of the coverage of the reads that served to construct that contig at each contig position; 2) to each TF is associated the sum of the abundances of the corresponding transcription factor binding motifs (TFBMs) associated with it at each intergenic zone of each contig. Each TFBM is counted only once in each intergenic zone that was found with the astringency parameters used.

Using this methodology we used the following abundance matrices, which were constructed at Mathomics Laboratory at CMM (credits: Ricardo Palma).

- **Matrix class M0:** Given a set of TFs and a gene list (basic M0 is with all annotated genes), we compute the abundance (coverage) of each TFM associated to it in all intergenic regiones of a given bacterial metagenome. If a TFM is found more than once in a given intergenic region, its coverage is counted only once.
- **Matrix class M1:** This is a subclass of M0 but with a particular class of genes. Given a set of TFs and a list of metabolic genes, defined by the fact that in KEGG belongs to class "1.X metabolism" (each will have an EC-number), we compute the abundance (coverage) of each TFM associated to it in all intergenic regiones of a given bacterial metagenome. If a TFM is found more than once in a given intergenic region, its coverage is counted only once.
- **Matrix class M2:** This is a subclass of M0 but with a particular class of genes. Given a set of TFs and a list of metabolic genes, defined by the existence of an EC number, we compute the abundance (coverage) of each TFM associated to it in all intergenic regiones of a given bacterial metagenome. If a TFM is found more than once in a given intergenic region, its coverage is counted only once. If one gene has multiple EC numbers, the coverage will be the same for all of them.

As mentioned in the introduction to this section, we can consider different lengths in the PRR and different cutoffs, therefore have different matrices for each described class. For this reason we will focus on presenting results corresponding to specific matrices properly described.

3.2.2. Early results and data visualization

Throughout our study, we will adopt the 'M0_300-30_TF-06' as our benchmark matrix. The nomenclature is instructive: it is an M0 class matrix, characterized by an upstream span of 300 bp and a downstream of 30 bp. The "06" in its name denotes $1-10^{-06}$ a cutoff threshold, restricting the TF to TFBM relation.

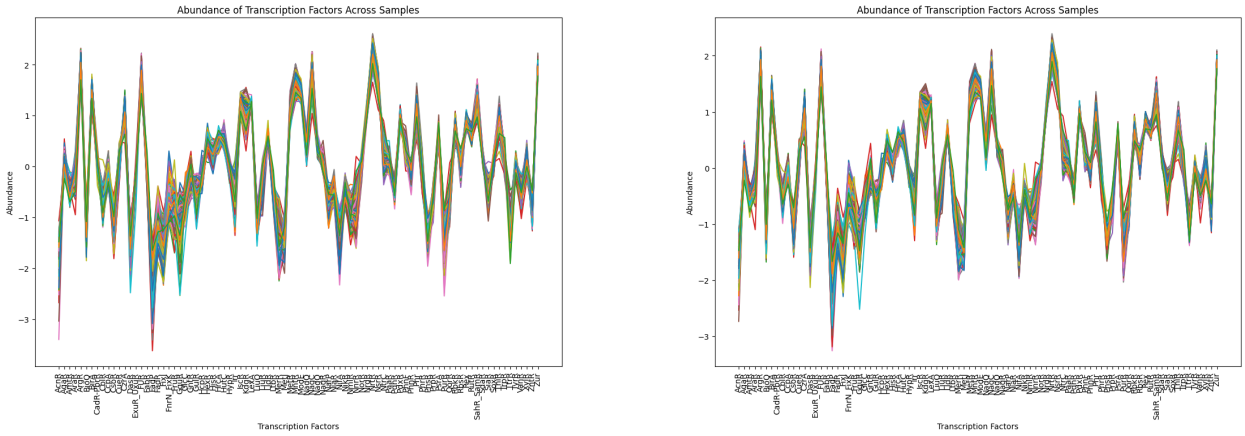
The choice of this specific matrix as a reference is based on the balance it offers in terms of drawing generalizable conclusions, since it is a M0 type matrix and has a "reasonable"

PRR size. While it forms a strong foundation for our initial analyses, we emphasize that our research is not limited to this matrix alone. Rather, we view it as a reliable springboard for our study, enabling us to validate and identify trends in various other matrices. Further analyses on this topic can be found in the appendix section.

In the following, we will illustrate some basic representations of the data extracted from our reference matrix, setting the stage for a more in-depth analysis of the binding abundance in subsequent sections.

To achieve that, we note that our biological datasets are compositional by nature, therefore, we find it appropriate to apply a centered log-ratio normalization, as recommended by Gloor et al. (2017) for compositional data [27]. This normalization will be consistently applied to all biological matrices in our study, unless specified otherwise.

In our initial analysis of these matrices, we will treat the abundance data of each transcription factor as a data point in a sequence, similar to a time series. In this context, 'time' is metaphorically represented by the order of the samples. The results shown in the Figure 3.32 are encouraging.



(a) 150-10 PRR, 10^{-06} cutoff

(b) 300-30 PRR, 10^{-06} cutoff

Figure 3.32: Binding abundance distribution across various samples, post-centered log ratio (CLR) normalization. Each curve represents a sample, showcasing the normalized abundance levels of transcription factors as they bind to transcription factor binding motifs (TFBMs). **(a)** Showcases a matrix with a PRR of 150-10 Upstream-Downstream with a 10^{-06} cutoff. **(b)** Showcases a matrix with a PRR of 300-30 Upstream-Downstream with a 10^{-06} cutoff. Both are an M0 class matrix.

From Figure 3.32 we find that the binding abundance across samples exhibits similar patterns for both the 150-10 and 300-30 PRR. Despite this overall similarity, there are discernible local variations within each individual sample. Furthermore, we observe transcription factors can be classified into high, medium, and low levels of abundance, providing a simplistic yet comprehensive classification approach. Figure 3.32 also puts in evidence a pronounced homogeneity in the binding abundance across samples.

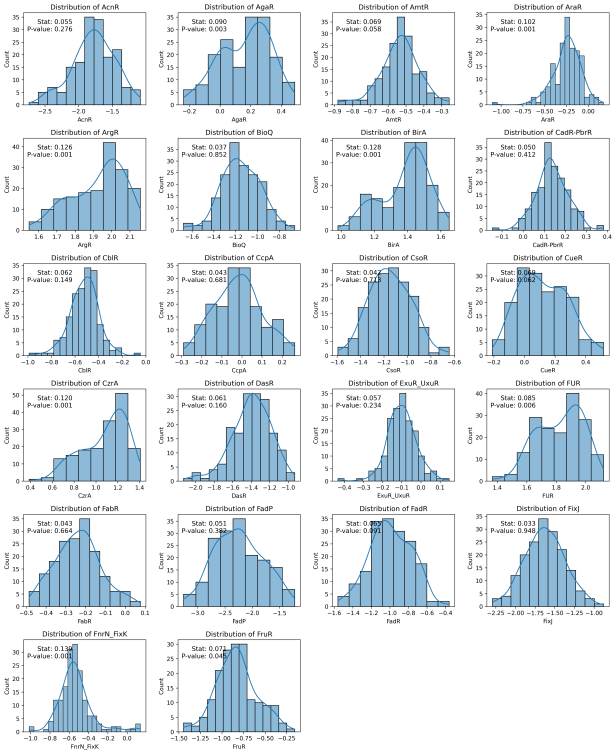
3.2.2.1. Distribution of abundances of the binding motifs associated to each transcription factor

To depict the distribution of each TF binding abundance within our data, we utilize the *Seaborn* library [28]. Moreover, we conduct a Lilliefors Test[9] from *Statsmodel's* library [29] to discern whether these distributions adhere to normality. Understanding these distributions is instrumental for multiple reasons:

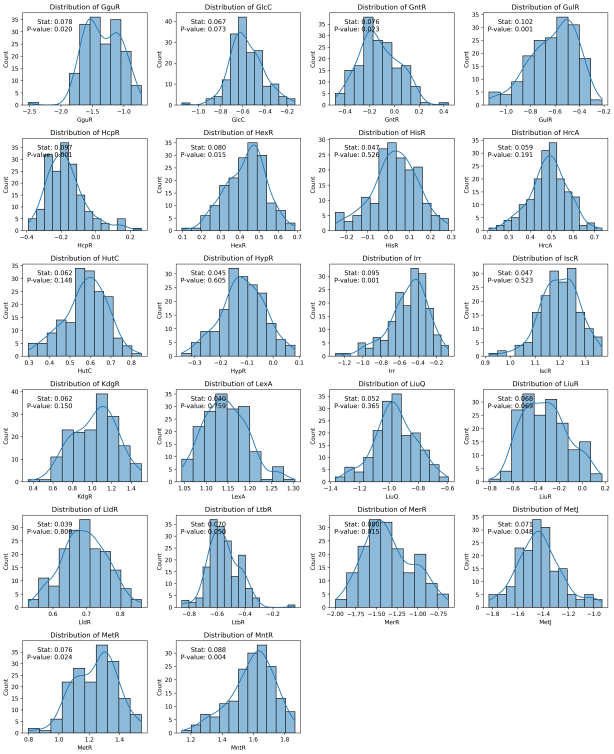
- **Machine Learning Models:** Some machine learning models assume that the input features are normally distributed, or at least have a Gaussian-like distribution.
- **Interpretability:** A normal distribution might also make results more interpretable, as it is a common and well-understood distribution.

The results for the M0 300-30 06 matrix can be found in the Figure 3.33.

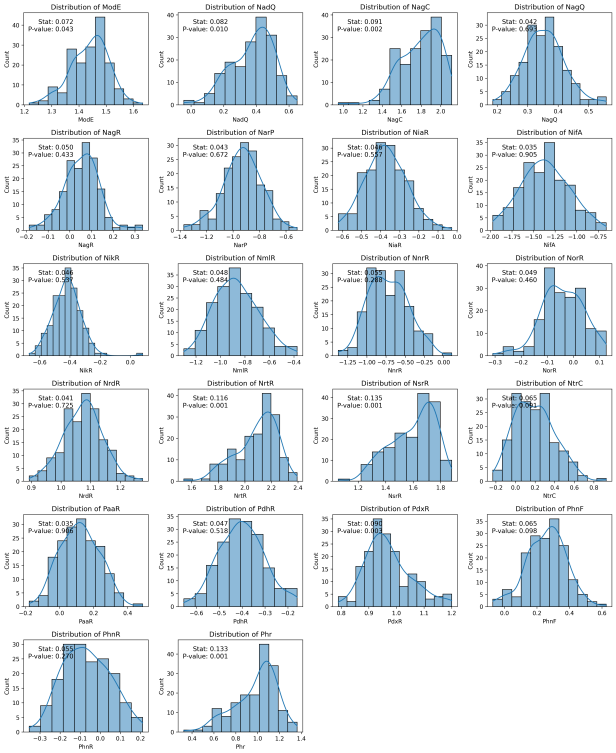
[9] This is a statistical procedure used to ascertain if a dataset follows a normal distribution. It modifies the Kolmogorov-Smirnov (K-S) test, making it more apt when the parameters (mean and standard deviation) of the normal distribution are inferred from the data itself.



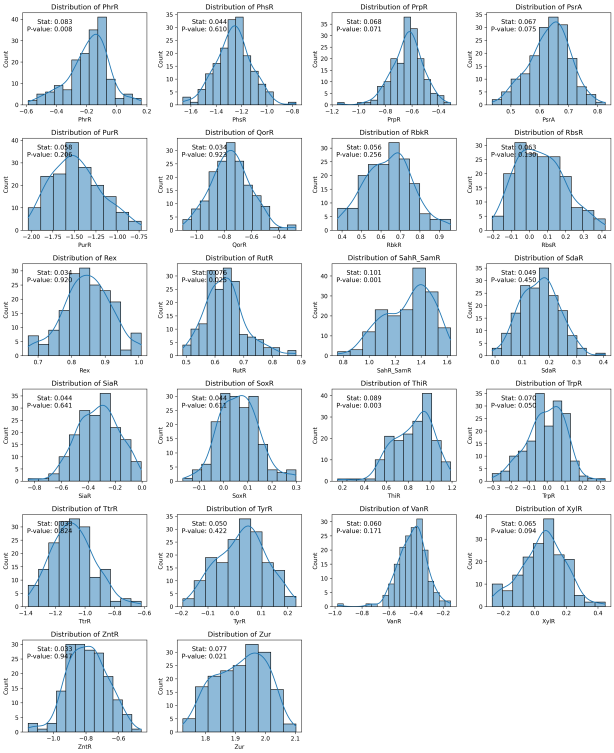
(a) Distribution plots of TFs from AcnR to FruR



(b) Distribution plots of TFs from GguR to MntR



(c) Distribution plots of TFs from ModE to Phr



(d) Distribution plots of TFs from PhrR to Zur

Figure 3.33: Distributions of Transcription Factors (TFs) in the M0 300-30 bp matrix with a 06 cutoff, partitioned into groups for clarity. Each subfigure represents a different set of TFs, illustrating the wide range of distribution patterns across all TFs.

Analyzing these results, we realize that there are 56 TFs that pass the Lilliefors statistical test with a significance of 0.05 and 32 that do not. In other words, there are 56 TFs for which we can say they follow a normal distribution.

Table 3.5: Results of Lilliefors Statistical Test for M0 300-30bp matrix

Test Result	**Count**
Pass	56
Fail	32

One could do the same with all the matrices availables, the results are shown in the Table 3.6

Table 3.6: Results of Lilliefors Statistical Test for various matrices

Matrix	**Pass**	**Fail**
M0 150-10	61	27
M0 300-30	56	32
M1 150-10	71	17
M1 300-30	59	29
M2 150-10	63	25
M2 300-30	63	25

Here are the transcription factors that we can say follow a normal distribution no matter what matrix we choose:

['BioQ', 'CadR-PbrR', 'CcpA', 'CsoR', 'CueR', 'DasR', 'ExuR_UxuR', 'FabR', 'FadP', 'FadR', 'GlcC', 'HrcA', 'HutC', 'HypR', 'KdgR', 'LiuQ', 'LiuR', 'LldR', 'NagR', 'NarP', 'NiaR', 'NifA', 'NmlR', 'NnrR', 'NorR', 'PaaR', 'PsrA', 'QorR', 'RbkR', 'RbsR', 'Rex', 'SdaR', 'SiaR', 'SoxR', 'ZntR']

The discovery that many transcription factors have a normal distribution in their abundance is significant within this thesis. Since the binding sites abundance matrix is introduced here, establishing a normal distribution for its elements suggests the patterns are not random.

However, our analysis extends beyond just normal distributions. Upon visual inspection, certain TFs' binding abundances, such as **GulR**, **HcpR**, **HexR**, **Irr**, **NrtR**, **NsrR**, **PdxR**, **Phr**, **RutR** and **Zur** appear to follow beta distributions.

Following the presentation of the distributions, it becomes natural to determine the means and variances (or standar deviations) for those transcription factors (TFs) that conform to a normal distribution. Figure 3.34 displays the values for these parameters.

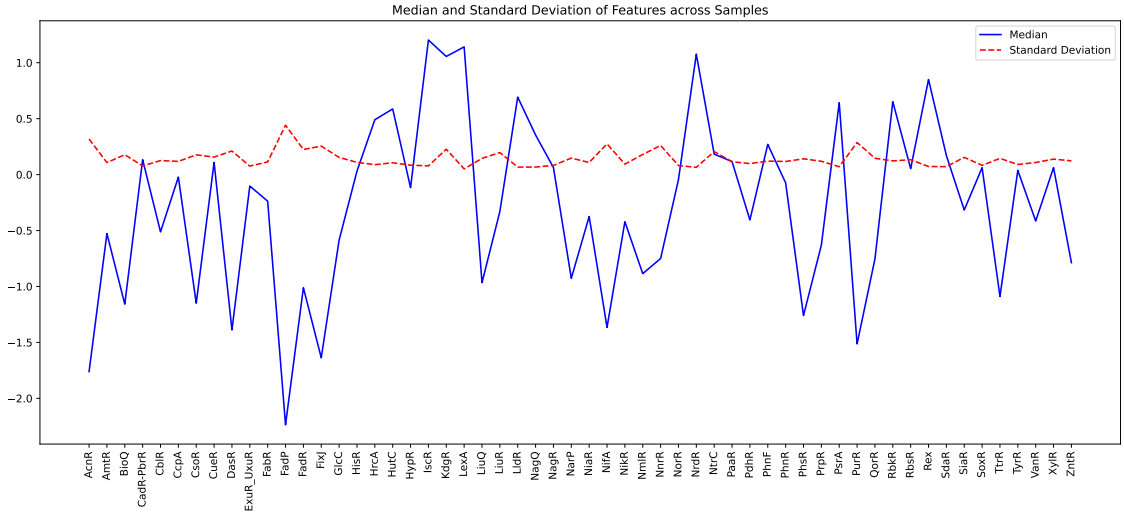


Figure 3.34: Median (in blue) and Standard Deviation (in red, dashed line) of transcription factor binding abundances for selected features across samples. The x-axis denotes the transcription factors, while the y-axis represents the magnitude of the median and standard deviation values.

From Figure 3.34, it is evident that while the medians vary considerably among the TFs, the standard deviation is consistently low. This observation aligns with the findings from Figure 3.32, which display limited fluctuations in abundance for each TF across the samples.

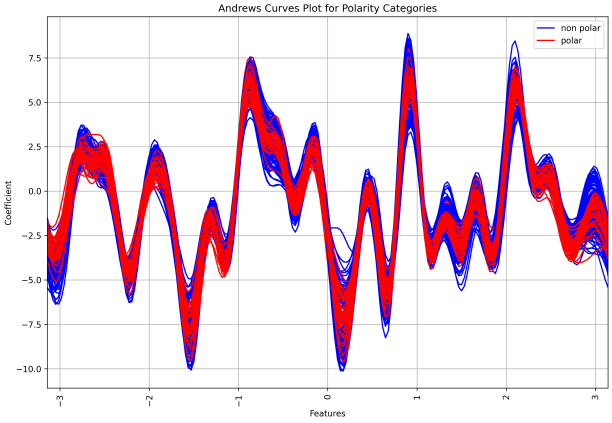
Divergent medians for each transcription factor highlight that regulatory processes are not executed uniformly; certain gene functions demand more frequent regulation than others. Furthermore, the consistent low variance across these factors suggests that these regulatory activities are largely unaffected by location.

3.2.2.2. Multivariate visualization of binding motifs abundance via Andrew's curves

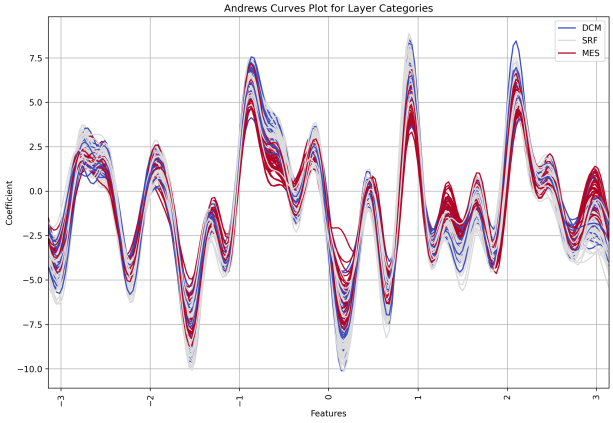
Andrews curves, provided by the *pandas* library, are a method for visualizing multivariate datasets. Each observation is transformed into a continuous function based on its features in high-dimensional space. Specifically, each function is represented as a sum of sinusoids, with each feature sequentially contributing to the sinusoids' coefficients. This approach allows for representation of the data's multidimensional nature in a 2D plot, with each observation shown as a curve.

Using the Andrews curves approach, we applied this technique to our abundance matrix. This offers a streamlined visualization of our multivariate dataset, helping to clarify the data relationships and structures.

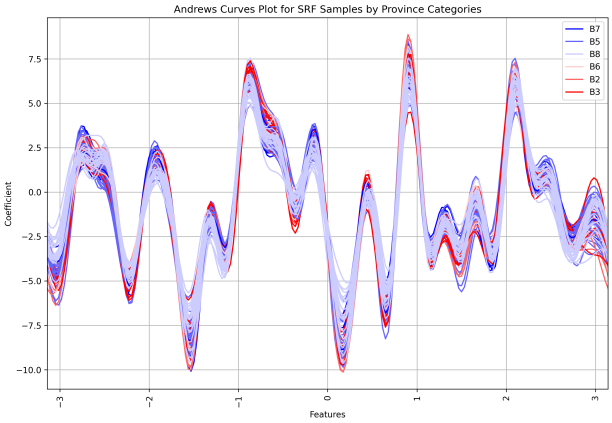
In our analysis, we consider six environmental features: polarity, layer, province, temperature, carbon export, and latitude. The visualization, presented in Figure 3.35, focuses on the Transcription Factors which binding abundances follow a normal distribution, facilitating our exploration of these features' interplay.



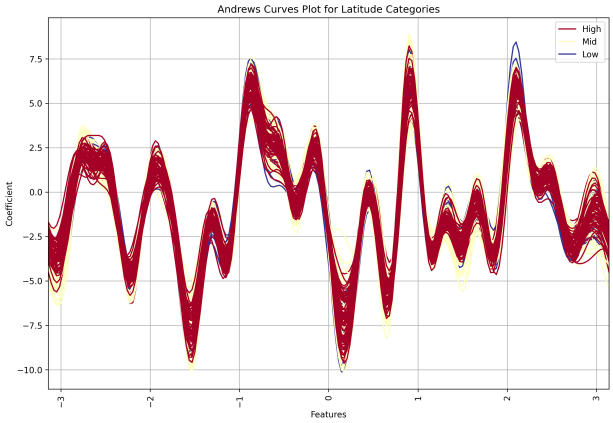
(a) Polar Category



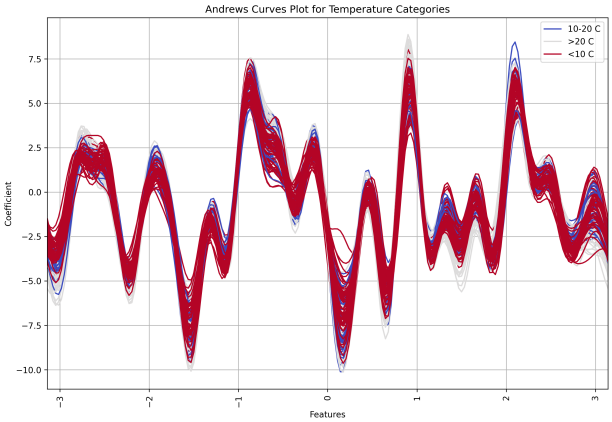
(b) Layer Category



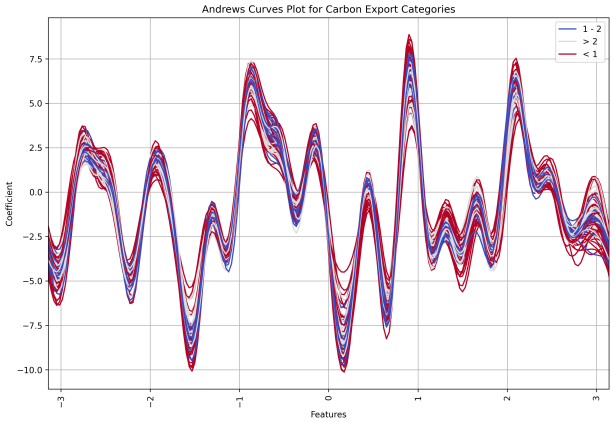
(c) Province Category



(d) Latitude Bins Category



(e) Temperature Bins Category



(f) Carbon Export Bins Category

Figure 3.35: Andrews Curves Visualization for Different Environmental Categories. **(a)** Andrew curves for inspecting the variations and patterns within the Polar category. **(b)** Andrew curves used to visualize the structure in the Layer category. **(c)** Curves revealing relations in the Province (or Bioprovince) category. **(d)** Andrew curves illustrating the diversity of the Latitude Bins category. **(e)** Curves demonstrating the spread in the Temperature Bins category. **(f)** Andrew curves highlighting the differences in the Carbon Export Bins category.

When applying Andrews curves to our dataset, we observed several things:

- In the Polar category the curves exhibited a distinct pattern characterized by a single prominent line that followed the 'polar' label. This finding suggests a strong correlation between the transcription factors and the polar environmental conditions. It might indicate a potential regulatory mechanism associated with the Polar category and understanding the underlying factors driving this pattern can provide valuable knowledge about the genetic regulatory processes in polar environments. Moreover, it suggests that using this biological data we are able to separate polar samples from non polar ones using proper machine learning models.

- Regarding the Layer category, the curves exhibited distinct patterns that varied based on the labels within the Layer category: SRF, DCM and MES. In the case of the SRF label, we observed that the lines tended to group together, forming a prominent line in specific zones. This suggests a strong correlation between the binding abundances of the TFs and the environmental conditions associatedd with the SRF layer. Conversely, for the MES label, we observed a similar behavior, with lines grouping together to form their own distinct prominent line in different zones. This indicates a different set of environmental conditions influencing the binding abundances in the MES layer. Remarkably, the lines from the DCM label displayed a different pattern compared to SRF and MES. Unlike the other labels, the lines from the DCM label did not tend to group together to form a prominent line. This intriguing observation suggests that the DCM layer may have unique characteristics or environmental conditions that result in a more dispersed distribution of transcription factor abundance.

- We observed that the B8 label displayed a distinct pattern compared to the other labels within the Province category. Specifically, the lines corresponding to the B8 label exhibited less local variation compared to the other labels. This suggests that the binding abundance in the B8 province is characterized by a more consistent and stable distribution, with less fluctuations across the dataset. On the other hand, the lines associated with the other labels in the Province category showed greater local variation, indicating a higher degree of heterogeneity in the binding abundance. One extra thing to note is that B8 corresponds to a polar bioprovince.

- Examining the Latitude bins category, we observed distinct patterns between the different latitude bins, specifically in the High and Mid latitudes bins. The lines corresponding to the High latitudes bin exhibited less dispersion, indicating a more consistent distribution of transcription factors. In contrast, the lines in the Mid latitudes bin showed greater variability. Moreover, we observed that in certain regions, the lines of the low latitudes bin acted as upper bounds for all the high latitudes lines, suggesting a limiting effect of low latitudes on transcription factor abundance in high latitudes. Conversely, in other regions, the lines of the low latitudes bin acted as lower bounds, indicating a minimal threshold in transcription factor abundance at low latitudes.

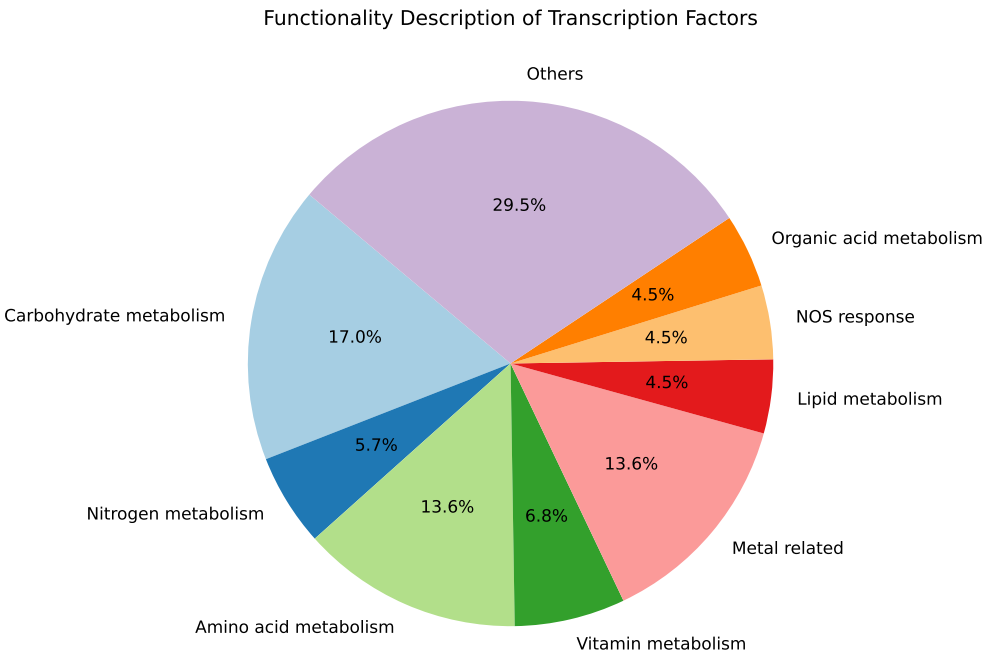
Beyond the direct insights from Figure 3.35, this analysis also states scenarios of particular interest that will be explored in detail later in this work. For instance, one of the most important ones is the polar / non polar scenario.

3.2.3. General description of transcription factor functionality.

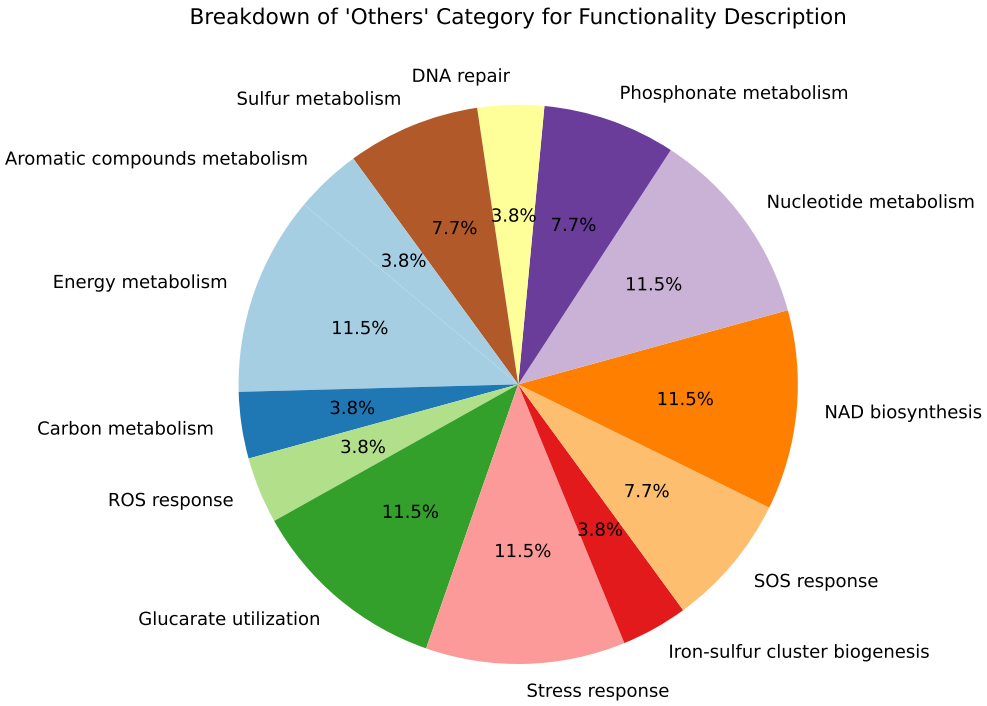
The RegPrecise database provides a notable category: the general description functionality linked to each transcription factor. This categorization is paramount as it offers deeper

insights into the specific roles and functions of each transcription factor, enriching our comprehension of their individual behavior.

Following this categorization, an immediate inquiry arises: Which functionality is predominant among the transcription factors under investigation? To address this, we will employ a pie chart, offering a visual representation of the relative occurrences of transcription factor's functionality.



(a) Pie chart representation of the transcription factor functionalities derived from the RegPrecise database



(b) Further exploration of the 'Others' segment.

Figure 3.36: Visual representation of transcription factors categorized by their general functionalities. **(a)** Categories representing less than 4% of the total transcription factors are collectively grouped under the 'Others' category **(b)** Detailed breakdown of the 'Others' category from the general functionality description of transcription factors.

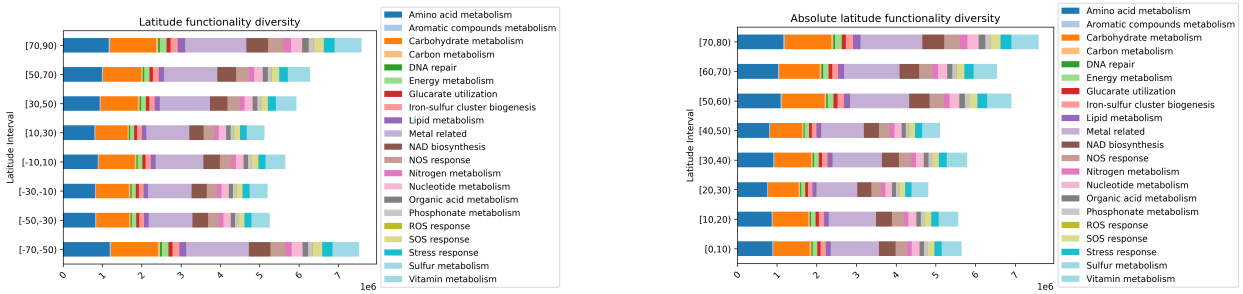
Figure 3.36 highlights three primary categories of transcription factors: Carbohydrate metabolism, Amino acid metabolism, and Metal-related. Most transcription factors are associated with carbohydrate metabolism, primarily due to the photosynthesis process performed by phytoplankton. The second most prevalent category of transcription factors is associated with amino acid metabolism. As amino acids are the foundational building blocks for proteins, their regulation is crucial for the growth, repair, and maintenance of phytoplankton cells. Lastly, transcription factors associated with metal responses are also significant. These factors allow phytoplankton to effectively manage metal concentrations, given the essential roles of trace metals like iron, zinc, and copper in various cellular processes. Such regulation ensures phytoplankton thrive despite fluctuating metal availability in marine waters.

3.2.4. Latitudinal diversity

Earlier in this chapter, we discussed the environmental and biological datasets, showcased the biotic component distributions, and employed Andrews curves for a multivariate analysis, aiming to understand biotic/abiotic relationships. While this provides an overview of our abundance matrix and its role in predictive model development, we need to compare functional aspects of it with recent literature results for a comprehensive evaluation. To accomplish this, we will refer to the 'General description' entry for each transcription factor in the RegPrecise database as we did for the previous chapter.

Expanding upon the findings of Ibarbalz [11] and Salazar [14] on plankton and genomic latitudinal diversity, we delved into the latitudinal functional diversity of transcription factors.

To undertake this analysis, we aggregated the binding abundance of transcription factors per sample based on shared functionality entries, resulting in a new dataset termed the 'functionality abundance matrix'. Subsequently, samples were grouped by latitude and absolute latitude categories. The outcomes of this approach are shown in Figure 3.37.



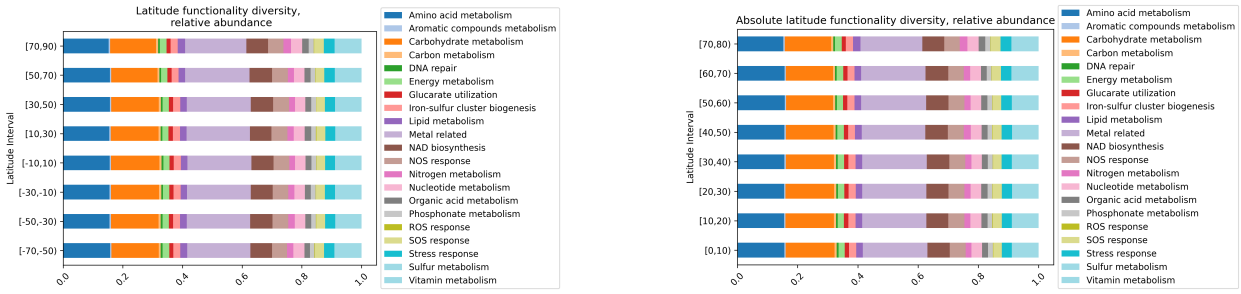
(a) (b)

Figure 3.37: Latitudinal diversity of the functionality of the transcription factors with the highest binding count corrected by quantity of samples in each latitude interval. **(a)** Shows the latitude diversity from top to bottom with latitude intervals of 20 degrees. **(b)** Shows the absolute latitude diversity, i.e., from the equator to the poles, with latitude intervals of 10 degrees.

Obs: *It is important to note that for each latitude interval, different quantities of samples were taken that directly affect the total abundance of measured transcription factors. For this reason, a correction for the number of samples in each interval was considered next.*

Figure 3.37 clearly illustrates that latitude does not govern the diversity in functionality

linked to the binding abundance. Further, it can be observed that the relative[10] standing of each functionality (regarding the other functionalities), remains almost constant across various latitudes. This can be better observed in the following figure:



(a) (b)

Figure 3.38: Relative latitudinal diversity of the functionality of the transcription factors with the highest binding count. **(a)** Shows the relative latitude diversity from top to bottom with latitude intervals of 20 degrees. **(b)** Shows the relative absolute latitude diversity, i.e., from the equator to the poles, with latitude intervals of 10 degrees

This outcome is noteworthy for two main reasons. Firstly, it contrasts with recent literature that identifies diversity gradients, whether of species or genes, across different latitudes. Here, we observe that each functionality is consistent throughout the ocean. Secondly, the findings suggest that regulatory mechanisms are largely uniform across the ocean, evidenced by the similarity in relative functional abundances for each latitude category, whether absolute or not.

[10] Relative in the sense that once the transcription factors' abundances are grouped by functionality (which I refer to as functionality abundance), I proceed to group them by latitude intervals. Then, for each of these intervals, the relative abundance of each functionality is calculated.

Chapter 4

Transcription factor binding structure and its relation with environmental variables

Understanding the correlations between variables is fundamental to elucidating the biological and environmental mechanics of the systems investigated in this study. This form of analysis can unveil key associations between biotic and abiotic variables.

To explore these relationships, we employ correlation analysis, specifically Spearman correlation, utilizing the *pandas* library in Python [30]. Unlike Pearson's correlation, which only captures linear relationships, Spearman's correlation is based on the rank order of values. This enables it to detect more complex, monotonous relationships, irrespective of whether they are linear. Given our aim to comprehend the structural relations between variables, not merely the linear associations, Spearman's correlation emerges as a more fitting choice for our study.

Our analysis starts with showcasing pairwise correlations for the abiotic and biotic data, followed by illustrating the relationship between the two.

Down below there is a guide to this chapter for enhanced comprehension and navigation:

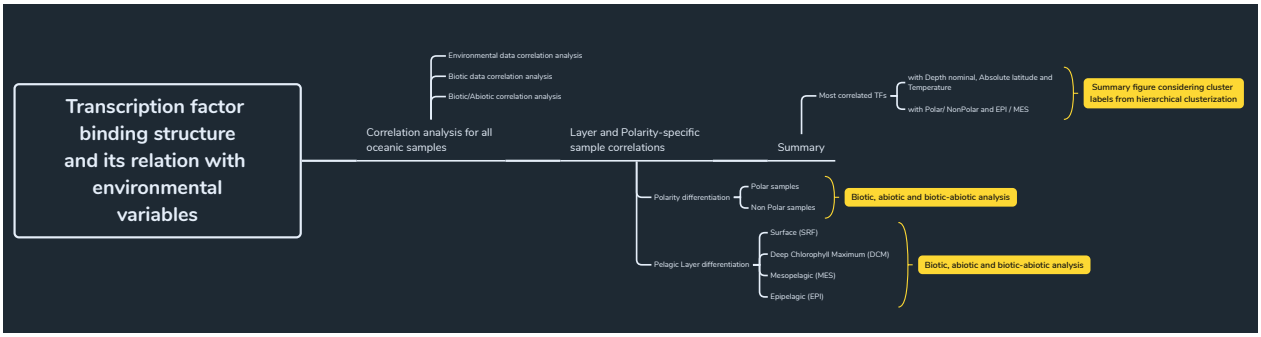


Figure 4.1: Chapter guide

4.1. Environmental data correlation

Making sense of the complex relationships between different environmental factors, such as geographic location, nutrient levels, and chemical properties, is a crucial part of understanding how our environment works. These factors can interact in complex ways that lead to a

variety of environmental processes and events.

By using a correlation's approach, we can estimate the strength of the relationships between these factors. This can help build a statistical model that predicts how the environment might behave under different circumstances. However, it is important to remember that correlation does not necessarily mean that one factor causes another.

In this section, we are focusing on the correlation analysis of various environmental factors. For a more detailed look, we haveve grouped these factors into specific categories: Geographic and Physical Features, Nutrient Availability Features, and Chemical Composition Features. Each category has unique effects on the environment that we are studying.

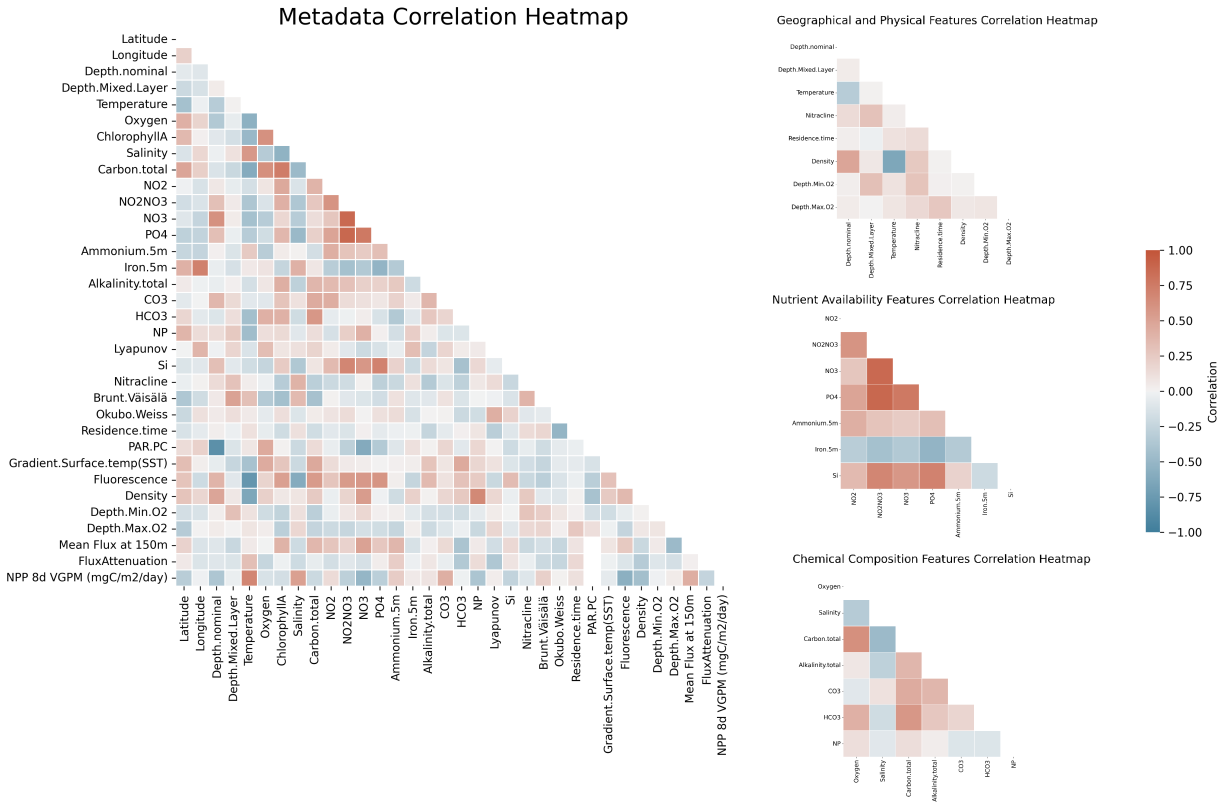


Figure 4.2: Heatmap analysis of environmental metadata. The large heatmap on the left represents the pairwise correlation of all environmental features, providing a holistic view of the intricate interconnections among all studied variables. To the right, we present three focused heatmaps, each showcasing a distinct group of environmental features: Geographic and Physical, Nutrient Availability and Chemical Composition. These focused heatmaps help illuminate the specific interrelations within each category, providing a more nuanced understanding of the environmental systems being studied.

From the heatmap presented in Figure 4.2 we note the following:

- Bicarbonate (HCO3) shows a strong correlation with both total carbon and alkalinity. These three features appear to have a synchronized variation pattern in the analyzed environment.
- Depth reveals a correlation with nutrient availability, specifically with nitrate (NO2NO3) and phosphate (PO4). The data suggests that the variations in these nutrients are

associated with changes in depth. The correlation is more significant with nitrate (NO3) than nitrite (NO2).

- Silica presents a relationship with most nutrients except for ammonium and iron. Its variation seems to parallel changes in the levels of nutrients like nitrate (NO2NO3) and phosphate (PO4).
- ChlorophyllA and total carbon exhibit a close correlation. Their levels seem to rise and fall in sync, suggesting a shared influence or interaction in the environmental system we are studying.
- Oxygen and temperature exhibit a negative correlation. As one increases, the other decreases, suggesting an inverse relationship between these two environmental features in the analyzed environment.
- Carbon export shows a strong correlation with bicarbonate (HCO3). This suggests that these two features could share some related mechanisms or processes in the environment under study.

While interpreting these results, it is crucial to bear in mind that correlation does not equate to causation. The fact that two variables are strongly correlated does not necessarily mean that changes in one are the cause of changes in the other. However, these correlations are still of great value, particularly in the context this thesis refers to, because we reveal a structural behavior of the environment as a whole.

A correlation heatmap is not the only tool that aids in understanding relationships among variables. A hierarchical clustermap, which transforms the visually complex correlation matrix into an easily interpretable visualization, proves to be invaluable. By grouping variables with similar correlation patterns, it provides a structured representation of potential hierarchical relationships among them.

By interpreting the clustermap, we gain insights about the interrelatedness of variables, identifying clusters that might exert similar influences on the phenomena under investigation. The color scheme within the map, ranging from red (indicating a strong positive correlation) to blue (a strong negative correlation), further facilitates the understanding of these relationships.

For the environmental features we have the Hierarchical Correlation Clustermap shown in Figure 4.3

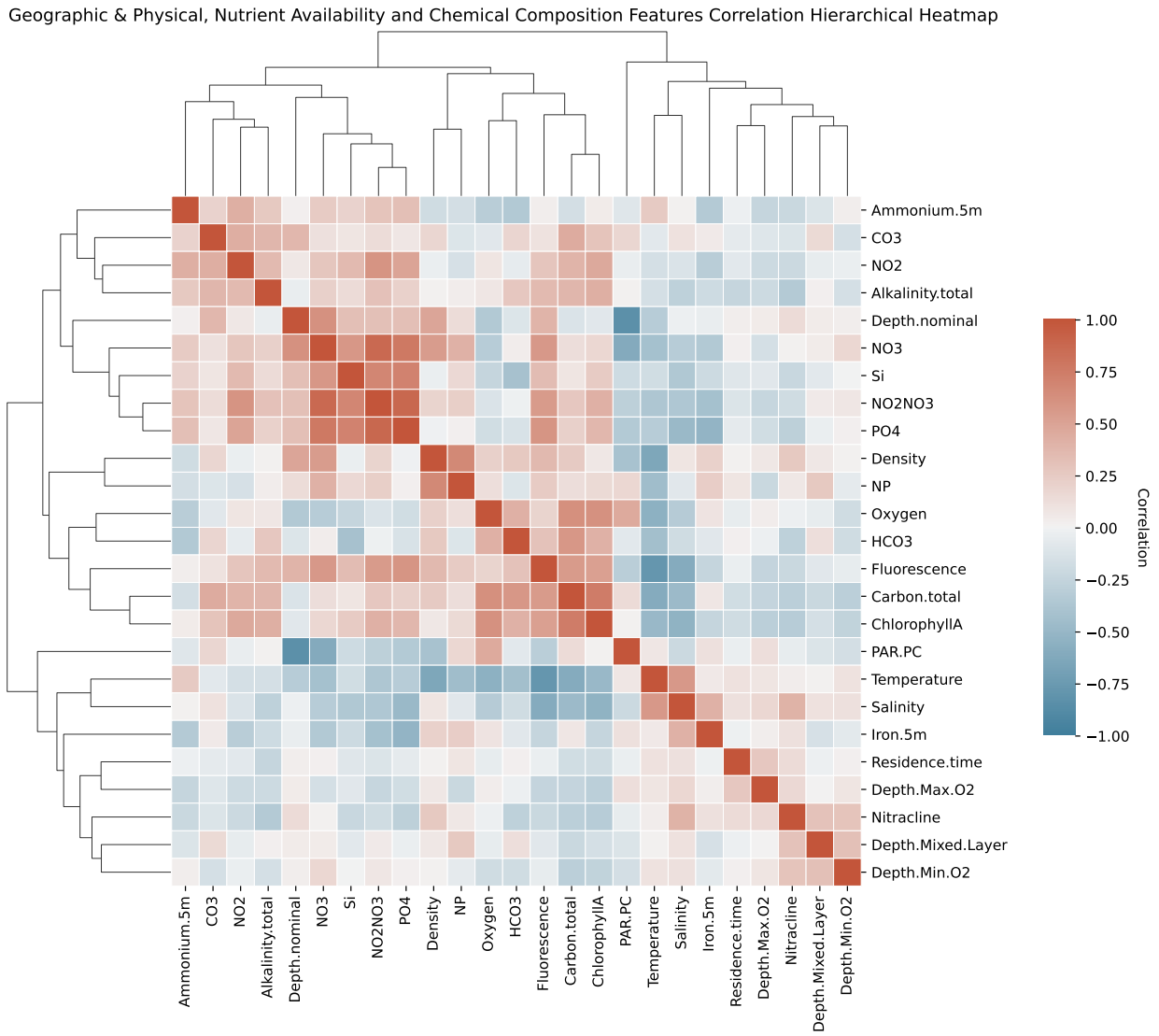


Figure 4.3: Hierarchical Correlation Clustermap of Various Environmental Variables Using the Average Linkage Method. The clustermap showcases the correlation between diverse geographic, physical, nutrient availability, and chemical composition features. Variables are hierarchically clustered based on their correlation patterns. The color gradient represents the Spearman correlation coefficients ranging from -1 (negative correlation) to +1 (positive correlation).

It illustrates three big clusters based on their correlation, these are 1) Ammonium, CO3, NO2, Alkalinity, Depth, NO3, Si, NO2NO3 and PO4, 2) Density, N:P ratio, Oxygen, HCO3, Fluorescence, Carbon total and Chlorophyll-a and 3) PAR.PC, Temperature, Salinity, Iron at 5m, Residence time, Depth of Max O2, Nitracline depth layer, Depth of mixed layer and Depth of Min O2. These can be further labeled as 1) Nutrient Cycling and Alkalinity Cluster; 2) Physical Properties and Primary Production Cluster and 3) Environmental Conditions and Depth-Related Features Cluster.

The observed clusters are not arbitrary. Indeed, the detailed descriptions provided in Chapter 3 align well with the representations in Figure 4.3.

The 'Nutrient Cycling and Alkalinity Cluster' encompasses features that are positively correlated. Notably, nitrate, a primary nutrient source, exhibits lower abundance in surface

waters, resulting in deeper waters having increased nitrate concentrations. This pattern holds true for other oceanic nutrients, such as phosphate and nitrogen-based compounds (Ammonium included), which tend to be more concentrated in the ocean's deeper layers. Additionally, this cluster includes carbonate ions. These ions predominantly originate from the dissolution of calcium carbonate, a process that mainly takes place below the carbonate compensation depth (CCD)[11], typically situated between 4,000 and 5,000 meters deep. As described earlier regarding carbonate, marine organisms like coccolithophores and pteropods utilize calcium carbonate to construct their shells and upon the death of these organisms, their calcium carbonate shells sink, promoting further dissolution of the molecule in the deep ocean, thereby enhancing carbonate concentration at greater depths. This phenomenon significantly influences seawater alkalinity, since hydrogen ions, introduced through the CO2 transfer from air to sea, interact with carbonate to produce bicarbonate, contributing to lowering the pH [31].

The hierarchical structure based on correlation for clusters 2 and 3 is not as pronounced as it is for cluster 1. However, we note that in the 'Physical Properties and Primary Production Cluster', environmental factors serve as foundational components for primary production features. For instance, water density is instrumental in vertical mixing, determining nutrient distribution. This nutrient availability, showcased by the N:P ratio, directly affects phytoplankton growth. This growth can be assessed through fluorescence and chlorophyll-a metrics. Furthermore, phytoplankton modulate oxygen concentrations via photosynthesis and play a role in the carbonate cycle, thereby influencing bicarbonate levels. Intricately, all these dynamics concurrently affect and are impacted by the ocean's total carbon content [31].

4.2. Biotic data correlation

Up next, we delve into exploring correlations within the biotic data[12]. It is important to remember that in this study, we standardize our approach by using the centered log normalization (CLR) of biological data.

In Figure 4.4 we present the correlation between the binding abundances associated to transcription factors.

[11] The CCD is the depth in the oceans below which the rate of dissolution of calcium carbonate exceeds the rate of its accumulation

[12] We recall that we are using the standar TF's abundance M0 class matrix with a 300-30 bp

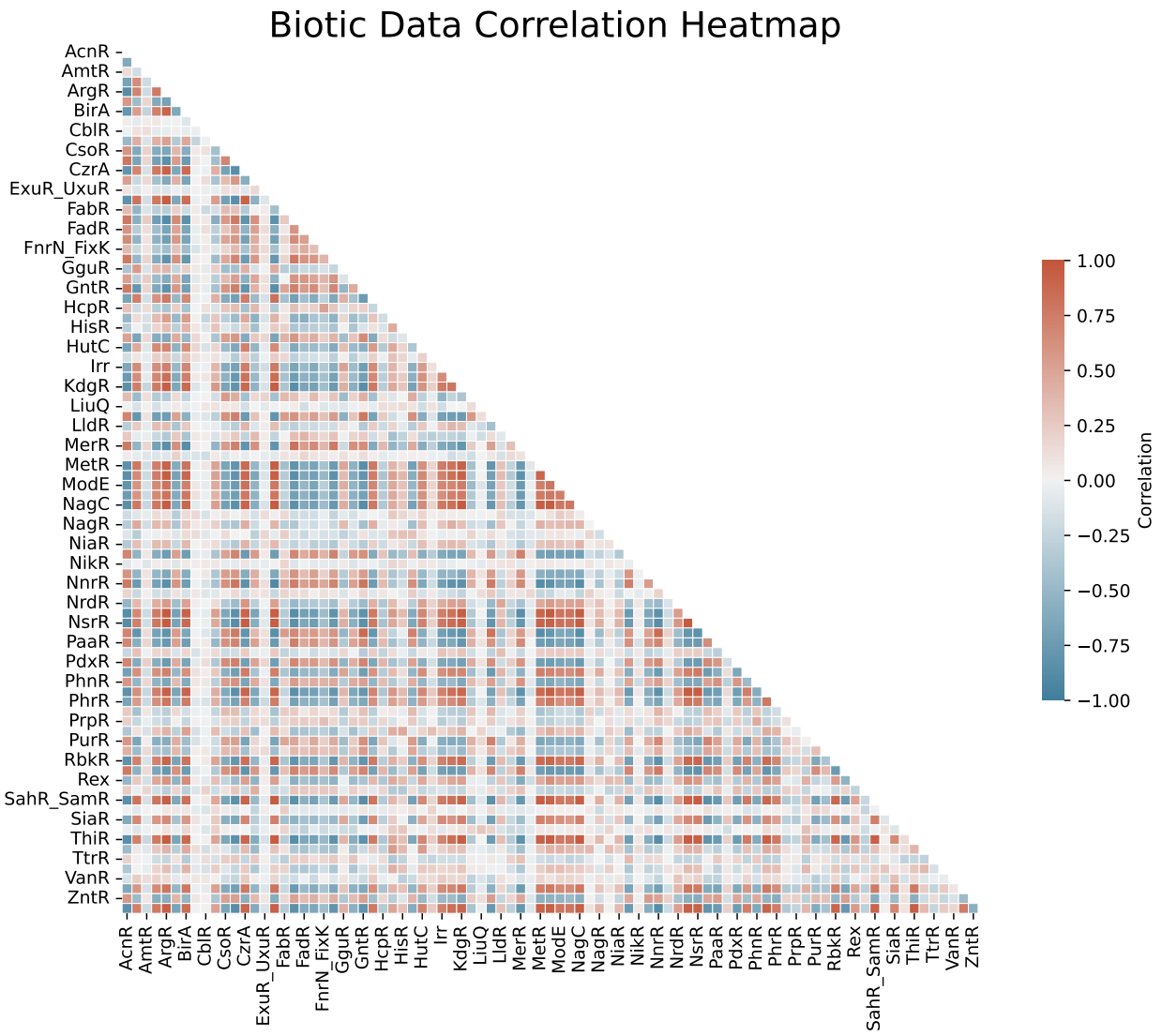


Figure 4.4: Biotic Data Correlation Heatmap. The heatmap displays the Spearman correlation coefficients between the abundances of Transcription Factors binding motifs. The color gradient ranges from -1 (indicating a negative correlation) to +1 (indicating a positive correlation). This visual representation helps discern the strength and direction of relationships between variables.

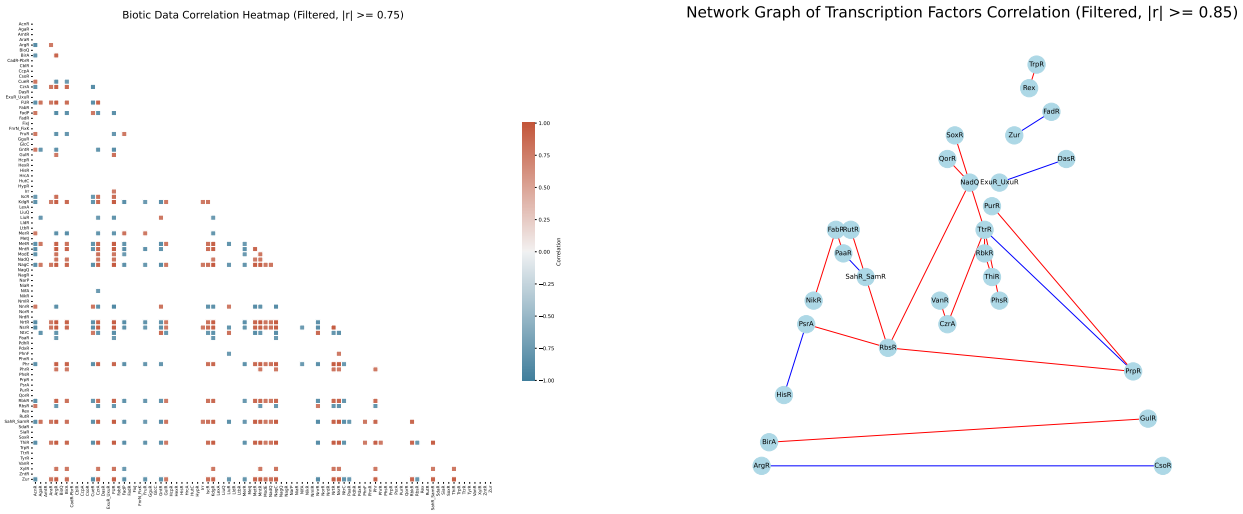
While the heatmap gives a broad perspective on the transcription factors' interconnections, a more nuanced understanding emerges when we focus on the strongest signals. By applying an absolute correlation threshold, we filter out the most influential relationships, sharpening our insights into the transcription factors' dynamics within the intricate marine environment.

From the original correlation matrix, we then applied a correlation filter to the genomic dataset, retaining only those transcription factors with an absolute correlation above 0.75. This minimizes redundancy and brings focus on strong, potentially meaningful associations between transcription factors. The chosen threshold strikes a balance between keeping meaningful data and reducing complexity, thus streamlining the interpretation process. This enables us to uncover significant biological relationships in an otherwise dense dataset.

We visualized the meaningful relationships between transcription factors using a network graph. To ensure a clear representation, we employed the *NetworkX* library [32] and set a

stringent correlation threshold of 0.85, ensuring a planar graph[13]. More so, it is the Planar Maximally Filtered Graph (PMFG) [33]. This graphical representation provides an intuitive way to understand complex relationships, illustrating links between highly correlated transcription factors as edges in the graph. By focusing on such high-correlation pairs, we are honing in on the most influential associations in our dataset, further refining our analysis and potential insights.

Both can be visualized in Figure 4.5



(a) (b)

Figure 4.5: **(a)** Heatmap of the Spearman correlation matrix computed from the CLR-normalized transcription factor's abundance dataset. The colormap ranges from -1 (blue, negative correlation) to 1 (red, positive correlation). Only transcription factors with absolute correlations above 0.75 are displayed, reducing complexity and highlighting strong associations. **(b)** A planar network graph representing the strong correlations (above 0.85) among the transcription factors. Nodes represent transcription factors, and edges represent the strong correlations between them. Red edges indicate positive correlations, and blue edges indicate negative correlations. The layout of the graph provides a visual depiction of the relationships among transcription factors, thereby assisting in discerning potential biological significance.

From Figure 4.4 and Figure 4.5 we can point out the following:

- The correlation heatmap for all transcription factors reveals a diverse set of interactions. A distinct group of transcription factors demonstrates particularly strong correlation, both positive and negative, with others in the same group. This suggests that these transcription factors might have closely intertwined roles in the regulatory network and might function together in a common biological process or pathway. On the other hand, some transcription factors show a relatively weaker correlation with the others, implying that their functions might be more independent or their interactions more nuanced.

- By imposing a correlation threshold of 0.75, we can further refine the interactions among the transcription factors. This filtering process effectively removes weaker and potentially less significant correlations, enabling us to focus on the transcription factors with

[13] A planar graph in simple terms is a graph that can be drawn on a flat surface, such as a sheet of paper, without any of its edges crossing each other.

the most robust and consistent relationships. This set of transcription factors could represent a core regulatory cluster within the broader transcriptional network.

- Further refining the correlation network with a threshold of 0.85 results in a tree-shaped planar graph. This graph presents a clearer picture of the strongest interactions, which can be viewed as the backbone of the transcriptional network. In this network, we can see distinct connected components which might represent sub-networks or modules within the overall regulatory network.

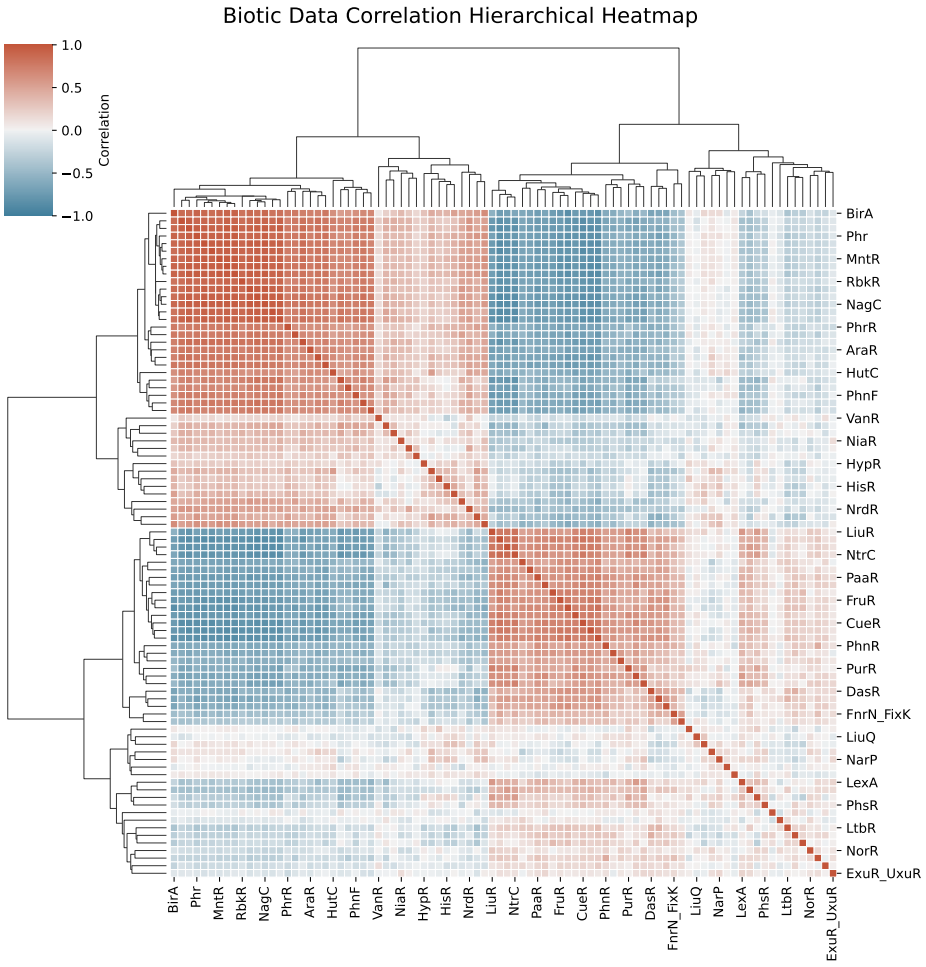
- It is interesting to note that the most significant sub-network exhibits an alternating correlation direction. The change in the color of the edges implies a cyclical behavior among the connected transcription factors. This could reflect the presence of feedback loops or cyclic pathways in the regulatory network, where the output of one factor affects its own regulation in a cyclic manner.

- The tree-shaped graph suggests a hierarchical structure in the correlations between the transcription factors. The direction and strength of these correlations could imply a certain order or priority in the execution of transcriptional regulation.

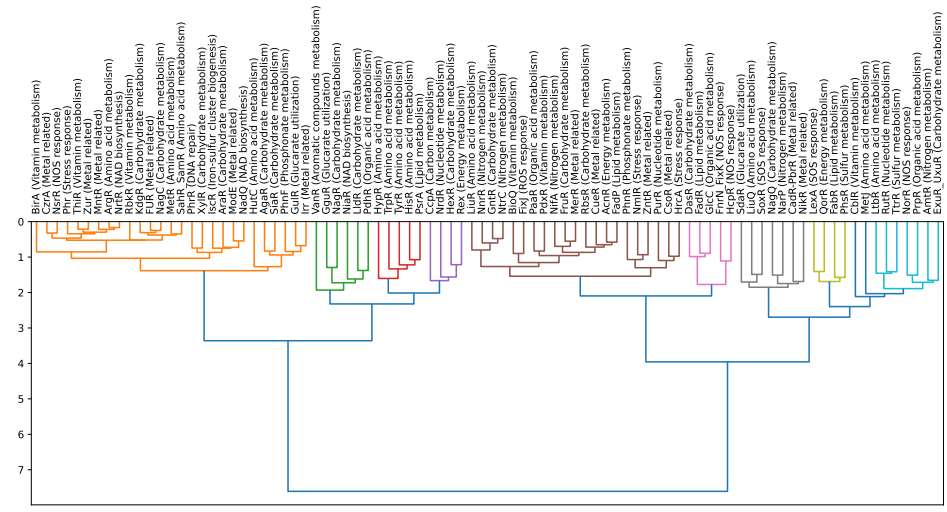
- In the correlation network, there are five transcription factors - 'TtrR', 'NadQ', 'RbsR', 'SahR_SamR', and 'PrpR' - that have a degree greater than 2. These transcription factors might be particularly influential within the regulatory network due to their extensive connections with other factors.

- These results open up new avenues for further research, including exploring the nature and implications of these strong correlations.

The patterns observable in the correlation heatmap can be more effectively visualized through a hierarchical clustermap, as displayed in Figure 4.6. (See Appendix B for more information).



(a) Hierarchical Clustermap



(b) Dendogram

Figure 4.6: Hierarchical Clustering and Dendrogram of Transcription Factors. **(a)** Depicts the clustermap of correlation coefficients among transcription factors, giving a vivid representation of the data structure and highlighting clusters of highly correlated transcription factors. **(b)** Shows the corresponding dendrogram where the distance threshold has been set to delineate distinct clusters. The dendrogram provides a visualization of the hierarchical clustering process, demonstrating the grouping of transcription factors based on their correlation coefficients.

From Figure 4.6.b, ten distinct clusters are evident. The orange and brown clusters are the most prominent, containing 27 and 21 TFs respectively. By adjusting the cluster color threshold, four major clusters emerge. For additional details, refer to Appendix B. Furthermore, transcription factors functionalities were annotated and we conclude that TFs are not grouped based on any specific functional criteria, except for the red cluster, which is made up of amino acid metabolism type. Referring to the dendrogram in Figure 4.6.b, it is essential to note that this was constructed based on the binding abundances associated with each transcription factor. Therefore, understanding the interconnections between them is essential to have a better understanding of the microbial communities transcriptomes. Moreover, in line with the insights from Luscombe et al. [9], we stratified our analysis by geographic locale and pelagic layer to account for environmental differences. This stratification allows us to explore how the interplay between biotic variables shifts across varied oceanic conditions.

4.3. Biotic/Abiotic relations through correlations

In complex marine ecosystems, the interplay between biotic and abiotic factors is crucial. Specifically, the correlation between transcription factors (TFs) and environmental parameters can greatly impact community structures and species distributions. In this section, our aim is to correlate the variations in TFs with changes in these abiotic variables. We hope to identify how environmental factors might impact genetic regulatory networks or vice versa. Our approach involves creating a correlation matrix, visualizing it through a heatmap, and constructing a correlation network to highlight the strongest associations.

In Figure 4.7 we display the correlation between the transcription factors and the environmental variables.

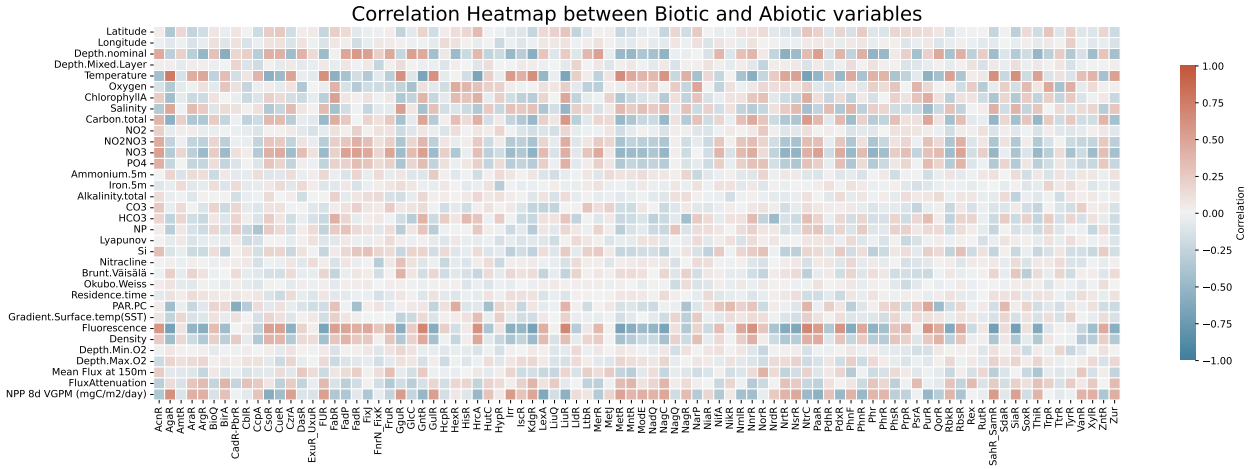


Figure 4.7: Biotic-Abiotic Correlation Heatmap. The heatmap depicts the Spearman correlation coefficients between the Transcription Factors (biotic data) and environmental variables (abiotic data). The color gradient spans from -1 (indicating a negative correlation) to +1 (indicating a positive correlation). This visualization aids in understanding the strength and direction of the relationships between these biological and environmental variables.

The heatmap of the correlation matrix offers initial insights into the interactions between abiotic variables and transcription factors. Upon visual assessment, certain abiotic variables seem to have a more pronounced correlation with the transcription factors, specifically Tem-

perature, Depth and Fluorescence. These parameters exhibit higher absolute correlation values, suggesting a stronger relationship with the biotic factors in our study.

While the heatmap offers a comprehensive picture of correlations, dissecting specific relationships among these variables is equally vital. In our pursuit to examine these connections further, we employ a correlation network—a powerful tool for visualizing intricate relationships. The forthcoming correlation network analysis is set to deepen our understanding of these associations, thereby offering a more defined impression of environmental factors' influence on genetic regulatory networks. As a testament to this, Figure 4.8 illustrates the correlation network, exclusively featuring robust correlations above 0.5:

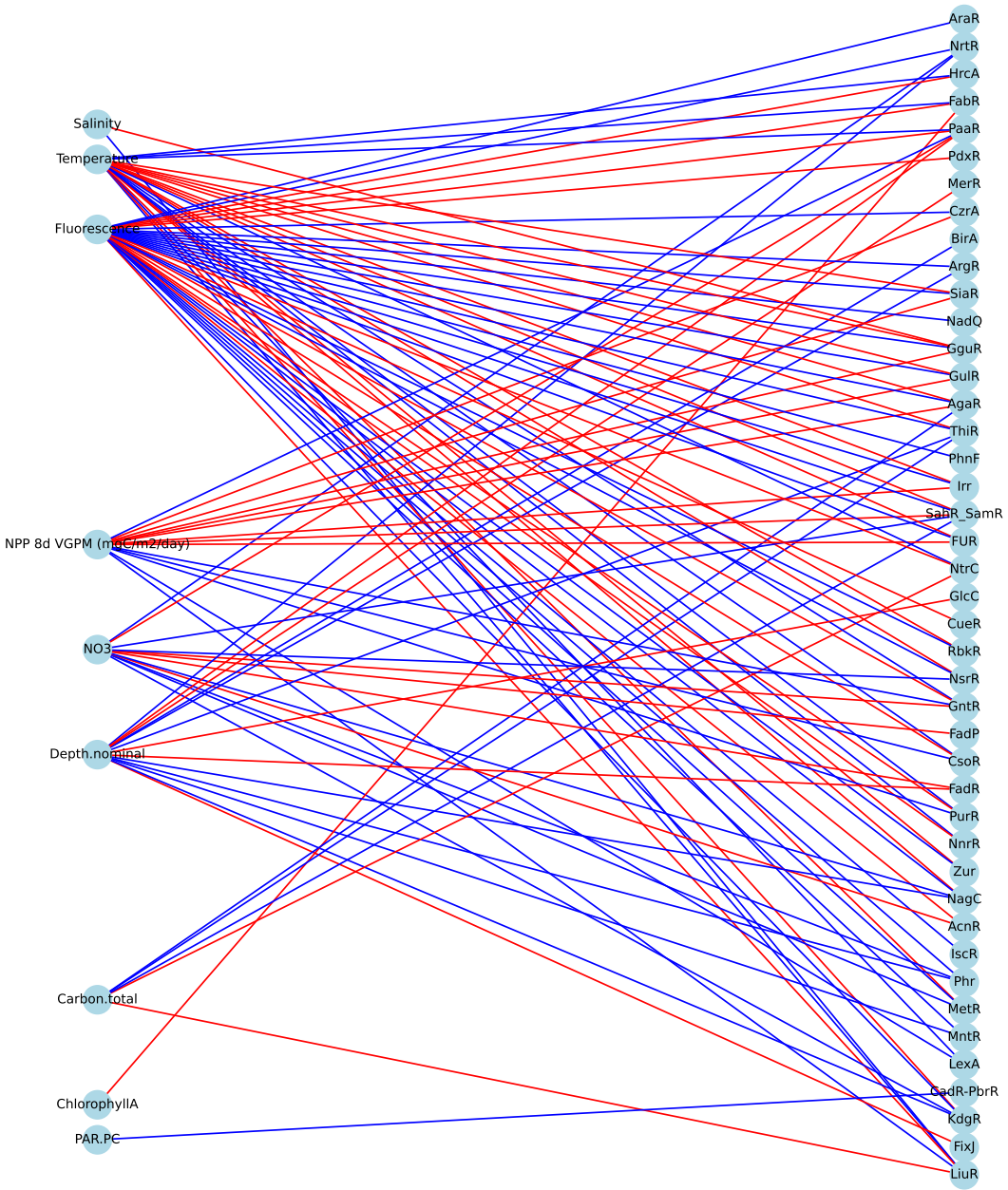


Figure 4.8: Biotic-Abiotic Correlation Bipartite Network. The network showcases the significant relationships (absolute correlation above 0.5) between abiotic variables and transcription factors. In this bipartite graph, nodes represent both biotic and abiotic factors, with each set on opposite sides of the graph to distinguish their categories. Edges indicate their correlations, with red signifying positive correlations, and blue signifying negative ones. This bipartite network visualization offers a deeper understanding of the intricate connections in marine ecosystems, emphasizing the direct associations that might impact genetic regulatory networks.

As anticipated from the correlation heatmap, key abiotic variables—Fluorescence, Temperature, Total Carbon, NPP, PAR.PC, NO3, Chlorophyll-a, Salinity, and Depth—manifest strong correlations with numerous transcription factors. For a detailed breakdown of their individual relationships with specific transcription factors, refer to Appendix C

4.4. Layer and Polarity-Specific Correlations

The analyses conducted up to this point have considered all samples as an integral whole. This approach provides an overarching view of the broad structure and dynamics within the oceanic biome. While this holistic perspective is insightful, it might inadvertently overlook the complexities inherent within different ocean strata or specific geographical locations. The ocean, as a biome, is not uniform—it is stratified and influenced by localized factors, necessitating an in-depth, segmented analysis to uncover the intricate nuances at these scales.

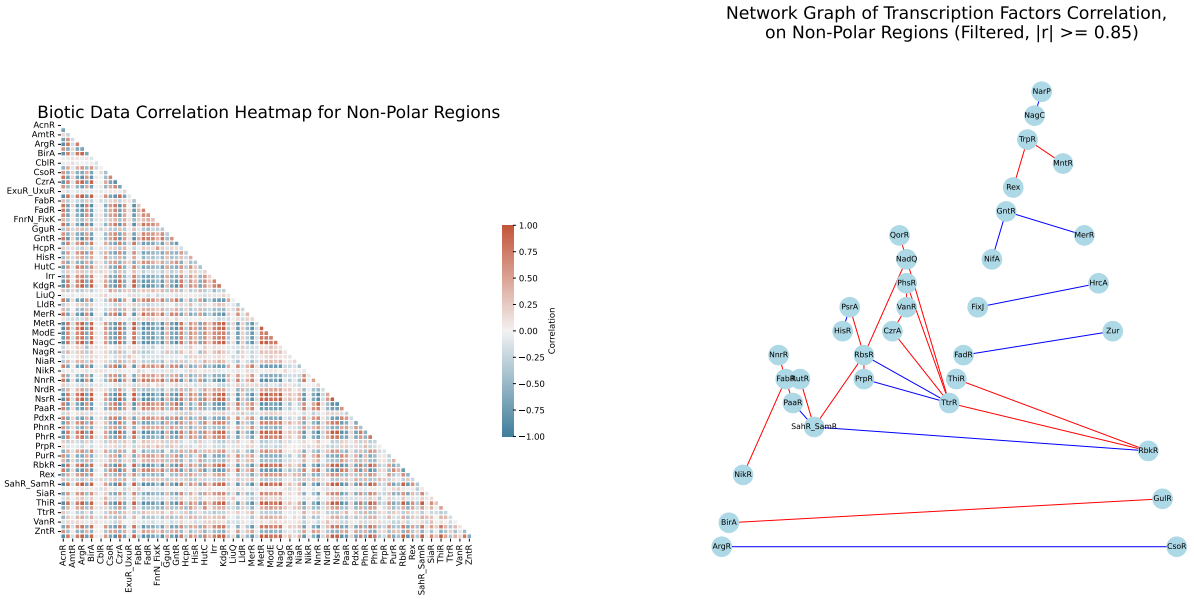
Pursuing a more granular, Layer/Polarity-specific investigation enriches our understanding of the oceanic ecosystem, providing a detailed portrayal of the biological and environmental interplays. This method aligns well with current literature that underscores the importance of understanding the ocean at these granular scales [34]. Accordingly, the subsequent sections undertake layer- and polarity-specific correlation analyses to delve into these finer details within our dataset.

4.4.1. Polar differentiation

In this subsection, we examine the clear distinctions observed in samples collected from polar and non-polar regions. Due to the significant differences in temperature, chemistry, and life forms between these environments, a separate analysis is essential.

Non-Polar regions

We begin this section by examining the Correlation Heatmap and Correlation Network for the non-polar regions. Using the biotic data, we apply a correlation threshold of 0.85 to the transcription factors, as illustrated in Figure 4.9:



(a) (b)

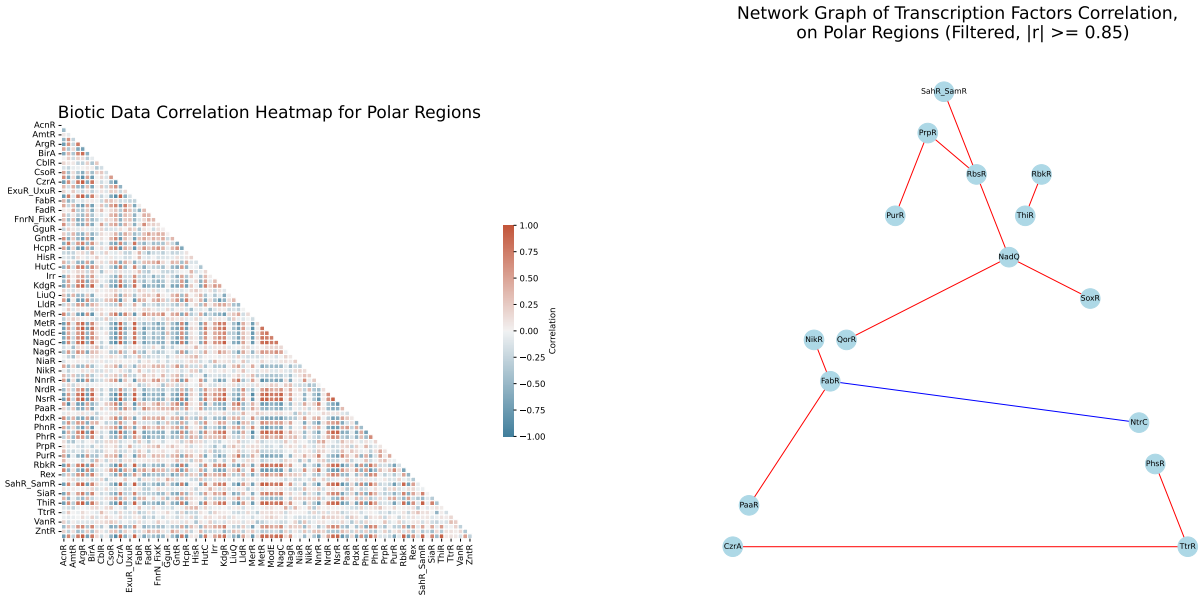
Figure 4.9: Detailed Visualization of Biotic Data Correlation for Non-polar Samples. **(a)** Shows a heatmap of the Spearman correlation matrix computed from the CLR-normalized transcription factors abundance dataset. The color scale represents correlation values, ranging from -1 (blue, indicating negative correlation) to 1 (red, indicating positive correlation). **(b)** Illustrates a planar network graph highlighting the strong correlations (above 0.85) among the transcription factors. The nodes represent transcription factors, while the edges symbolize significant correlations, with red signifying positive correlations and blue indicating negative correlations.

From the Figure 4.9, it is evident that there are several connected components[14]. Notably, one major component mirrors the pattern observed in the comprehensive analysis of the entire sample set, indicating a consistent connectivity structure regardless of the non-polar sample consideration. However, due to potential sample distribution imbalances between nonpolar and polar regions, drawing definitive conclusions requires caution.

Polar regions

We continue with the same analysis, but for the polar regions, shown in Figure 4.10:

[14] In graph theory, a connected component is a group of nodes in a graph that are linked to each other by paths, and isolated from other such groups.



(a) (b)

Figure 4.10: Detailed Visualization of Biotic Data Correlation for polar Samples. **(a)** Shows a heatmap of the Spearman correlation matrix computed from the CLR-normalized transcription factors abundance dataset. The color scale represents correlation values, ranging from -1 (blue, indicating negative correlation) to 1 (red, indicating positive correlation). **(b)** Illustrates a planar network graph highlighting the strong correlations (above 0.85) among the transcription factors. The nodes represent transcription factors, while the edges symbolize significant correlations, with red signifying positive correlations and blue indicating negative correlations.

From the Figure 4.10 examining Polar regions, we observe a sparse correlation network with small connected components at a high threshold of 0.85. This indicates robust yet limited strong relationships among certain variables. These tight correlations might be due to unique Polar environmental conditions. Comparatively few strong interconnections suggest that many variables might operate independently in these regions. Understanding this network's topology offers insights into the Polar ecosystem's resilience and adaptability, emphasizing the need for comparative studies with non-Polar regions and further targeted research on the identified components.

Correlation Analysis of Biotic and Abiotic Variables Across Oceanic Regions

Expanding our analysis, we now aim to understand how the variations in transcription factors correlate with changes in the abiotic variables within non-polar and polar regions.

We start with the non-polar regions, illustrating the relationship between these two realms by plotting their respective correlation heatmap, shown in Figure 4.11 and network, shown in Figure 4.12

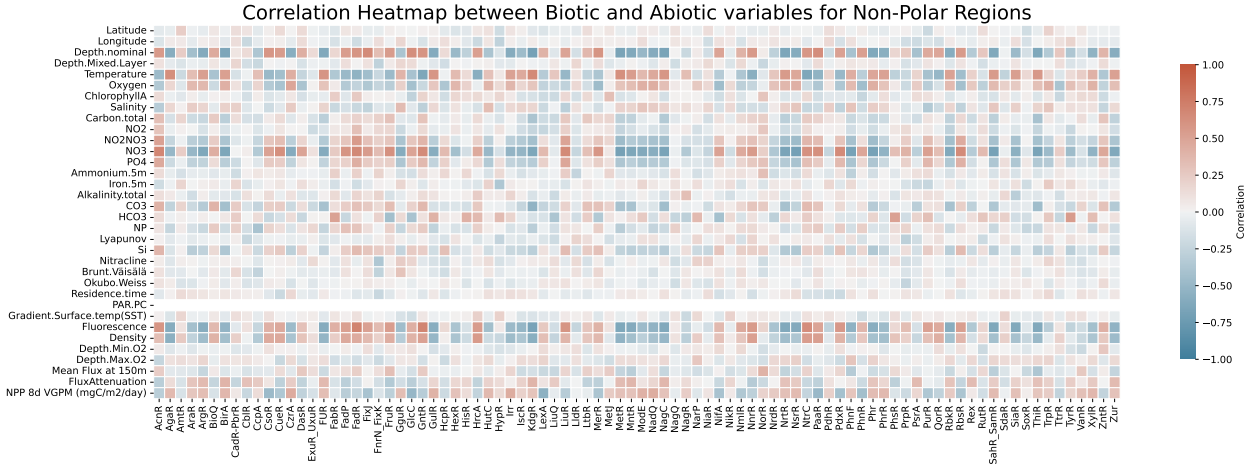


Figure 4.11: Biotic-Abiotic Correlation Heatmap for Non-Polar Samples. The heatmap illustrates the Spearman correlation coefficients between Transcription Factors (TFs, representing biotic data) and environmental parameters (abiotic data) for non-polar marine samples. The color gradient ranges from -1 (indicating a negative correlation) to +1 (indicating a positive correlation).

The abiotic variable *PAR.PC* is not correlated with any transcription factor, due to missing data. This observation is clearer in Figure 3.30. While one could consider using an imputation algorithm to address this, our preference for this analysis was to work with the most unaltered data possible. This approach was adopted to maintain consistency with the available literature.

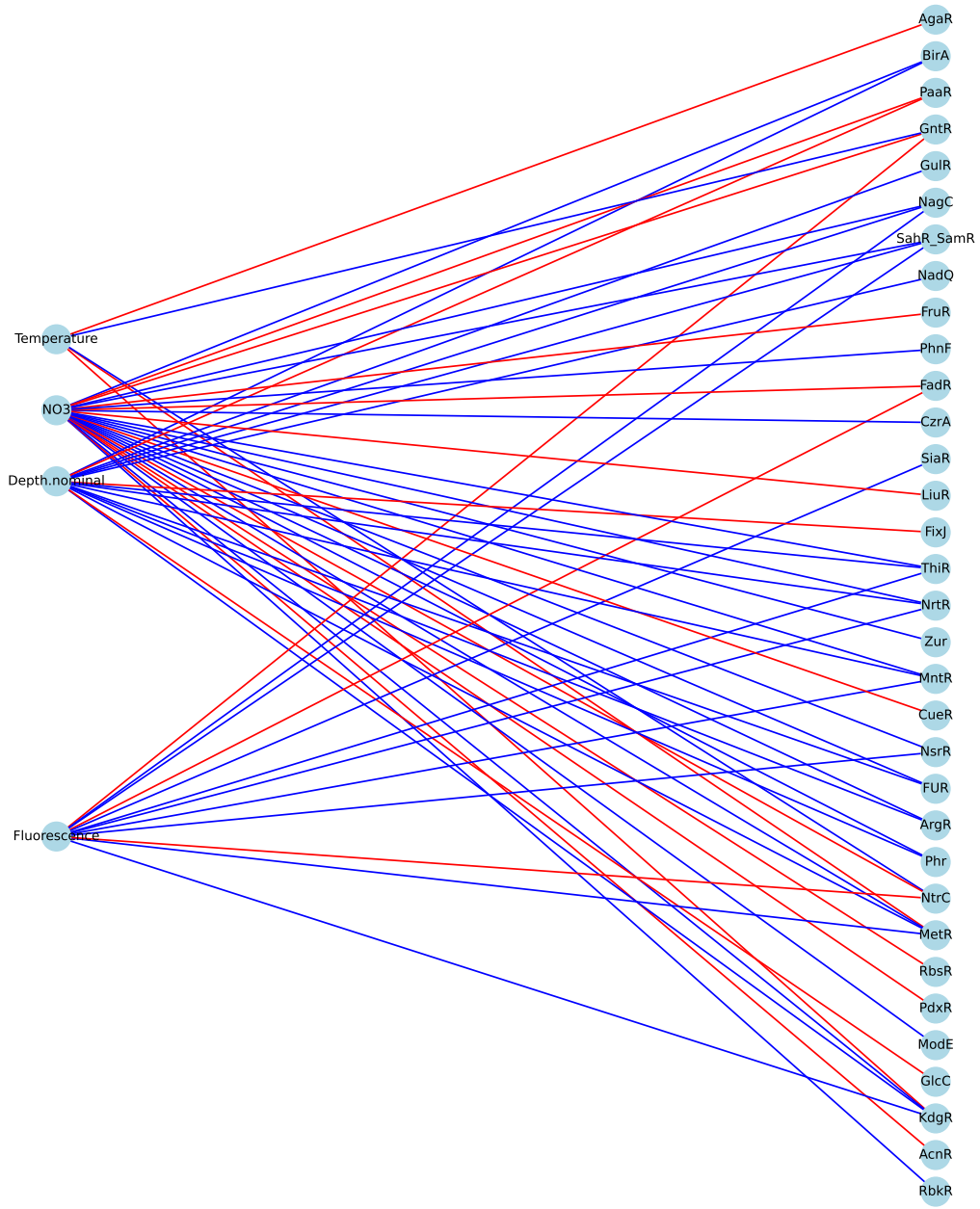


Figure 4.12: Biotic-Abiotic Correlation Bipartite Network for Non Polar regions. The network showcases the significant relationships (absolute correlation above 0.6) between abiotic variables and transcription factors. In this bipartite graph, nodes represent both biotic and abiotic factors, with each set on opposite sides of the graph to distinguish their categories. Edges indicate their correlations, with red signifying positive correlations, and blue signifying negative ones.

Depth, NO3 concentration, fluorescence, and temperature stand out as the primary abiotic variables exhibiting a strong correlation (abve 0.6) with transcription factors. This suggests that these environmental parameters play pivotal roles in influencing gene expression in marine organisms. By focusing on the transcription factors correlated with depth, we may gain a deeper understanding of how the stratification and ecological zones of the ocean influence marine biology and biogeochemistry, which will be later discussed in more detail.

Now we do the same for the polar regions:

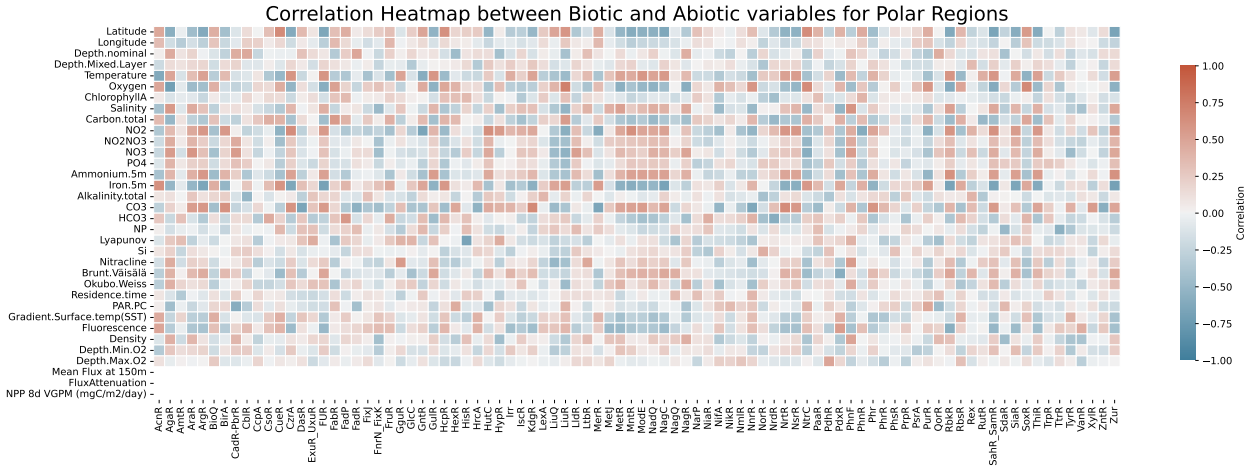


Figure 4.13: Biotic-Abiotic Correlation Heatmap for Polar Samples. The heatmap illustrates the Spearman correlation coefficients between Transcription Factors (TFs, representing biotic data) and environmental parameters (abiotic data) for non-polar marine samples. The color gradient ranges from -1 (indicating a negative correlation) to +1 (indicating a positive correlation).

From the heatmap presented in Figure 4.13, the heterogeneity in correlations between biotic and abiotic variables in polar regions becomes strikingly evident. This diverse interplay underscores the complex nature of interactions within polar marine ecosystems.

The intricacies of these relationships can be more effectively visualized and dissected when portrayed in a correlation network (Figure 4.14), providing a clearer understanding of the interconnected dynamics present in such extreme environments:

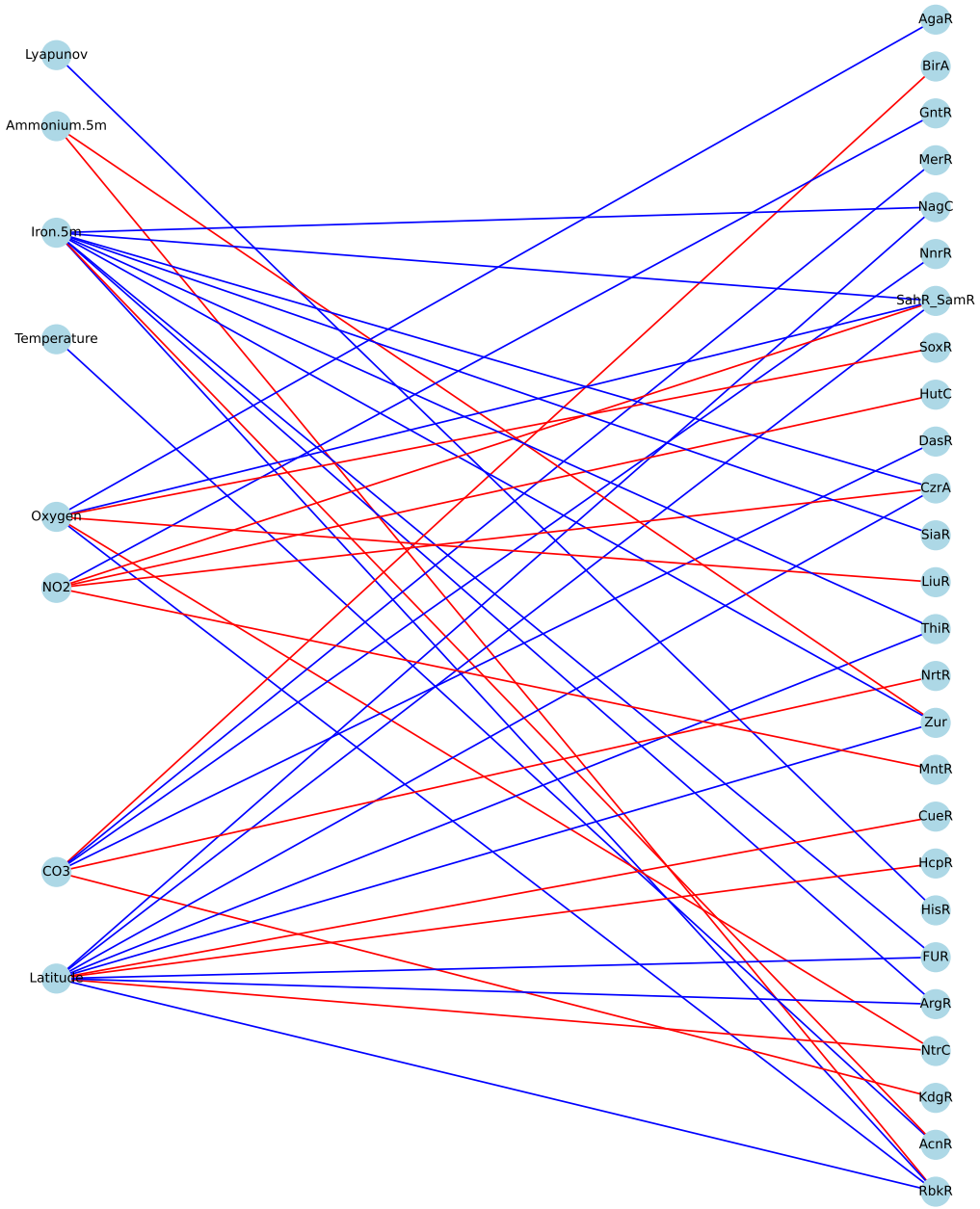


Figure 4.14: Biotic-Abiotic Correlation Bipartite Network for Polar regions. The network showcases the significant relationships (absolute correlation above 0.5) between abiotic variables and transcription factors. In this bipartite graph, nodes represent both biotic and abiotic factors, with each set on opposite sides of the graph to distinguish their categories. Edges indicate their correlations, with red signifying positive correlations, and blue signifying negative ones.

In the correlation network illustrated in Figure 4.14, the marked dispersity—or heterogeneity—becomes more pronounced. There is a noticeable trend of fewer transcription factors being correlated with a broader array of environmental variables. Notably, latitude stands out as having a strong correlation with certain transcription factors. Given that our focus is on polar regions, it can be deduced that the abundance of these biotic variables in these regions is particularly sensitive to latitude shifts.

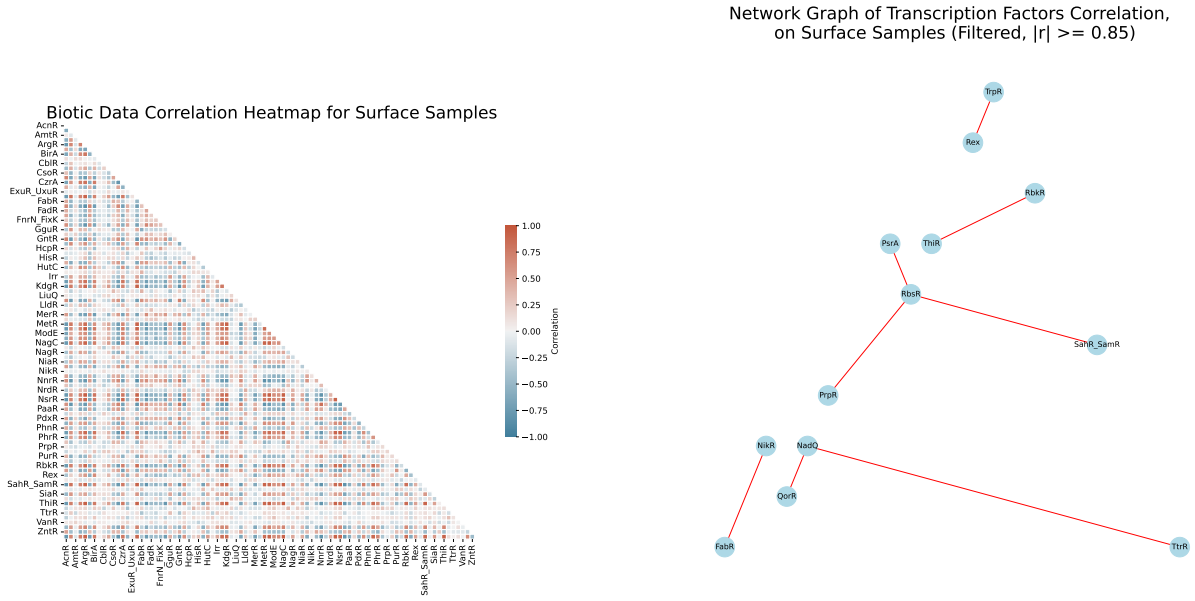
4.4.2. Layer Stratification in Oceanic Samples

The ocean is a layered realm, with each stratum reflecting unique biological and environmental contexts, primarily determined by depth. In our study, we have segmented these depths into three prominent layers: the Surface (SRF), the Deep Chlorophyll Maximum (DCM), and the Mesopelagic (MES). For a broader perspective, these individual layers can be further integrated into zones: the Epipelagic (EPI), composed of the SRF and DCM, and the aforementioned Mesopelagic region.

Each layer's distinctive features will be explored in isolation using correlation studies, aiming to discern how oceanic stratification shapes our data and brings out the peculiarities of each stratum.

Surface Layer (SRF)

Figure 4.15 portrays the biotic data correlation analysis for surface samples:



(a) Heatmap of Spearman Correlation for SRF samples.

(b) Network Graph for SRF samples.

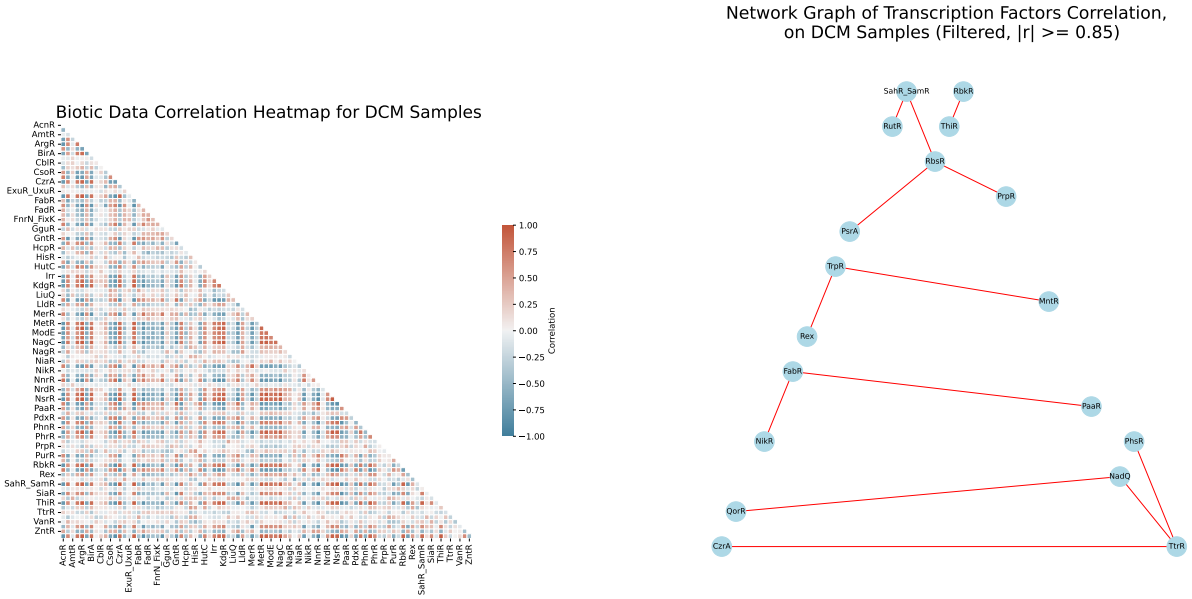
Figure 4.15: Biotic Data Correlation for Surface (SRF) Samples: **(a)** Heatmap displaying the Spearman correlation matrix from the CLR-normalized transcription factor abundance dataset. Correlation values vary from -1 (blue) for negative correlation to 1 (red) for positive. **(b)** Network graph showing significant correlations (above 0.85) among transcription factors. Nodes depict transcription factors; edges represent correlations, colored red for positive and blue for negative associations.

In observing the correlation network presented in Figure 4.15, its topology bears a striking

resemblance to the network derived from the biotic correlations in polar samples shown in Figure 4.10.b. Notably, they both highlight several shared transcription factors. This observation suggests that transcription factors such as 'TtrR', 'RbsR', 'NadQ', 'QorR', 'FabR', and 'NIkR' play a vital role not just in specific regions but on a global scale. However, given the underrepresentation of mesopelagic samples in polar regions, it is imperative to approach further hypotheses with caution. Nevertheless, the recurring prominence of these transcription factors across diverse datasets underscores their significance and positions them as focal points in future research endeavors.

Deep Chlorophyll Maximum Layer (DCM)

The DCM layer's results are encapsulated in Figure 4.16:



(a) Heatmap of Spearman Correlation for DCM samples.

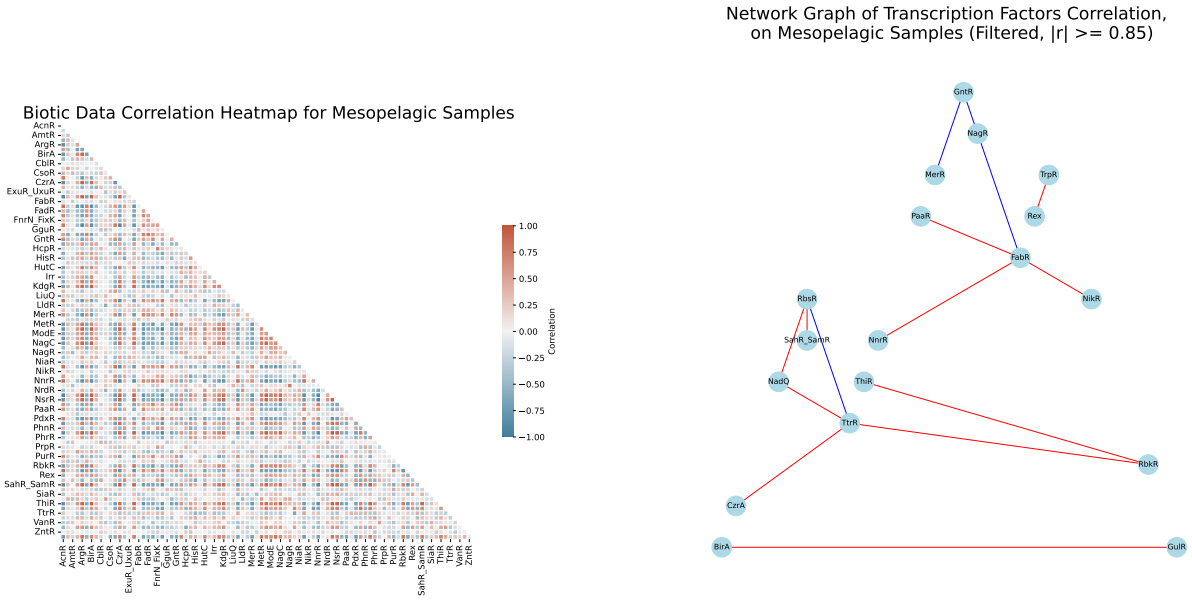
(b) Network Graph for DCM samples.

Figure 4.16: Biotic Data Correlation for Deep Chlorophyll Maximum (DCM) Samples: **(a)** Heatmap of the Spearman correlation matrix derived from CLR-normalized transcription factor abundances. Correlation values span from -1 (blue, negative) to 1 (red, positive). **(b)** Network graph spotlighting marked correlations (greater than 0.85) among transcription factors. Nodes symbolize transcription factors; edges signify their correlations, with red for positive and blue for negative ties.

From Figure 4.16.b, it is evident that the homogeneity comes from the epipelagic zone in terms of its correlations, as its network topology closely aligns with that of Figure 4.15.b. Additionally, there is a noticeable overlap in the transcription factors between the two figures, further emphasizing this similarity.

Mesopelagic Layer (MES)

Mesopelagic results are illustrated in Figure 4.17:



(a) Heatmap of Spearman Correlation for MES samples.

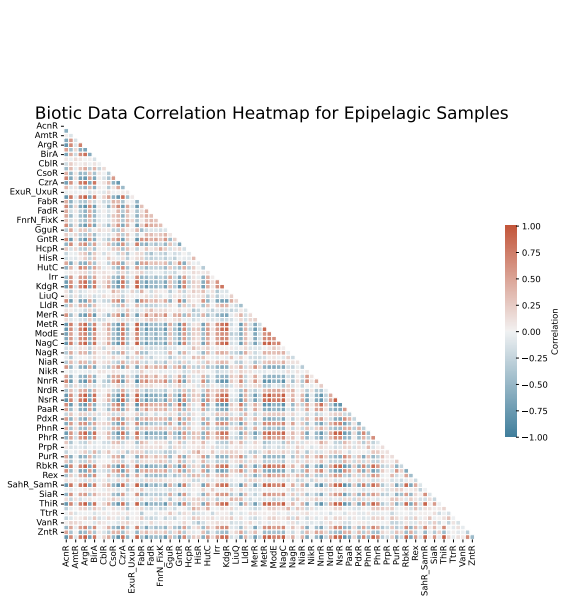
(b) Network Graph for MES samples.

Figure 4.17: Biotic Data Correlation for Mesopelagic (MES) Samples: **(a)** Heatmap presents the Spearman correlation matrix from the CLR-normalized transcription factor dataset. The color gradient from blue (-1) to red (1) shows negative to positive correlations. **(b)** Network graph illustrating pronounced correlations (exceeding 0.85) among transcription factors. Nodes stand for transcription factors, while edges indicate their significant correlations; red lines suggest positive while blue signifies negative connections.

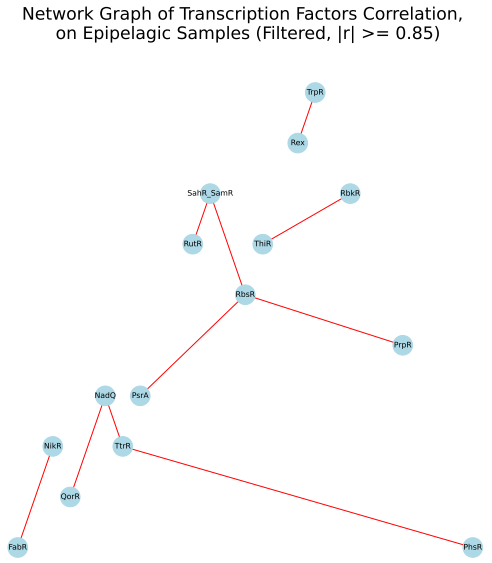
From Figure 4.17.b, a striking distinction in the network topology becomes evident when compared to Figures 4.15.b and Figure 4.16.b. Yet, what is particularly remarkable is the persistence of certain overlapping transcription factors that exhibit strong correlations within this oceanic layer. This consistency may suggest that these transcription factors play a foundational role in maintaining the biotic balance and interactions, irrespective of the varying environmental conditions across different layers. Further research could delve into understanding the underlying mechanisms that make these transcription factors ubiquitous and crucial in the ocean's stratified environment. These are: 'TtrR', 'FabR', 'NikR' and 'SahR_SamR'.

Epipelagic Zone (EPI)

Lastly, the Epipelagic zone, encompassing both SRF and DCM layers, is captured in Figure 4.18:



(a) Heatmap of Spearman Correlation for EPI samples.



(b) Network Graph for EPI samples.

Figure 4.18: Biotic Data Correlation for the Epipelagic (EPI) Zone: **(a)** Heatmap portrays the Spearman correlation matrix for CLR-normalized transcription factor abundance. The spectrum goes from blue (-1, negative correlation) to red (1, positive correlation). **(b)** Network graph delineating strong correlations (above 0.85) among transcription factors. Nodes represent transcription factors, with edges drawing their significant correlations; edges in red mark positive while those in blue denote negative associations.

In both the SRF correlation network (Figure 4.15.b) and the DCM correlation network (Figure 4.16.b), the topologies are strikingly similar to those observed in the collective epipelagic samples. This consistency hints at the epipelagic zone's relative uniformity in terms of the regulatory network. The inherent homogeneity within this zone might be a significant pointer, suggesting it could be pivotal for uncovering new insights in subsequent research.

Correlation Analysis of Biotic and Abiotic Variables Across Oceanic Layers

Understanding the relationship between transcription factors is essential, as highlighted in the previous section. However, it is equally crucial to examine their interactions with the environment. Given that the ocean is characterized by its stratified layers, our interest lies in exploring these relationships within each of the aforementioned layers.

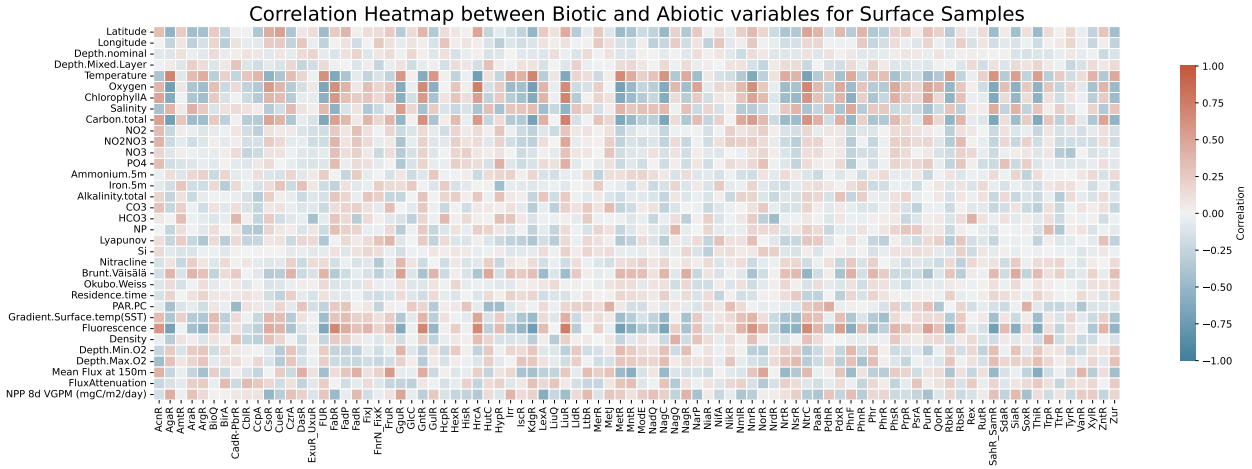


Figure 4.19: Spearman correlation heatmap visualizing relationships between biotic and abiotic variables in the Surface (SRF) ocean layer. Each cell's color intensity represents the strength and direction of the correlation, with a color scale ranging from -1 (blue) indicating a perfect negative correlation to +1 (red) indicating a perfect positive correlation.

Figure 4.19 offers an overarching perspective on the interplay between biotic and abiotic variables. Distinctly, certain environmental data exhibit a pronounced association with all transcription factors.

To explore these connections in more detail, as we have previously undertaken, it is essential to present the correlation network.

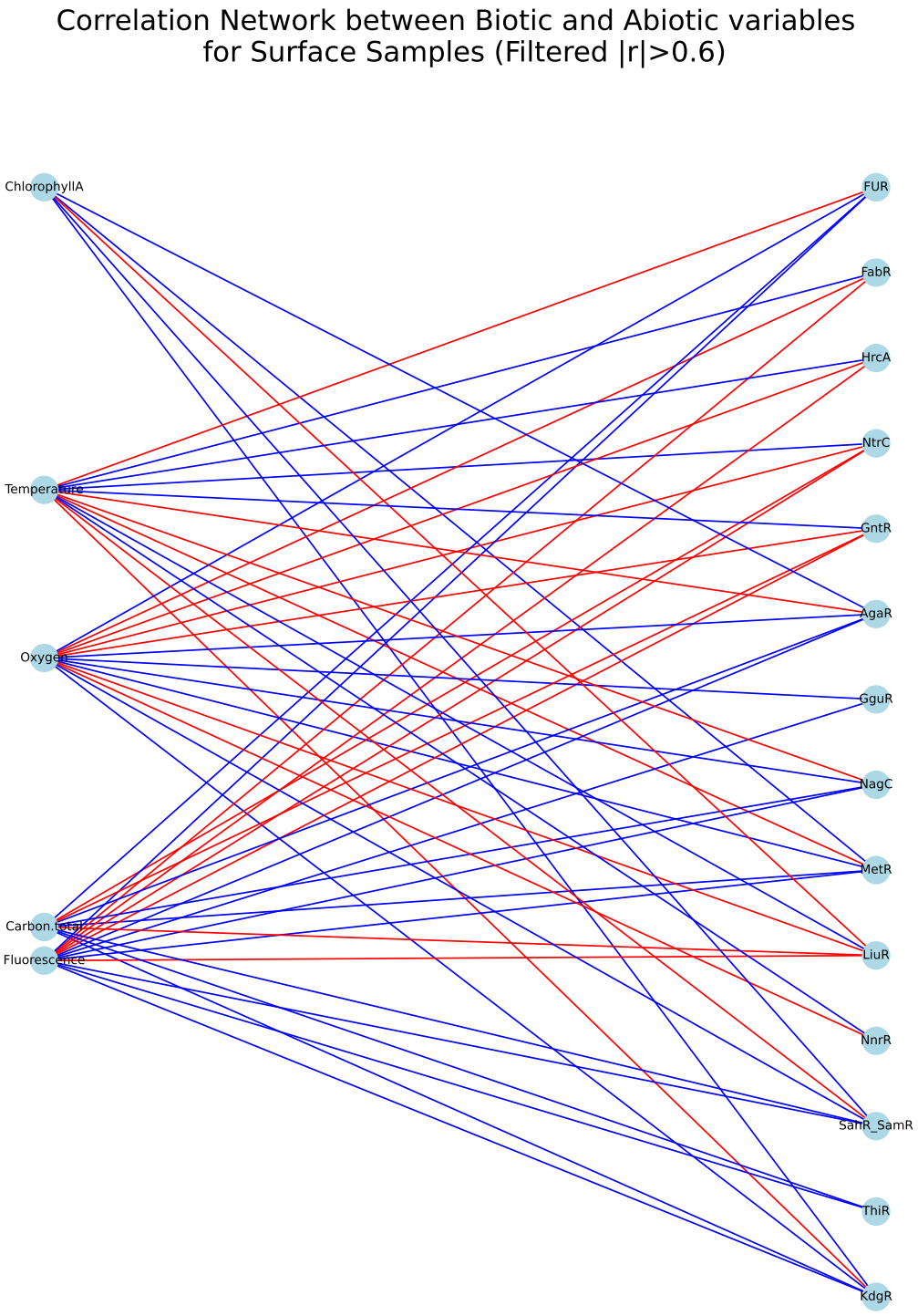


Figure 4.20: Bipartite correlation network depicting relationships between biotic (transcription factors) and abiotic (environmental features) variables in the Surface (SRF) ocean layer. Nodes represent environmental features (LHS) and transcription factors (RHS), while edges, colored either red (positive correlation) or blue (negative correlation), signify strong correlations with an absolute value greater than the set threshold of 0.6. Only nodes with at least one significant correlation are displayed for clarity.

In the analysis of surface samples, it becomes evident that temperature significantly correlates with certain transcription factors. Furthermore, core components of the ocean ecosystem, such as oxygen levels and total carbon, emerge as pivotal players. This interplay between biotic and abiotic factors is something to consider: for instance, recent studies, like that of Guidi 2016 [10] and Kaneko 2021 [35], have underscored that biotic components can even steer certain abiotic phenomena. Guidi's groundbreaking research highlighted that specific plankton communities, a biotic component, can correlate strongly with carbon export in the epipelagic layer. More so, Kaneko's findings assert that eukaryotic virus composition can predict the efficiency of carbon export. These revelations underscore the intricacies of these interactions and the necessity of delving deeper into understanding them. Recognizing such relationships provides a holistic view of how marine ecosystems function, adapt, and influence abiotic processes, making it paramount to keep these transcription factors central in our ongoing research and discussions.

The DCM layer's correlation heatmap results are encapsulated in Figure 4.21

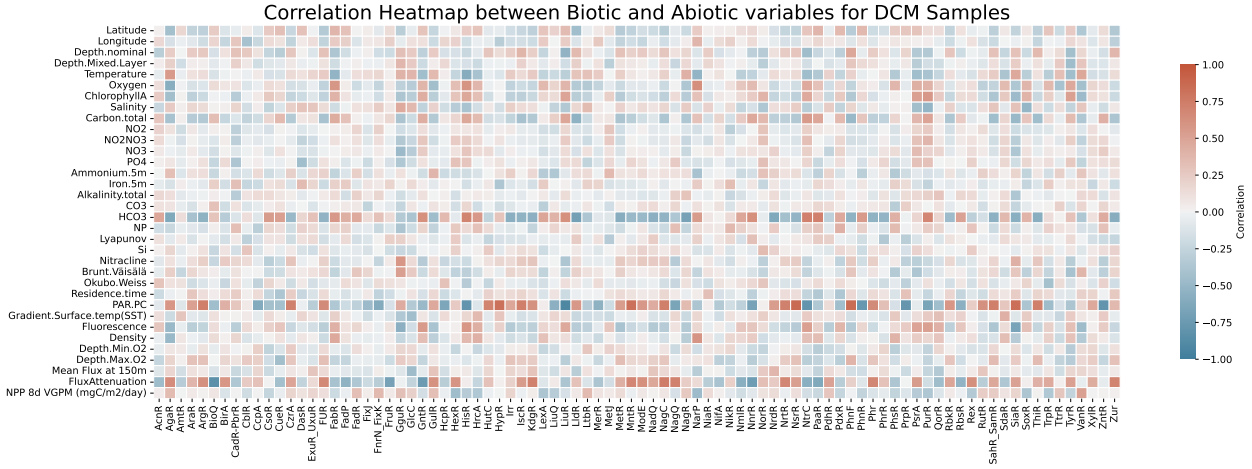


Figure 4.21: Spearman correlation heatmap showcasing the relationships between biotic (transcription factors) and abiotic (environmental features) variables within the Deep Chlorophyll Maximum (DCM) ocean layer. Each cell in the heatmap represents the correlation coefficient between a given pair of variables, with the color intensity and direction (blue for negative and red for positive) indicating the strength and nature of the correlation.

To delve deeper into the associations presented in Figure 4.21, we will showcase the corresponding correlation network.

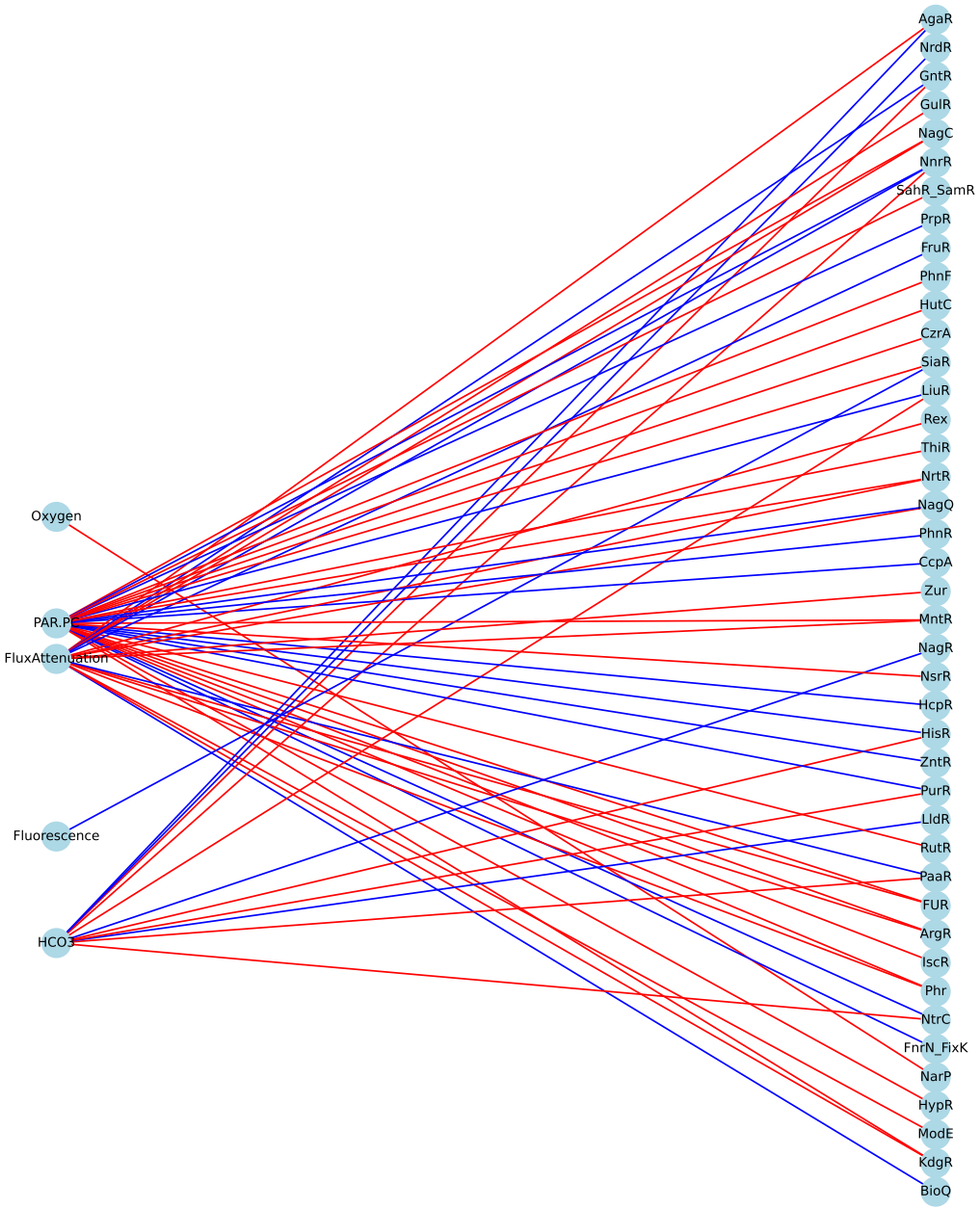


Figure 4.22: Bipartite correlation network depicting the relationships between biotic (transcription factors) and abiotic (environmental features) variables within the Deep Chlorophyll Maximum (DCM) layer. Nodes represent environmental features (LHS) and transcription factors (RHS), while edges signify correlations with magnitudes greater than the threshold of $|r| > 0.6$. The edge color indicates the nature of the correlation: red for positive and blue for negative.

In our observations from Figure 4.22, temperature no longer emerges as an abiotic variable that correlates with transcription factors when compared to Figure 4.20. Additionally, there is a noticeable decrease in environmental variables that maintain such correlations. Interestingly, Flux Attenuation stands out as it connects with certain transcription factors. This is particularly significant as Flux Attenuation pertains to the attenuation of carbon export in the ocean.

Mesopelagic results for the correlation heatmap are illustrated in Figure 4.23:

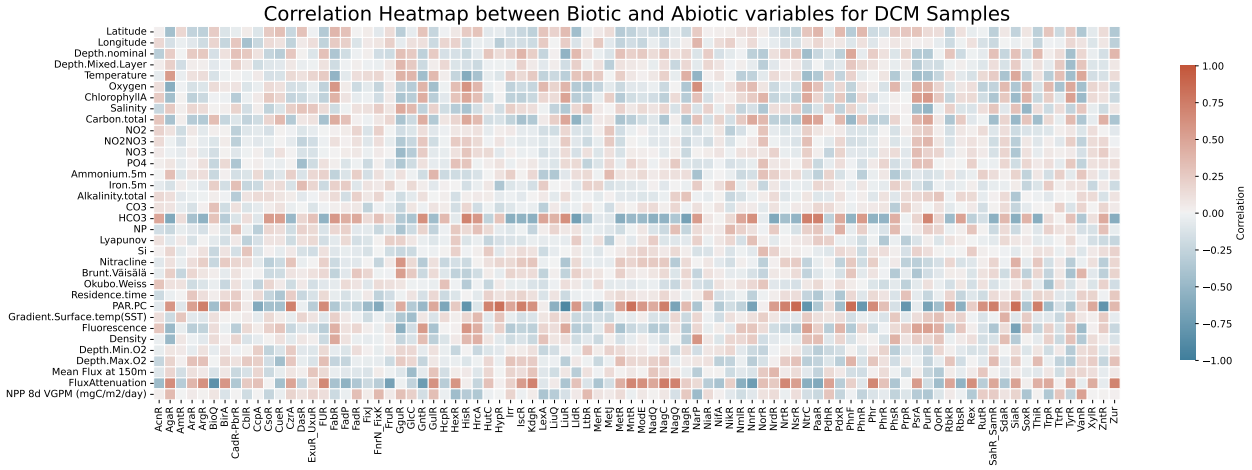


Figure 4.23: Heatmap representation of the correlations between biotic and abiotic variables within the mesopelagic (MES) layer. Each cell in the heatmap illustrates the Spearman correlation coefficient between corresponding biotic (e.g., transcription factors) and abiotic (e.g., environmental features) variables. The color gradient, spanning from blue (negative correlation) to red (positive correlation), provides a visual cue for the strength and direction of each correlation.

Once more, due to the absence of data for Total Alkalinity in mesopelagic samples, its correlation remains undisplayed. However, it is noteworthy that CO3 emerges as a predominant factor correlating with transcription factors (Figure 4.23). To delve deeper into these findings, we will present a correlation network graph for a more detailed exploration.

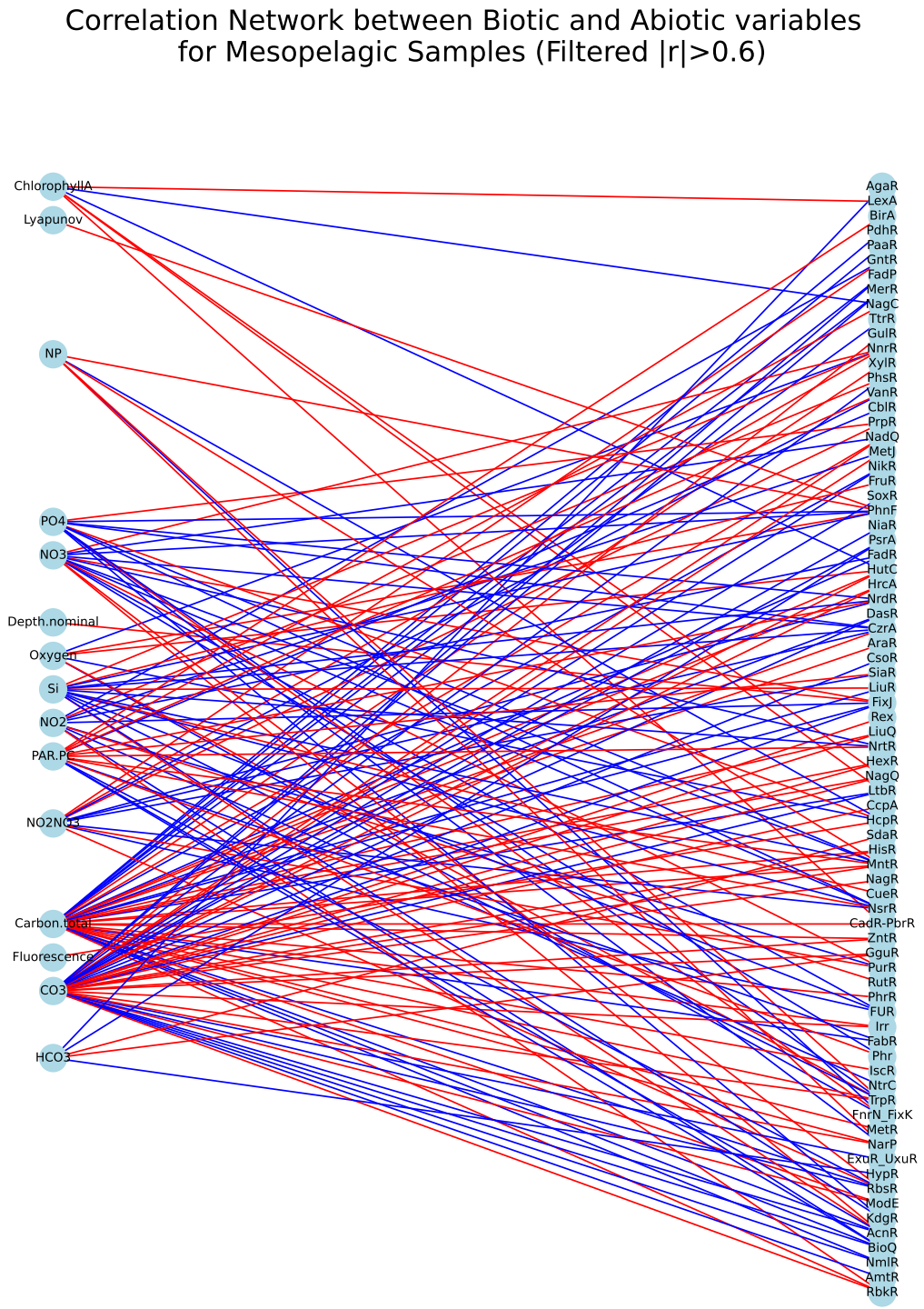


Figure 4.24: Network visualization of strong correlations ($|r| > 0.6$) between biotic and abiotic variables in the mesopelagic (MES) zone. Nodes in the graph represent environmental features (LHS) and transcription factors (RHS), and the edges between them indicate significant correlations. The edge colors differentiate positive (red) from negative (blue) correlations.

In Figure 4.24, it becomes evident that environmental variables related to nutrient availability stand out for their strong correlations with various transcription factors. This observation aligns with the known characteristics of mesopelagic zones, which are renowned for being nutrient-rich due to the significant respiration and remineralization of organic particles that occur within them [36].

Finally we address the same analysis in the Epipelagic zone.

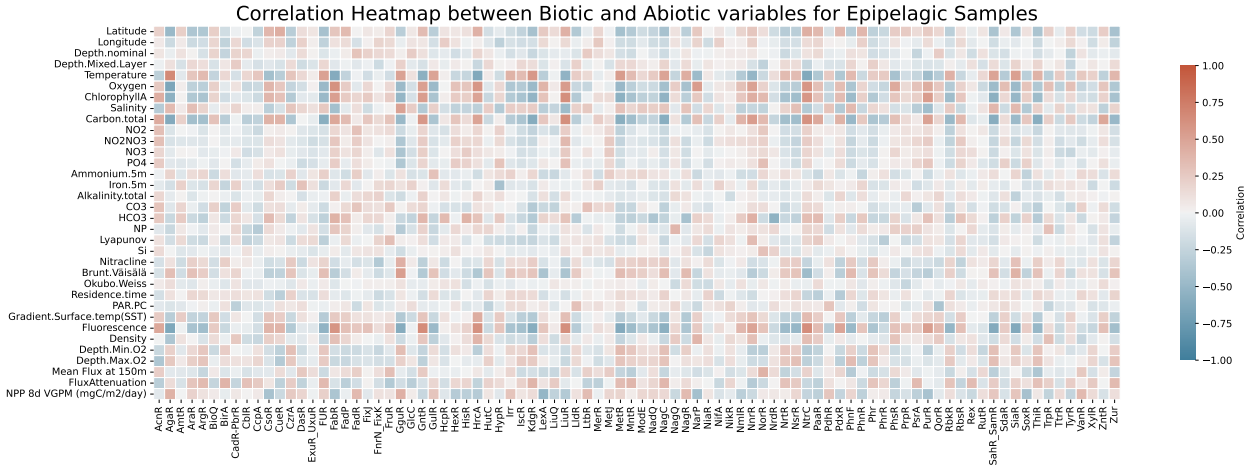


Figure 4.25: Correlation heatmap highlighting associations between biotic and abiotic variables within the epipelagic (EPI) zone, derived from surface (SRF) and deep chlorophyll maximum (DCM) samples. The colormap ranges from -1 (blue) indicating strong negative correlations, to 1 (red) indicating strong positive correlations, with neutral associations in white.

From Figure 4.25, it is apparent that there is a relatively weak overall correlation between the biotic and abiotic variables in the Epipelagic (EPI) samples, contrasting sharply with the pronounced correlations observed in Figure 4.19 and Figure 4.21 for SRF and DCM samples respectively. This disparity can be attributed to the unique environmental variables in each oceanic layer that strongly correlate with different transcription factors. As previously done, we will plot the correlation network

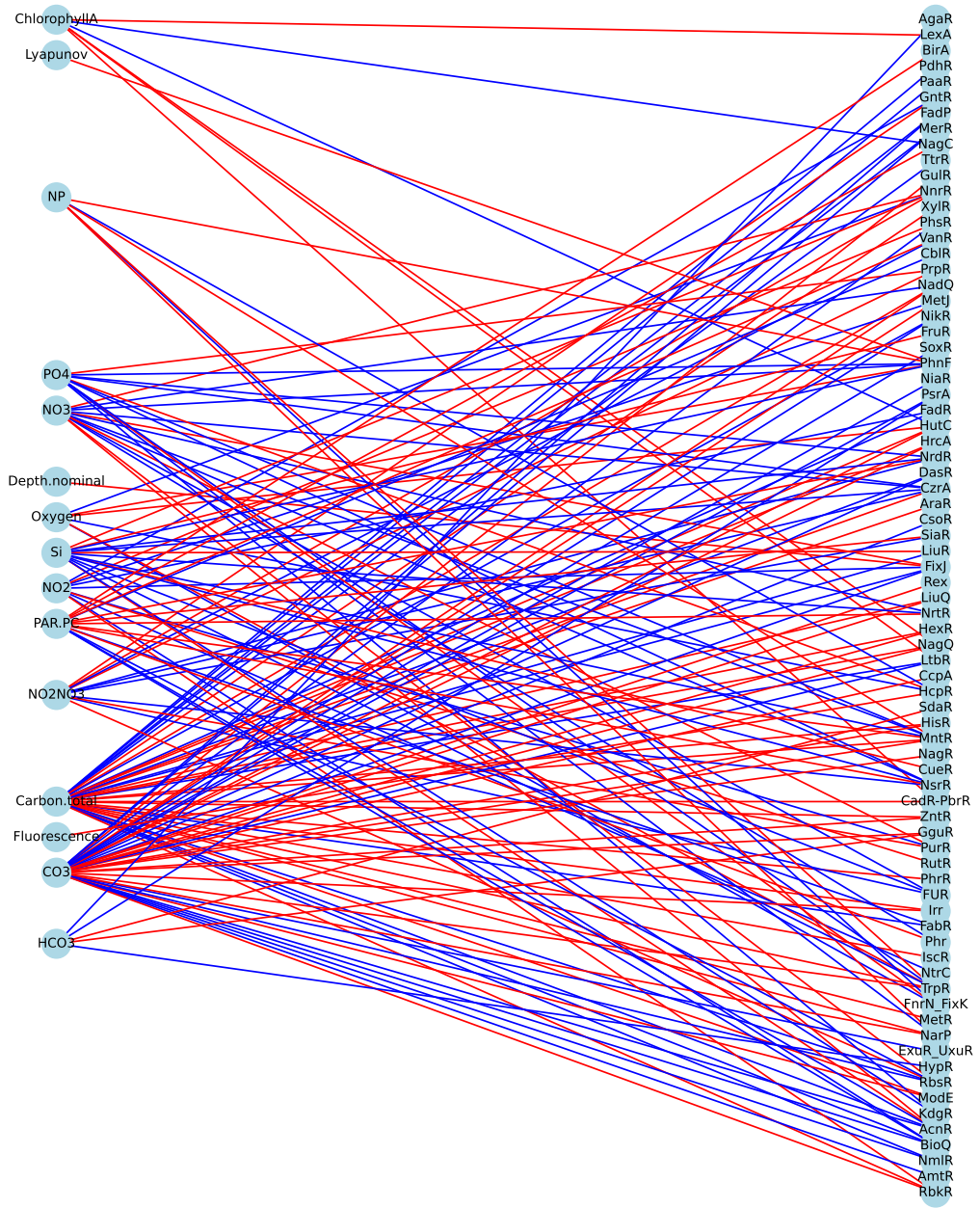


Figure 4.26: Bipartite correlation network visualization for the mesopelagic (MES) zone, showcasing the significant relationships ($|r| > 0.6$) between environmental features and transcription factors. Nodes represent both the biotic (transcription factors in the RHS) and abiotic (environmental in the LHS) variables, while edges colored in red indicate positive correlations and those in blue represent negative correlations.

Consistent with our anticipation from Figure 4.25, the correlation network depicted in Figure 4.26 reveals weaker correlations between biotic and abiotic variables. However, it is noteworthy that the environmental features identified here overlap significantly with those highlighted in the surface correlation network.

4.5. Summary

The correlations obtained in this chapter allow us to get an idea of how the variables are related, obtaining the first structural result of this work. However, the large amount of information makes it difficult to interpret and easy to explain. For this reason, we will focus on environmental variables that we have mentioned earlier are of special interest: latitude, temperature and depth; and their categorical counterparts: polarity and layer.

To produce this result, we will condense the information by showing those transcription factors most correlated[15] ($|\text{correlation}| > 0.5$), in different scenarios, with the aforementioned variables. Those who surpass this threshold will be marked as shown in Figure 4.27 and Figure 4.28. In addition, we've color-coded the transcription factors based on their cluster label, restricting our focus to four main clusters (See Appendix B for more information). Their functionalities are provided alongside for clarity, as illustrated in Figure 4.6.b.

[15] The correlation with the categorical variables was calculated using a point-biseral correlation

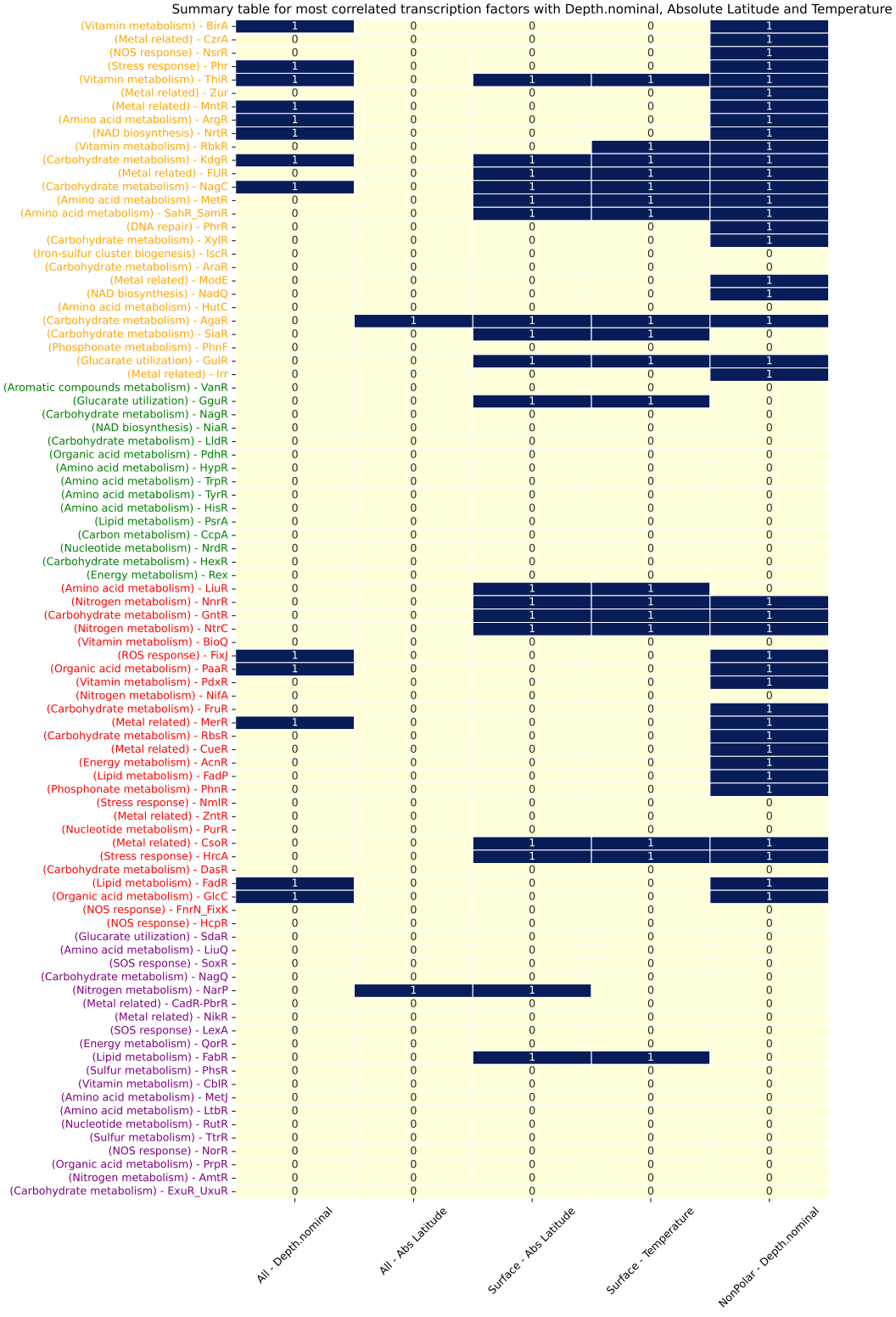


Figure 4.27: Visualization of transcription factors with strong correlations ($|\text{correlation}| > 0.5$ marked with a '1') vs continuous variables (depth, absolute latitude and temperture) across different sample categories (All: all samples; Surface: surface samples; NonPolar: non polar samples). Next to the transcription factors, their functional category is annotated.

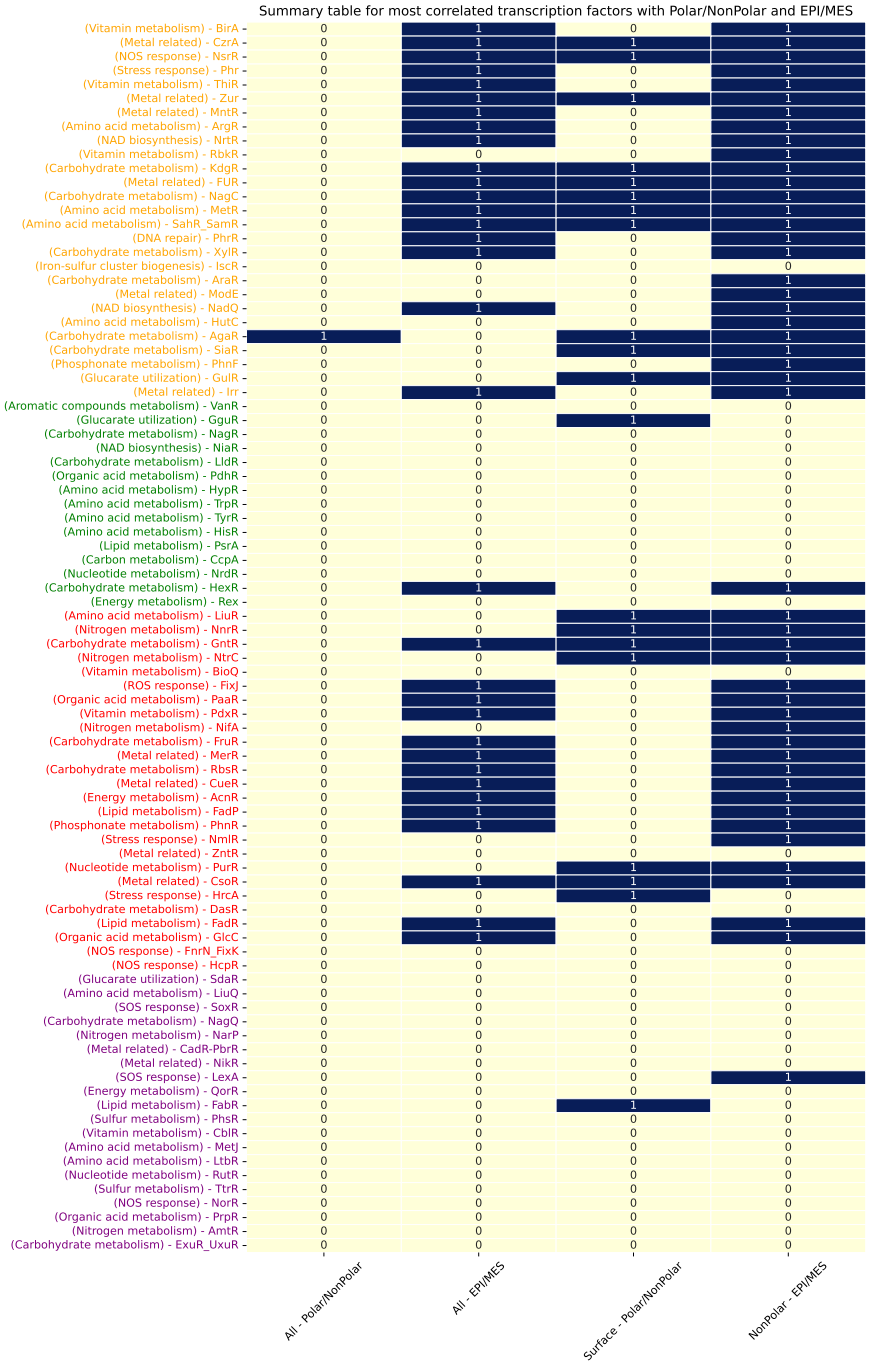


Figure 4.28: Visualization of transcription factors with strong point-biseral correlations (|correlation| > 0.5 marked with a '1') vs categorical variables (polarity and pelagic zone) across different sample categories (All: all samples; Surface: surface samples; NonPolar: non polar samples)

Clearly, 'Depth.nominal' and 'Abs latitude' are related to 'Layer' and 'polar' respectively, the latter ones being the categorical counterparts of the former. Temperature stands out since it drops in polar samples and goes down in deeper layers like mesopelagic samples. Therefore, we looked at temperature only in surface samples to avoid noise from the deeper layers.

Using the condensed information, we can conclude the following:

- When considering all the samples, the absolute latitude is strongly correlated (s.c.) with **AgaR** and **NarP**. However, the Polar / Non Polar category is only s.c. with **AgaR**, meaning that **NarP** is a good indicator of latitudinal changes in non polar samples only.

- When we look at all the samples, and how the transcription factors correlate with Depth.nominal and EPI/MES category, we see that there are much more biotic variables s.c. with the latter category than with the former continuous one. We can deduce that binding abundances change abruptly when considering different layer stratifications, in contrast when considering smooth changes in depth. This conclusion is remarkable, since climate change drives changes in the ocean stratification, thus, given this result, could also change the regulation mechanisms carried out in the ocean.

- It is evident that for surface samples, we have almost the same strongly correlated transcription factors with temperature and absolute latitude (except for **RbkR** and **NarP**, the first one s.c. with temperature and the second s.c. with absolute latitude). This emphasized the fact that temperature (in surface samples) is able to encapsulate the information of the absolute latitude. Moreover, temperature is also being able to encapsulate polar / non polar structural relations with transcription factors, since they differ only in a few s.c. transcription factors.

- Finally, when looking at Non Polar samples, for Depth.nominal and EPI/MES category, we deduce the same as when we considered all the samples, due to the higher amount of s.c. TFs with the latter category when compared to the former one. One last thing to notice tho, is that the structural relationship is enriched in this sub category of samples, since when all samples were considered, we had fewer s.c. TFs.

In adittion to the previous conclusions, we also managed to produce a list of transcription factors that are structural attached to the environmental variables of interest (however, this does not imply a direct cause-and-effect relationship). This is remarkable, since the abundance matrix starts showcasing the potential as a biological characterizer of the environment. Notably, only two clusters of TFs are s.c. with environmental data.

Chapter 5

Robust prediction of environmental variables of the ocean from transcription factor bindings

In this chapter, we delve deeper into the results, utilizing an array of analytical techniques to reveal patterns and insights within our data. Our focus is comprehensive, as we employ both supervised and unsupervised learning methods to understand the multifaceted nature of marine ecosystems.

We begin by implementing dimensionality reduction techniques, which allow us to simplify our intricate datasets without losing critical information. These methods are vital in condensing complex, high-dimensional data into more manageable forms, ensuring that the most salient features of our data are highlighted. This lays the groundwork for a clearer understanding of the complex dynamics at play within our marine ecosystems.

Next, we turn to unsupervised learning via clustering algorithms. Rather than imposing a structure on the data, these techniques allow the inherent patterns within the data to guide our analysis. By grouping similar instances together, we may uncover unanticipated relationships and connections, offering valuable insights that might not surface in a more directed analysis.

Lastly, we transition into a supervised learning approach by employing classification techniques. These methods provide a structured means of understanding how different factors contribute to the health and function of marine life, drawing on predictive models to delve into the underlying relationships among our variables.

Together, these methods form a comprehensive toolkit for exploring our data, shedding light on the complex dynamics of marine ecosystems. This chapter takes us on a thorough journey through our data, revealing the key patterns and trends that underpin our marine environments.

5.1. Dimensionality reduction

Understanding high-dimensional biological datasets, such as transcription factors, presents a significant challenge due to the 'curse of dimensionality.' This term refers to the complexities of interpreting and visualizing data in high-dimensional spaces. To address this, we utilize dimensionality reduction techniques. Our first approach is to employ Principal Component Analysis (PCA) using *Scikit-learn* library [37], one of the most widely accepted dimensionality

reduction techniques within the biological community. PCA simplifies the data structure by projecting it into a lower-dimensional space while preserving the original data's structure and relationships. This approach is particularly effective with compositional data, as it unveils crucial patterns in the relative abundance of transcription factors, thereby facilitating easier visualization and interpretation.

We will commence our analysis by scrutinizing the ratio of explained variance contributed by each individual principal component in the PCA transformed[16] dataset. This is shown in Figure 5.1

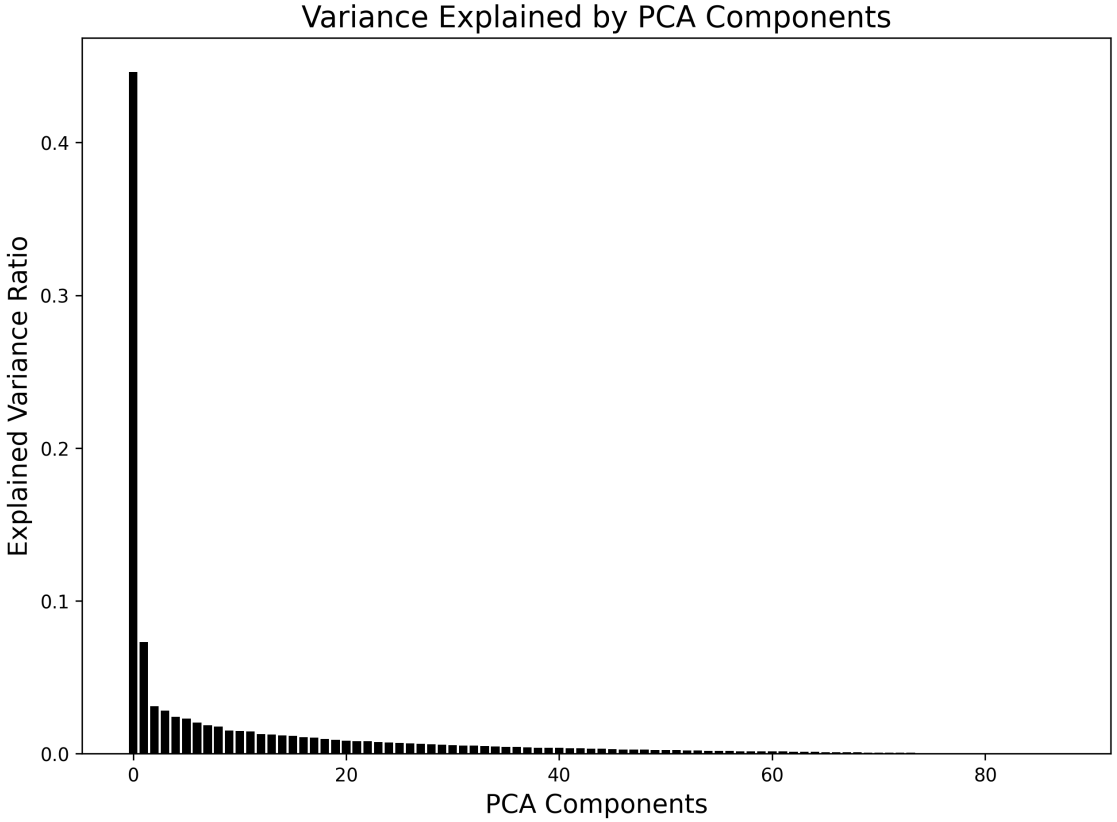


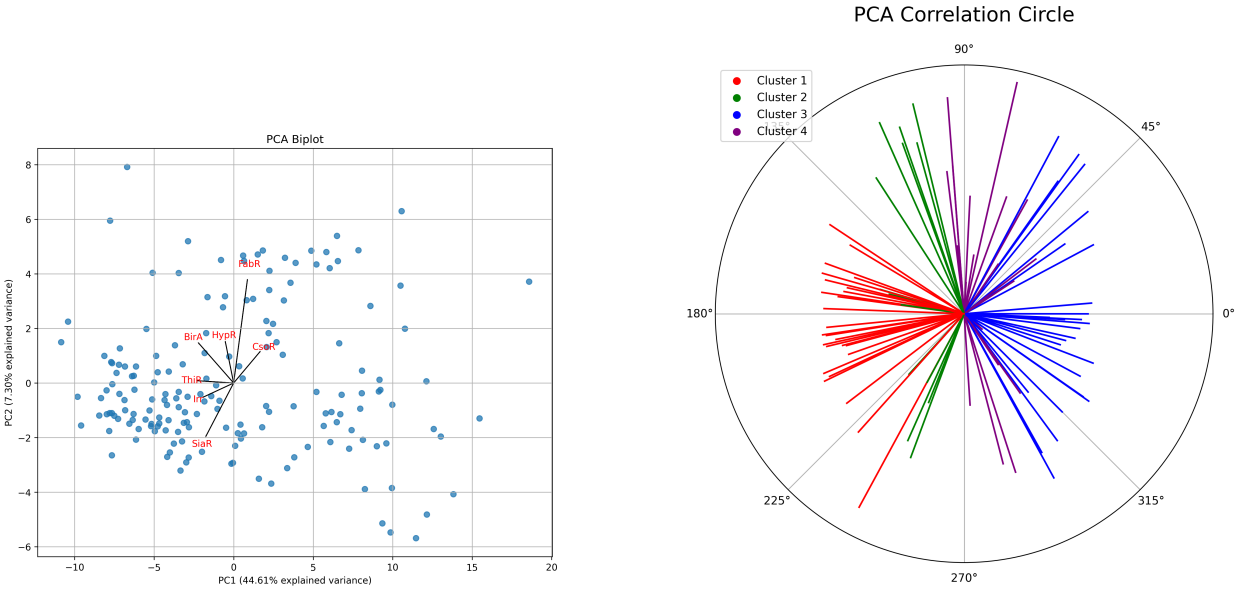
Figure 5.1: Explained Variance Ratio of PCA. Bar chart depicting the explained variance by each Principal Component (PC) in the PCA transformed dataset. The x-axis enumerates the PCs while the y-axis quantifies the variance explained. This visualization aids in understanding information encapsulated by each PC and in determining the optimal number of PCs for subsequent analysis.

The PCA analysis illustrates the high-dimensional structure in our dataset, as the first two principal components account for roughly more than 50% of the variance. This indicates that the complexity of the data cannot be simply (linearly) captured by a few dominant features. The third and subsequent components each explaining around 5% of the variance further underlines this point. Therefore, any substantial dimension reduction might cause a significant loss of critical information. This understanding will guide further steps of our analysis, since we are not keen to loss any information.

In Figure 5.2, we present a biplot based on the standard two Principal Component Analysis (PCA) components, alongside a PCA correlation circle. These visuals aid in comprehending

[16] The dataset was standarized for this purpose

the underlying structure and relationships within our dataset.



(a) (b)

Figure 5.2: Visualization of the Principal Component Analysis (PCA) using biplots for the biotic dataset. **(a)** PCA biplot presenting correlations between seven transcription factors and the first two principal components in the biotic dataset. The vectors symbolize the transcription factors, their orientation and magnitude suggesting their contribution to the principal components. The total variance explained by each principal component is indicated on the axes. **(b)** PCA Correlation Circle diagram visualizing the correlation of transcription factors (TFs) grouped into four different clusters with the first two Principal Components. Each vector's orientation, length, and color represent the TFs' contribution to the Principal Components and their cluster affiliation. For more information about the clusters see Appendix B

In the PCA correlation circle, distinct spatial patterns are evident when visualizing the clusters along the principal components. The red cluster, representing the first group, is primarily oriented along the negative side of the first principal component (PC1). Conversely, the blue third cluster is oriented towards the positive side of PC1. The green second cluster aligns vertically, indicating its characteristics are predominantly captured by the second principal component (PC2), and it leans a little towards the negative side of PC1. Meanwhile, the purple fourth cluster also aligns vertically along PC2 but leans more towards the positive side of PC1. This distribution underscores the distinct variance and relationships inherent in each cluster[17] when considering the dominant components of the dataset.

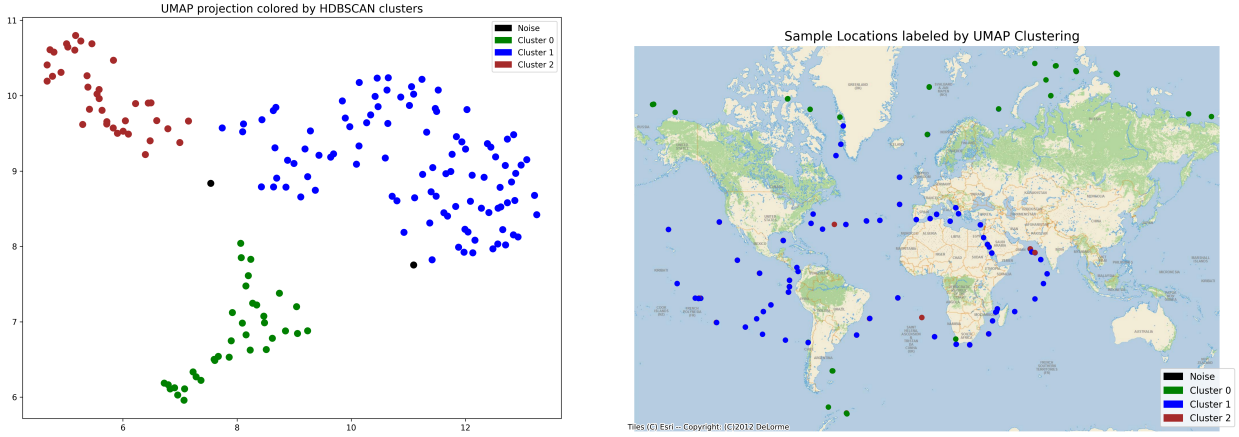
Capturing non-linear relations

While PCA is commonly used in the biology community for dimensionality reduction, it mainly captures linear variations in data. Given the complex, non-linear interactions often seen between biotic and abiotic variables in biological systems, a more advanced technique is required. UMAP [38], unlike PCA, captures non-linear structures and offers greater flexibility, ensuring both local and global data relationships are preserved. Therefore, while PCA is a

[17] For more information, see Annexed B

useful preliminary method, UMAP provides a more in-depth exploration, allowing us to uncover nuanced relationships and deeper insights that linear methods might miss.

Initially, we will perform an unsupervised 2D UMAP projection on the biotic data using a Euclidean metric. Subsequently, the HDBSCAN clustering algorithm [39] will be employed to provide insights into the data's structure.

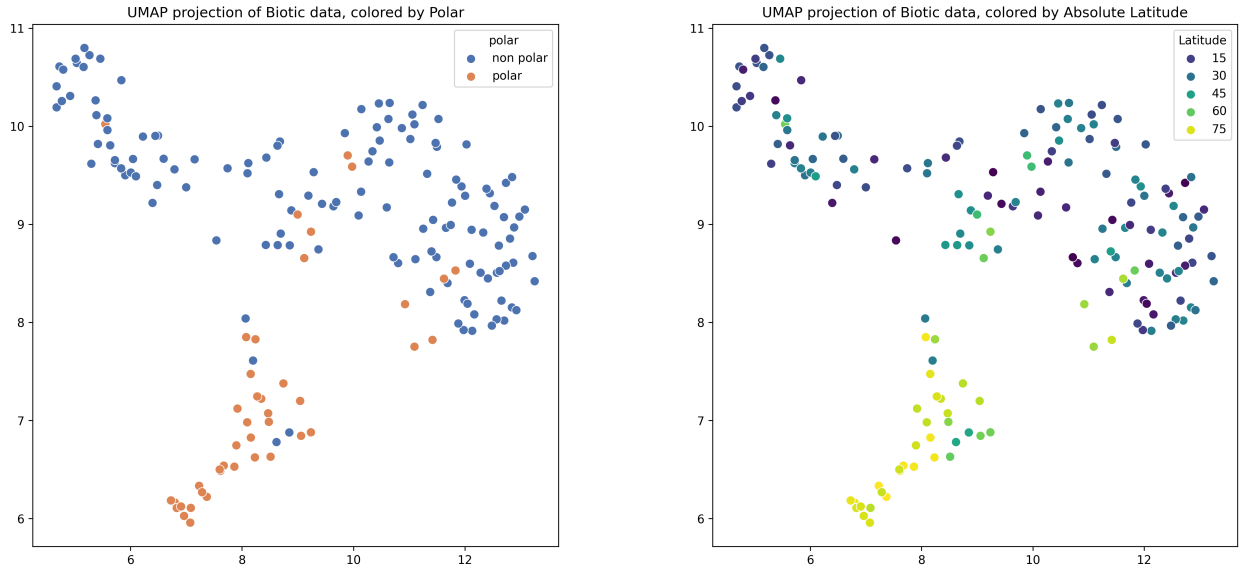


(a) (b)

Figure 5.3: Analysis of biotic data with HDBSCAN clustering on the 2D UMAP Projection. **(a)** 2D UMAP projection of the biotic data using the Euclidean metric, where each point represents a sample, colored by its HDBSCAN cluster assignment. Noise data points are represented in black. **(b)** Geographic distribution of the samples based on their longitude and latitude, colored by their respective HDBSCAN cluster.

From Figure 5.3, it becomes evident that biological factors distinctly differentiate between polar and non-polar samples (or absolute Latitude aswell), with only a few exceptions. What is particularly noteworthy is not merely the differentiation between polar and non-polar samples based on biological factors but the fact that this distinction is achieved through the lens of genomic regulation and its abundance. This approach sheds light on an aspect that has remained unexplored until now, reinforcing and complementing existing findings in the literature.

The insights drawn earlier are further illustrated in Figure 5.4. Our visual observations confirm our previous findings about the separation based on polarity and absolute latitude. Distinctly, aside from a few outliers, two clear clusters emerge, delineating polar and non-polar samples. Moreover, when considering latitude, a prominent cluster is evident, representing samples from regions with high absolute latitudes, specifically around 60 to 70 degrees. This underlines the significance of our data representation in capturing essential environmental gradients.



(a) (b)

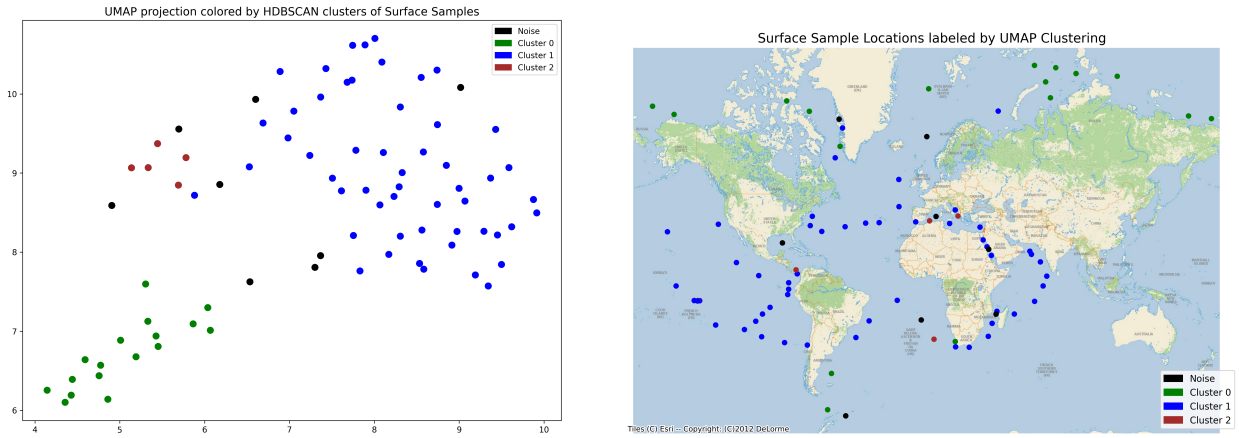
Figure 5.4: UMAP projections of biotic data using the Euclidean metric. (a) Samples color-coded based on their polarity. (b) Samples color-coded according to their abolsute latitude.

With the understanding that transcription factors serve as vital environmental sensors of the ocean [6], we have approached our analysis with a comprehensive perspective. Up to this point, the studies in this chapter have considered all samples in aggregate, which, while useful for broad observations, may not fully account for the intricate complexities of ocean systems. Recognizing this, we now aim to refine our approach by segmenting our analysis based on specific location and layer-based samples, thus enabling a more nuanced understanding of the environmental factors at play.

2D UMAP Projection of Surface Samples

Building on our earlier emphasis on the importance of the surface layer in marine research, we conduct an analysis of surface samples. This layer not only boasts the most comprehensive data but is also frequently identified in the literature as both an influencer and a respondent to biotic factors [13, 1]. Recognizing the mutual influences at play, an intricate exploration of these samples stands to illuminate the overarching marine ecosystem dynamics.

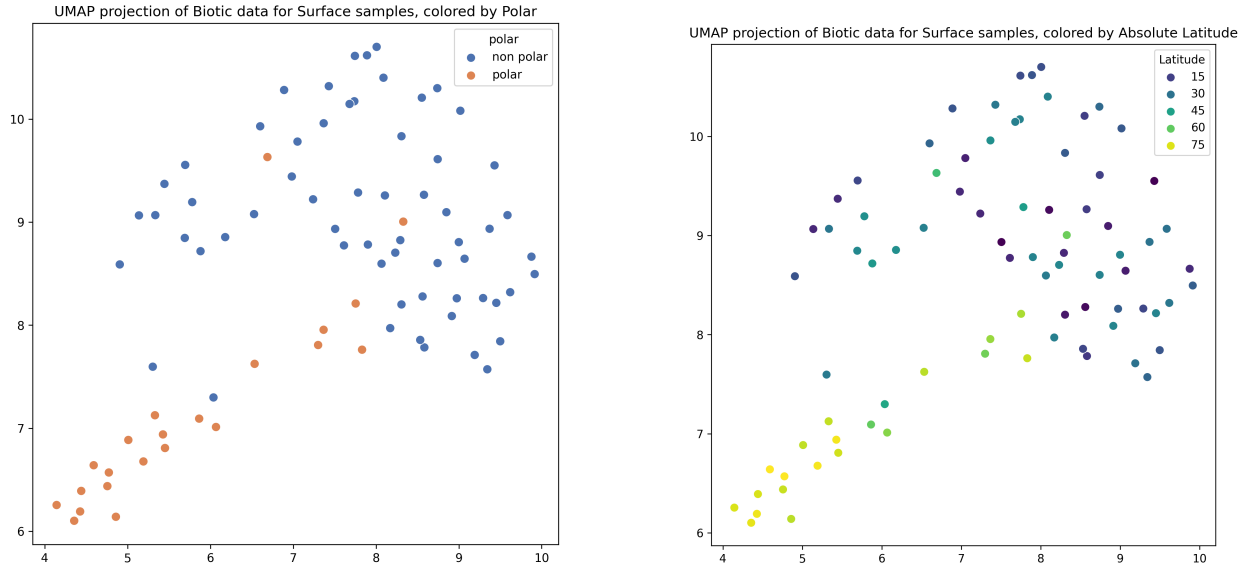
We will begin by replicating the analysis we conducted for all samples. This approach allows us to gain a comprehensive understanding of the specific dynamics within the surface samples.



(a) (b)

Figure 5.5: Analysis of biotic data with HDBSCAN clustering on the 2D UMAP Projection of Surface Samples. **(a)** 2D UMAP projection of the biotic data using the Euclidean metric, where each point represents a sample, colored by its HDBSCAN cluster assignment. Noise data points are represented in black. **(b)** Geographic distribution of the samples based on their longitude and latitude, colored by their respective HDBSCAN cluster.

From Figure 5.5, it is evident that even without the interference of the deeper mesopelagic samples, the abundance of transcription factors distinctly differentiates between polar and non-polar samples. To provide further clarity on this distinction, we will color the 2D projection based on polarity and latitude:



(a) (b)

Figure 5.6: UMAP projections of biotic data using the Euclidean metric of Surface samples. (a) Samples color-coded based on their polarity. (b) Samples color-coded according to their abolsute latitude.

Delving further into environmental factors and assessing their delineation solely based on the biological information within the dataset would be intriguing. For additional details, refer to Appendix D.

5.2. Comparative predictive modeling of environmental targets using biological features across different marine samples

Given our prior emphasis on the pivotal role of biotic-abiotic interactions in biological systems, it becomes crucial to employ advanced techniques for unravelling the subtleties within. Classification stands out as one of these powerful tools, adept at discerning intricate patterns and connections in vast datasets. This method, when applied judiciously, enables us to capture and understand the multifaceted associations between living organisms and their physical-chemical environments, enriching our comprehension of the dynamic interplay at hand.

The decision to utilize a classification algorithm in our study was driven by empirical evidence, not mere chance. Insights from the Andrew curves (Figure 3.35), along with the inherent patterns revealed by our dimensionality reduction (Figure 5.4), have underscored a pronounced distinction in environmental features. These analyses collectively pointed towards the potential of classification to further decipher these relationships.

In the realm of data science, a plethora of classification algorithms exist, each tailored for diverse applications. Broadly, these can be categorized into linear models, such as Support Vector Machines, and ensemble methods, which encompass techniques like Decision Trees, Random Forests, Neural Networks, and Gradient Boosting Machines. All these algorithms operate by 'learning' patterns from data, subsequently forming a model capable of predicting or classifying new data points. Notably, among the ensemble methods, Extreme Gradient Boosting (XGBoost) [40] has risen to prominence. Renowned for its efficiency and adaptability, XGBoost is an optimized distributed gradient boosting library that has garnered significant attention in recent times [41].

For our analysis, we chose XGBoost algorithm [42] because of its proven track record in delivering high performance in classification tasks. Given the complexity of our dataset (more unknown than complex), it is essential to use an algorithm that can robustly handle the intricacies and provide reliable results.

In the forthcoming sections, we delve into the construction and evaluation of predictive models under three distinct scenarios that separate the ocean environment, targeting key variables, for instance, Polar versus Non-Polar, Epipelagic (EPI) versus Mesopelagic (MES), etc. to understand different environmental conditions. The scenarios are as follows: 1) Utilizing the entire dataset encompassing all samples, 2) Focusing exclusively on surface (SRF) samples, and 3) Isolating Non-Polar samples. The selection of these target variables is consistent with our earlier correlation analysis and aligns with findings in the literature, where polarity and ocean layers are identified as critical elements of ocean stratification [11, 14, 1]. Additionally, we consider bioprovinces [13] (described in the third chapter), which were built based on taxonomic composition data.

All samples considered to predict target environmental variables.

We will begin with basic classification tasks to establish their merit and potential value for our study. Results are shown in Table 5.1:

Table 5.1: Classification scores for various target locations based on different metrics. The table compares the classification performance of predicting polar versus non-polar samples, surface (SRF) versus deep chlorophyll maximum (DCM) versus mesopelagic (MES) layers, epipelagic (EPI) versus mesopelagic (MES) layers, Ocean regions, and Provinces. For multilabel classifications, metrics were adjusted to be weighted, and the ROC AUC was calculated using a One-vs-Rest approach.

Target	Accuracy	Precision	Recall	F1	ROC AUC
Polar / Non Polar	0.87	0.78	0.68	0.73	0.92
SRF / DCM / MES	0.63	0.61	0.63	0.61	0.74
EPI / MES	0.72	0.74	0.71	0.72	0.81
Ocean region	0.46	0.44	0.46	0.43	0.75
Province	0.60	0.57	0.60	0.57	0.79

The Table 5.1 showcases the classification outcomes across diverse environmental and biological categories. Most prominently, the differentiation between 'Polar / Non-Polar' samples achieves an impressive accuracy of 87%, underscoring the stark contrasts between these groups. On the other hand, discerning among 'Ocean regions' poses more difficulty, reflected in the lowest accuracy rate of 46%. This might suggest overlapping characteristics or the presence of numerous confounding elements among the regions. However, given the challenge of nine target variables, this result is commendable. Also, the 60% in the prediction of provinces is important to be highlighted since this partition of the ocean was defined only using taxonomic information. The high ROC AUC values across most categories, including those with multilabel classifications, demonstrate the model's robust ability to rank samples effectively. This proficiency extends to the One-vs-Rest distinction, particularly evident in the ROC AUC values for multilabeled targets. Some categories display a variance in precision and recall, emphasizing the model's challenge in maximizing true positive predictions while minimizing false negatives. In sum, while certain environmental categories can be confidently predicted, others warrant deeper exploration and potential refinement in subsequent research.

Surface samples considered to predict target environmental variables.

As in earlier sections, our analysis will primarily concentrate on surface samples. The findings are presented in Table 5.2.

Table 5.2: Classification scores for surface samples across different environmental categories. The table highlights the performance of predicting 'Polar vs. Non-Polar', 'Ocean region', and 'Province'. As with multilabel classifications in previous tables, metrics were adjusted to be weighted, and the ROC AUC was calculated using a One-vs-Rest approach.

Target	Accuracy	Precision	Recall	F1	ROC AUC
Polar / Non Polar	0.90	0.86	0.80	0.80	0.95
Ocean region	0.39	0.35	0.39	0.36	0.68
Province	0.63	0.65	0.63	0.62	0.85

Table 5.2 emphasizes the classification metrics specifically for surface samples. When we juxtapose this with the comprehensive results from Table 5.1, several trends emerge.

For the 'Polar vs. Non-Polar' classification, the model's accuracy climbs to 90% for surface samples, a slight increase from the 87% achieved in the overarching dataset. This uptick in

accuracy underscores that surface samples more sharply delineate the disparities between the two categories. Notably, across all evaluation metrics, surface samples consistently outperform their broader counterparts.

'Ocean region' classification for surface samples presents a more complex picture. With an accuracy of 39%, it falls slightly below the 46% recorded for all samples. It hints at the inherent challenges associated with classifying oceanic regions based purely on surface samples. In any case, it is evident from this study that the oceans category does not tak into account the complex oceanographic processes along the sea that tend to mix oceans.

Lastly, the 'Province' classification results are comparably stable between both tables, albeit with a minor uptick in accuracy for surface samples. But as mentioned above, this category was defined only considering surface distribution of bacterial taxonomy and no functional trait.

In brief, surface samples tend to offer a clearer differentiation for 'Polar vs. Non-Polar' category.

Non Polar samples considered to predict target environmental variables.

Throughout this document, we have observed distinct biological differences between polar and non-polar samples. Recognizing this distinction, we will now delve into an analysis of the non-polar samples, aiming to glean insights specific to this classification.

Table 5.3: Classification scores for non-polar samples across various environmental categories. The table illustrates the performance of predicting different layering (SRF, DCM, MES), Ocean, and Province labels. As with other classifications, metrics were weighted, and the ROC AUC was computed using a One-vs-Rest method.

Target	Accuracy	Precision	Recall	F1	ROC AUC
SRF / DCM / MES	0.69	0.69	0.69	0.67	0.81
EPI / MES	0.64	0.68	0.61	0.64	0.76
Ocean	0.38	0.38	0.38	0.37	0.66
Province	0.63	0.65	0.63	0.62	0.85

The Table 5.3 displays classification metrics for non-polar samples across environmental categories like layering (SRF, DCM, MES), Ocean regions, and Provinces. Notably, the SRF, DCM, and MES layering metrics are consistent around 0.69, while the EPI/MES classification exhibits a slight variance with accuracy at 0.64. The Ocean region category presents a challenge, as all metrics cluster around 0.38. Meanwhile, the Province category metrics hover around 0.63, but it boasts a strong ROC AUC at 0.85, reflecting a high model discriminative capacity. Collectively, the results suggest a decline in classification effectiveness compared to when all samples were analyzed. This dip in performance may arise from the inherent ease associated with classifying polar ocean regions.

5.3. First approach to feature selection

Following the preliminary classification outcomes, there arises an imperative to further refine and enhance our model's accuracy and interpretability. A pivotal aspect of this refining process is 'Feature Selection'. We employed the 'Recursive Feature Elimination with Cross

Validation' (RFECV) technique from the *scikit-learn* library [37] using a *F1 scoring* and *XGBoosting estimator* to sieve out the most pertinent transcription factors. This will also help us determine whether certain transcription factors carry any irrelevant information.

Consequently, Table 5.4 displays the F1-scores and the optimal number of transcription factors required for classification of the target variables across various sample compositions, namely all samples, surface samples, and non-polar samples. The F1-score metric is not randomly selected and it is tied to the imbalance presented in the samples, regarding polarity or layer.

Table 5.4: Summary of the average F1-scores achieved following feature elimination with cross validation, alongside the number of selected features in parenthesis. Each row represents a distinct sample type, while columns specify the target variable.

Samples/Target	Polar	SRF/DCM/MES	EPI/MES	Ocean	Province
All	0.71 (24)	0.62 (84)	0.80 (18)	0.44 (80)	0.60 (12)
Surface	0.88 (6)	-	-	0.45 (19)	0.68 (56)
Non-Polar	-	0.65 (20)	0.90 (11)	0.41 (55)	0.50 (39)

Table 5.4 demonstrates a consistent pattern in the F1-score metric, mirroring our findings when evaluating all transcription factors. This table also sheds light on the transcription factors most relevant to the target variables. Yet, the reliability of this method is significantly influenced by train-test splits and other random elements intrinsic to the process. Although our results validate our preliminary assertions about the pivotal role of transcription factors in environmental contexts, pinpointing the 'preeminent' transcription factors remains a task at hand. Given the insights from Table 5.4 and the constraints of train-test splits and randomness, we have formulated a refined methodology that we will call *Robustness*. This strategy is meticulously tailored to reliably identify the transcription factors that have the most bearing on environmental target variables.

5.4. Robustness

Henceforth, our focus will be on the classification task for the target variables: polar, layer (SRF/DCM/MES), and layer2 (EPI/MES), owing to the suboptimal scores observed earlier for the other target variables.

In this section, we will introduce two critical parameters to assess the predictive strength of the transcription factors: 1) feature importance stability and 2) permutation importance. The former measures the consistency of feature importances in the face of the algorithm's inherent randomness, while the latter quantifies the permutation importance of the biotic features. Considering these parameters, we can rank the features based on specific metrics to be explained, enabling us to identify the optimal features for each classification task.

5.4.1. Feature importance stability

Earlier, we highlighted the intrinsic unpredictability of the algorithms driving our models, which inhibits our ability to definitively identify the most effective predictive features or establish a consistent ranking. To counter this challenge, we will implement a Monte Carlo

cross-validation method onto the feature importance[18] metric provided by the XGBoost algorithm [42]. This method involves multiple iterations, during which the feature importance of each transcription factor is documented for every run and each division generated by the stratified cross-validation. Let $F \in \mathcal{F}$ be a transcription factor in the set of all transcription factors and let us consider $m(F)$ the mean of the feature importances of F and $s(F)$ the standard deviations of the feature importance of F (calculated across all runs and each division). Let us now consider three metrics ϕ_1, ϕ_2 and ϕ_3 defined as:

$$\begin{aligned} \phi_1 : \mathcal{F} &\to \mathbb{R} \\ F &\mapsto \phi_1(F) = m(F) \end{aligned}$$

$$\begin{aligned} \phi_2 : \mathcal{F} &\to \mathbb{R} \\ F &\mapsto \phi_2(F) = \frac{m(F)}{s(F)} \end{aligned}$$

$$\begin{aligned} \phi_3 : \mathcal{F} \times [0,1] \times [0,1] &\to \mathbb{R} \\ (F, \alpha, \beta) &\mapsto \phi_3(F, \alpha, \beta) = \alpha \cdot \bar{m}(F) + \beta \cdot \bar{s}(F) \end{aligned}$$

were

$$\bar{m}(F) = \frac{m(F)}{\sum_{G \in \mathcal{F}} m(G)}, \quad \bar{s}(F) = \frac{1}{s(F)} \cdot \left(\sum_{G \in \mathcal{F}} \frac{1}{s(G)} \right)^{-1}$$

and $\alpha + \beta = 1$. We note that $\bar{m}(\cdot)$ and $\bar{s}(\cdot)$ are normalized mean and normalized inverse standar deviation respectively.

In our study, we have chosen $\alpha = 0.8$ and $\beta = 0.2$ because we prioritize feature importance over the weight of the error.

Subsequently, we obtain a set of scores for each metric, denoted as $\{\phi_i(F)\}_{F \in \mathcal{F}}$ for $i = 1, 2, 3$. We then apply a sorting function, $\Phi_i : \mathcal{F} \to \{1, ..., |\mathcal{F}|\}$, which ranks each transcription factor F based on the score $\phi_i(F)$. Specifically, $\Phi_i(F) > \Phi_i(G)$ if and only if $\phi_i(F) > \phi_i(G)$ for all $F, G \in \mathcal{F}$.

Finally, we consider the following ranking $r(F) = \sum_{i=1,2,3} \Phi_i(F)$ and obtain the set $\{r(F)\}_{F \in \mathcal{F}}$ which can be sorted.

5.4.2. Permutation importance

Analogously to the approach in the previous *Feature importance stability* section where we considered the feature importance metric, we will now focus on the permutation importance[19] metric, obtainable through the *Scikit-learn* library. Drawing from the analogy, we obtain the set $\{t(F)\}_{F \in \mathcal{F}}$ where each $t(F) = \sum_{i=1,2,3} \Psi_i(F)$ is a permutation importance-based ranking analogously to the $r(F)$.

[18] Feature importance assigns scores to features based on their relevance in predicting a target variable, highlighting the most influential ones for model decision-making.

[19] Permutation importance measures the decrease in a model's performance when a specific feature's values are randomly shuffled, indicating its significance in predictions.

5.4.3. Robust list of transcription factors

Having detailed the methodology for ranking transcription factors, we now transition to creating a robust list for classification. To achieve this, let us define $\kappa(\cdot)$ as follows:

$$\kappa(F) = r(F) + t(F),\ \forall F \in \mathcal{F}$$

One can readily sort the combined ranking set $\{\kappa(F)\}_{F\in\mathcal{F}}$ in order to obtain the most stable and *inelastic*[20] transcription factors. Given the established hierarchy trought the ranking system for the biological features, one can define thresholds of importance to glean deeper insights into these features. Notably, we selected the top 20% of features as the most crucial for classification and in Table 5.5 we show the results.

Table 5.5: Transcription factors constituting the top 80%, identified using the robustness modeling technique, for classifying target variables by sample categories.

Sample Category → Target	**Transcription Factors**
All samples → polar	BirA, CadR-PbrR, CcpA, DasR, ExuR_UxuR, FabR, FadR, FnrN_FixK, GguR, GulR, HrcA, LexA, LiuR, NagQ, PhnR, PurR, QorR, SdaR
All samples → SRF/DCM/MES	BirA, ExuR_UxuR, FabR, FadR, FixJ, GlcC, HrcA, LexA, LtbR, MetJ, NifA, NmlR, PaaR, PdxR, PsrA, Rex, TyrR, VanR
All samples → EPI/MES	CcpA, ExuR_UxuR, FadR, FruR, GlcC, HutC, Irr, IscR, LexA, LiuQ, LtbR, NmlR, NrdR, PaaR, PdxR, PhnF, PhnR, SdaR
Surface samples → polar	DasR, ExuR_UxuR, FabR, GlcC, GulR, HisR, HrcA, Irr, LexA, MetJ, NorR, PrpR, PurR, QorR, RutR, SiaR, TrpR, VanR
Non Polar samples → SRF/DCM/MES	BirA, CadR-PbrR, CblR, CzrA, FabR, FadR, FruR, GlcC, GntR, NifA, NikR, PaaR, PhnR, PrpR, QorR, RbsR, Rex, TrpR
Non Polar samples → EPI/MES	AmtR, CzrA, ExuR_UxuR, FadP, FadR, GguR, GlcC, Irr, PaaR, PhnR, ZntR

For a better visualization and a comprehensive comparison with the list of strongly correlated transcription factors with environmental data, we will display Table 5.5 as a figure. We decide to drop the three layer target variable and focus on the EPI/MES target only for clarity:

[20] Inelastic in the sense that they are not permutable.

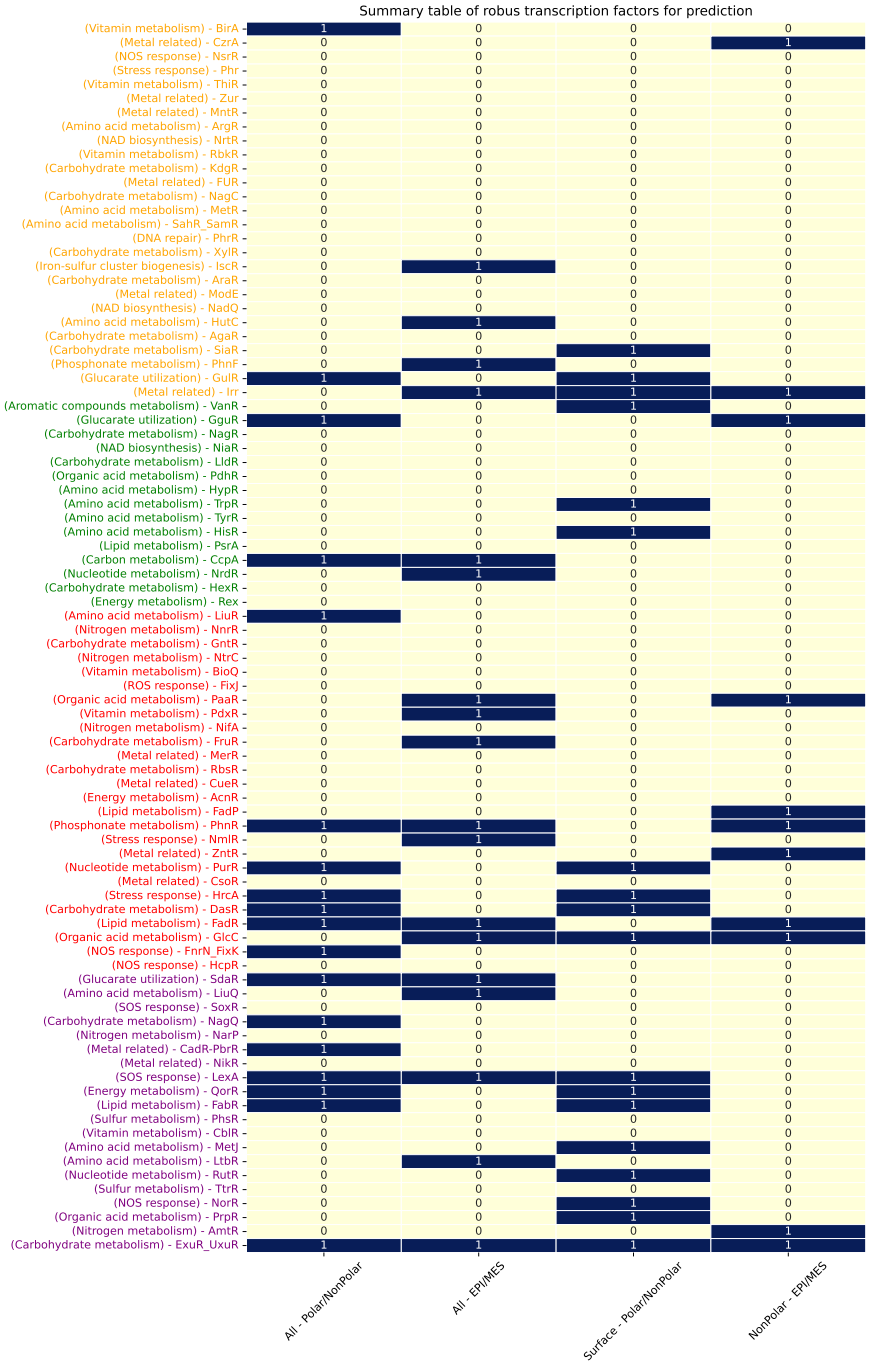


Figure 5.7: Transcription factors constituting the top 20%, identified using the robustness modeling technique, for classifying target variables by sample categories (All: all samples; Surface: surface samples; NonPolar: non polar samples). Transcription factors are colored by their cluster label (Figure B.1) and have their functionality annotated.

At this point, we identify a key set of transcription factors that predict environmental variables using biotic data. Upon comparison with Figure 4.28, it becomes evident that transcription factors with strong predictive power (Figure 5.7) are not always those with the highest correlation to environmental variables. This finding is significant as it suggests that certain transcription factors may have low correlation coefficients but are nonetheless effective predictors due to their roles in complex transcriptional regulatory networks, which may not

be fully appreciated by correlation analysis alone. Therefore, we conclude that correlation metrics are insufficient for a comprehensive analysis of biotic-abiotic interactions. Predictive metrics are necessary to provide deeper insights into the underlying biological regulatory mechanisms.

This observation extends to the outcomes of cluster analysis; it becomes evident that membership within a specific TF cluster does not inherently imply superior predictive capability. Furthermore, when considering the functional annotations associated with each TF, a direct correlation between predictive efficacy and designated functionality is not substantiated. Therefore, it cannot be concluded that a particular function is indicative of enhanced predictive performance.

It is pertinent to note that while transcription factors are evaluated and assigned a rank based on their predictive performance, the final compilation is presented as an unranked list. This approach is adopted due to the marginal variations in predictive efficacy among top-tier transcription factors as determined by statistical analysis. Consequently, delineating a hierarchy among these factors could misrepresent their relative predictive capabilities. Thus, the provision of a collective list of optimal predictors is deemed more appropriate for representing the findings. With this context established, we will proceed to engage with existing literature to elucidate the interactions between these transcription factors and environmental factors. We acknowledge the presence of certain transcription factors for which the current literature does not offer extensive insights. The implications of this are twofold: firstly, it presents an opportunity for future research to bridge this knowledge gap; secondly, it underscores the importance of experimental validation to substantiate the predictive relevance of these factors in environmental contexts.

We will compare our results with findings in the literature. While many of the 88 transcription factors are not well documented in terms of their relationship with physicochemical variables, there are some that are.

FabR

The transcription factor FabR emerges as a pivotal feature in predicting polar versus non-polar samples. In the RegPrecise database, this factor is identified as a repressor of the fatty acids biosynthesis genes [43]. Additionally, several studies in the literature have established a relationship between temperature and fatty acid synthesis. Notably, Seung-Hwi Lee et al. [44] conducted an investigation on the effects of temperature shifts on *Paracyclopina nana*. They observed that at temperatures below the standard 15°C, there was an increase in mRNA expression associated with lipogenesis, a larger area of lipid droplets (intracellular compartments storing lipid reserves in organisms), and alterations in fatty acid composition. Conversely, all these markers decreased significantly at temperatures exceeding the standard 25°C. In a separate study conducted by Olson & Ingram [45], the fatty acid composition of *Agmenellum quadruplicatum* was shown to be significantly influenced by temperature. Specifically, as the growth temperature increased from 20°C to 43°C, there was a notable rise in saturated fatty acids in log-phase cells, from 38.4% at 20°C to 63.6% at 43°C. In another study, researchers explored how *Thermobifida fusca* produces fatty acids under varying conditions. The article from Wilkelman & Nikolau [46] details their findings when the organism is grown on different carbon sources, such as glucose, cellobiose, cellulose, and avicel. Additionally, they considered two distinct growth temperatures: an optimal temperature of 50°C and a suboptimal one at 37°C. The conclusions drawn indicate that while both

carbon sources and growth temperatures influence the fatty acid profiles of *T. fusca*[21], it is the temperature that has a more pronounced impact on these traits.

In conclusion, the empirical evidence firmly establishes that shifts in temperature play a critical role in fatty acid synthesis. Consequently, the regulation of genes associated with this is essential. Figure 5.8 illustrates the relationship between FabR abundance[22] and temperature.

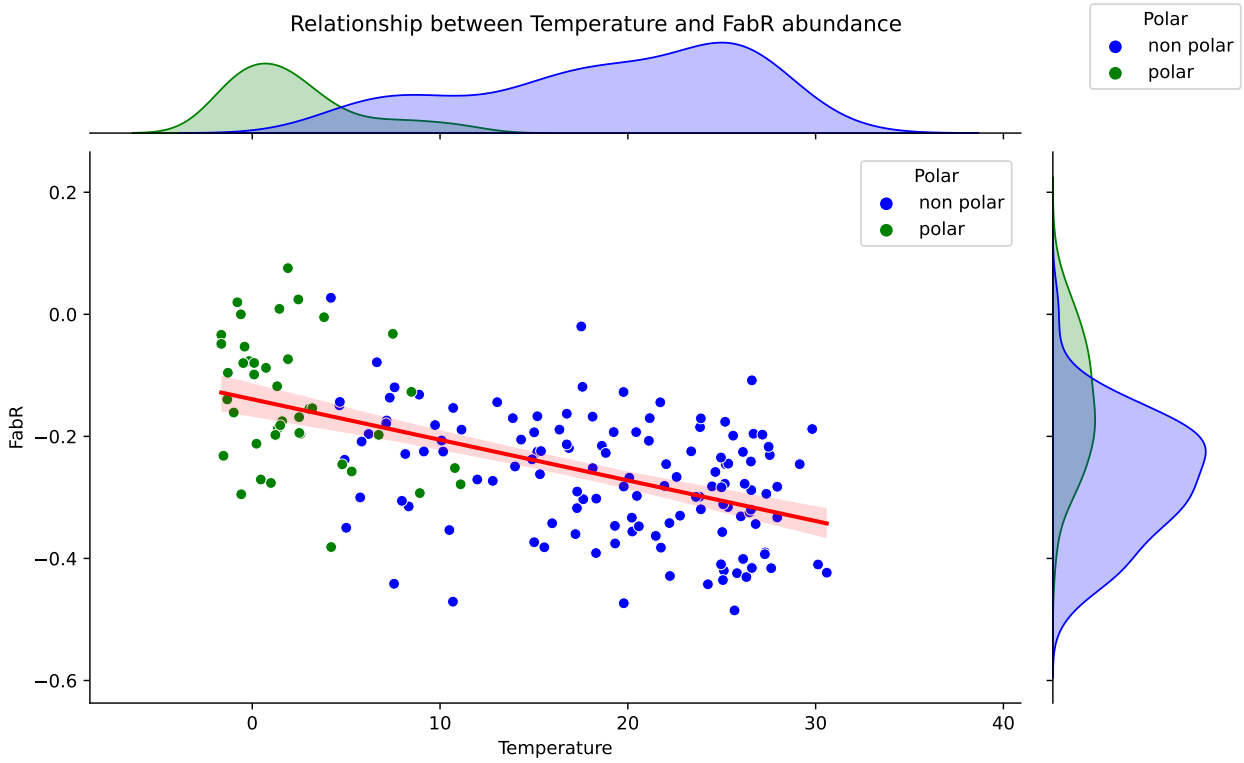


Figure 5.8: Scatter plot illustrating the correlation between FabR abundance and environmental temperature, with overlaid distributions indicating the prevalence within polar and nonpolar regions. The plot underscores potential patterns of FabR abundance in response to temperature variances.

Inspection of Figure 5.8 reveals a discernible trend: an inverse relationship between temperature and FabR expression levels, where an increase in FabR abundance corresponds to a decrease in temperature. Utilizing this series as a sole predictor in our model, we achieved a prediction accuracy exceeding 70% for classifying samples into polar and non-polar categories.

These insights affirm the FabR transcription factor as a robust predictor for distinguishing between polar and non-polar samples. The validity of FabR as a predictive marker holds not only in the context of the entire sample set but is particularly strong when analyzing surface samples.

CcpA

The catabolite control protein **CcpA** is a pleiotropic transcription factor that mediates the global transcriptional response to rapidly catabolizable carbohydrates such as glucose

[21] Thermobifida fusca is a thermophilic bacterium known for its ability to degrade plant biomass, particularly cellulose, making it relevant for biofuel research.

[22] We recall that by TF abundance we mean the abundance of binding sites associated to a TF

in Gram-positive bacteria from the Firmicutes phylum. CcpA belong to the LacI protein family and controls its target genes either positively or negtively, depending on the position of CcpA-binding sites (or Catabolite Responsive Elements, CRE) in the promoter regions of its target genes [43]. Since CcpA controls many carbohydrate metabolism, it suggets that the role of CcpA as a link between carbon and nitrogen metabolic pathways. The Redfield and Sterner C:N ratios are different in oceans, including polar ones [47]. So, contrary to FabR, this predictor is based on the nutrient availability of the ocean to capture the relation between polarity and the biology.

On the side of Layer prediction, we have the following:

BirA

BirA is a bifunctional protein that acts as a repressor of biotin biosynthesis genes and as an enzyme, biotin-protein ligase, that charges the biotin-dependent enzymes with the co-factor. The BirA repressor binds to the promoter regions of the regulated genes in the presence of the biotinyl-5'-adenylate [43].

There is a study by Erin M. and Andrew E. [48] which studies the influence of **vitamin B** auxotrophy on nitrogen metabolism in eukaryotic phytoplankton. They claim that the rate, magnitude, and species composition of marine primary production has a profound influence of global carbon cycling and therefore climate. As a result, factors controlling the growth of marine primary producers are of considerable interest. While nitrogen and iron availability are often considered the primary bottom-up controls on short-term marine primary productivity, the importance of organic growth factors received considerable early attentionand is the subject of renewed interest. Furthermore, recent developments in analytical techniques, application of trace metal clean bioassay experiments and culture-based surveys of vitamin requirements have identified B_12 (cobalamin) and B_1 (thiamine) as highly important growth factors for eukaryotic phytoplankton and suggest that these micronutrients have the potential to broadly influence marine productivity and species composition. Since nutrient availability is different in oceanic layers, we confirm BirA as a robust predictor of Layers in All samples and Surface samples scenarios.

Although we could continue this analysis for each transcription factor, the available literature for the others is not sufficiently clear to allow us to relate these factors to environmental changes based on existing research. Nevertheless, the transcription factors mentioned in the robust list 5.5 are, from our perspective, the best predictors, setting aside the inherent random noise of the algorithms. In this context, this list contributes to future scientific studies aiming to relate, from an experimental and biological viewpoint, the details of these existing relationships.

We acknowledge that the method we used to construct this robust list is just one approach among many. In this regard, SHAP values emerge as an intriguing option to add to the algorithm, possibly as a complement or even replacing a parameter currently in use, such as feature importance or permutation importance. From our perspective, incorporating only these two metrics addresses a need for explainability of the robust list. Furthermore, we observed that the transcription factors linked to the highest SHAP values for each scenario and target variable do not significantly differ from the robust list obtained in this study, which is an additional reason why we decided not to incorporate SHAP values as an extra parameter for the robust model.

Chapter 6

Conclusions

The profound expanse of the ocean biome, accounting for approximately 70% of Earth's surface, remains a subject of intense scientific curiosity, given its intricate ecosystems and the plethora of life forms it harbors. Central to this biome is the ocean microbiome, particularly planktonic bacteria, which play pivotal roles in biogeochemical cycles and significantly influence global climate patterns. This research embarked on the journey to delve deeper into the genomic regulation of these bacteria, specifically exploring the intricate relationship between the binding motifs of transcription factors in bacterial metagenomes and the environmental factors of the ocean biome..

Revisiting our primary question, "How are the abundances of binding motifs of transcription factors in oceanic bacterial metagenomes related to environmental factors in the ocean biome?", this study, rooted in a robust interdisciplinary approach, has made several groundbreaking contributions. It has provided a comprehensive exploration of the binding motifs abundance associated to 88 transcription factors across the ocean, revealing insightful patterns in the clr-normalized distribution of transcription factors. Specifically 2/3 of the distributions were normal and the majority left were beta distributions. The median of these distributions were diverse across the ocean, but the standar deviation was small. More so, minimal variations were observed between the abundances of binding motifs across the samples, meaning that abundances were bounded at each TF. This suggests that local variations, rather than broader geographical differences, account for the true variability in abundance. The consistency of these patterns across diverse oceanic regions underscores the importance of the regulatory mechanisms of genes on a global scale.

Furthermore, we were able to identify clusters of transcripcion factors revealing biological relations using a correlation metric. Specifically, we proceeded with hierarchical clustering using a distance metric defined by $1 - |\text{spearman}\rho|$. Two distinct clusters with a correlation score $|\rho| > 0.5$ emerged: one encapsulating positively correlated transcription factors and the other bringing together those that are negatively correlated. These insights give a comprehension of the intricate interactions within marine ecosystems. More so, we also identified key transcripcion factors correlated with the environment, emphasizing how external environmental factors can 'influence' bacterial transcriptional regulation. Although it does not mean causality, given the context we could say that environment is affecting the biological dynamic present in the regulation.

Building on the understanding that correlations do not imply causality, our focus shifted to constructing predictive models for environmental variables using the binding motif abundance matrix. Initially, the performance metrics for each target variable exceeded expectations, par-

ticularly notable in the classification of polar versus non-polar regions, achieving an F1-score close to 0.8. Additionally, the prediction of bioprovinces, defined based on taxonomy, yielded an F1-score around 0.6. This led to the development of a robust method for identifying key contributors to prediction, termed 'robust predictors'. These predictors are adept at capturing the biological relationship with the environment. This robust list of transcription factors, tailored for predictions in various scenarios - encompassing all samples, surface samples, or exclusively non-polar ones, and for each target variable such as polar/non-polar, epipelagic/mesopelagic - provides a tool through which we can narrow down variables to understand the dynamics between biology and the environment.

Our findings were juxtaposed with existing literature studies that link transcription factors, or the processes they regulate (functionality), with variations in environmental variables. Of the 88 transcription factors analyzed, only a few are sufficiently documented to provide a comprehensive overview. However, we highlight FabR, a 'repressor of the fatty acids biosynthesis genes'. It has been documented that temperature changes are linked to alterations in fatty acid synthesis, which corroborates the significance of this TF in predicting polarity. This alignment with established research not only validates our results but also underscores the potential of our approach in elucidating the intricate dynamics of transcription factors in relation to environmental changes. Another transcription factor, also robust in predicting polarity, is CcpA. Unlike FabR, CcpA controls many aspects of carbohydrate metabolism, further suggesting its role as a link between carbon and nitrogen metabolic pathways. Considering that the carbon:nitrogen (C:N) Redfield ratio varies across different oceans [47], including polar and non-polar regions, we can reaffirm the importance of this transcription factor in the prediction of polarity. This finding underscores how CcpA might play a critical role in bacterial community adaptation to marine environments, regulating nutrient assimilation and utilization in accordance with the specific environmental conditions of these regions.

The ramifications of our findings extend well beyond the academic sphere. With a more nuanced understanding of genomic regulation in marine bacteria, policymakers and conservationists are better equipped to devise strategies pertinent to marine conservation, climate change adaptation, and biodiversity management. Moreover, the potential insights our studies could provide on functions like the oceanic carbon pump emphasize the environmental and climatic significance of this research.

In sum, this research not only broadens our knowledge horizon of the oceanic ecosystem's adaptability and resilience in the face of environmental fluxes but also underscores the pressing need and importance of conserving and understanding this vast biome. The oceans remain a rich tapestry of mysteries, and as this thesis demonstrates, every thread unraveled leads to deeper insights and revelations that have ramifications for the planet at large. Through the lens of genomic regulation and the myriad interconnections of the marine biome, we take a step closer to understanding the depths and expanse of our blue planet.

Bibliography

[1] S. Sunagawa, S. G. Acinas, P. Bork, C. Bowler, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels, L. Karp-Boss, E. Karsenti, M. Lescot, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes, C. Sardet, M. Sieracki, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, D. Eveillard, G. Gorsky, L. Guidi, D. Iudicone, E. Karsenti, F. Lombard, H. Ogata, S. Pesant, M. B. Sullivan, P. Wincker, C. de Vargas, and T. O. Coordinators, "Tara oceans: towards global ocean ecosystems biology," *Nature Reviews Microbiology*, vol. 18, no. 8, pp. 428–445, 2020. Note: The author list seems to have duplicates. Please ensure accuracy.

[2] M. FREDERIKSEN, M. EDWARDS, A. J. RICHARDSON, N. C. HALLIDAY, and S. WANLESS, "From plankton to top predators: bottom-up control of a marine food web across four trophic levels," *Journal of Animal Ecology*, vol. 75, no. 6, pp. 1259–1268, 2006.

[3] L. Polimene, S. Sailley, D. Clark, A. Mitra, and J. I. Allen, "Biological or microbial carbon pump? The role of phytoplankton stoichiometry in ocean carbon sequestration," *Journal of Plankton Research*, vol. 39, pp. 180–186, 12 2016.

[4] E. J. Biers, S. Sun, and E. C. Howard, "Prokaryotic genomes and diversity in surface ocean waters: Interrogating the global ocean sampling metagenome," *Applied and Environmental Microbiology*, vol. 75, no. 7, pp. 2221–2229, 2009.

[5] E. Davidson, *Genomic Regulatory Systems: Development and Evolution*. Academic Press, 2001.

[6] G. Balázsi and Z. N. Oltvai, "Sensing your surroundings: How transcription-regulatory networks of the cell discern environmental signals," *Science's STKE*, vol. 2005, no. 282, pp. pe20–pe20, 2005.

[7] F. Aitken and J.-N. Foulc, *From deep sea to laboratory. 1: The first explorations of the deep sea by H.M.S. Challenger (1872-1876)*. London: Wiley-ISTE London, 2019. 1 online resource.

[8] J. C. Venter and D. E. Duncan, *The Voyage of Sorcerer II*. Cambridge, MA and London, England: Harvard University Press, 2023.

[9] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein, "Genomic analysis of regulatory network dynamics reveals large topological changes," *Nature*, vol. 431, pp. 308–312, Sep 2004.

[10] L. Guidi, S. Chaffron, L. Bittner, D. Eveillard, A. Larhlimi, S. Roux, Y. Darzi, S. Audic, L. Berline, J. R. Brum, L. P. Coelho, J. C. I. Espinoza, S. Malviya, S. Sunagawa, C. Dimier, S. Kandels-Lewis, M. Picheral, J. Poulain, S. Searson, L. Stemmann, F. Not,

P. Hingamp, S. Speich, M. Follows, L. Karp-Boss, E. Boss, H. Ogata, S. Pesant, J. Weissenbach, P. Wincker, S. G. Acinas, P. Bork, C. de Vargas, D. Iudicone, M. B. Sullivan, J. Raes, E. Karsenti, C. Bowler, G. Gorsky, and T. O. C. Coordinators, "Plankton networks driving carbon export in the oligotrophic ocean," *Nature*, vol. 532, no. 7600, pp. 465–470, 2016.

[11] F. M. Ibarbalz, N. Henry, M. C. Brandão, S. Martini, G. Busseni, H. Byrne, L. P. Coelho, H. Endo, J. M. Gasol, A. C. Gregory, F. Mahé, J. Rigonato, M. Royo-Llonch, G. Salazar, I. Sanz-Sáez, E. Scalco, D. Soviadan, A. A. Zayed, A. Zingone, K. Labadie, J. Ferland, C. Marec, S. Kandels, M. Picheral, C. Dimier, J. Poulain, S. Pisarev, M. Carmichael, S. Pesant, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, M. Babin, E. Boss, D. Iudicone, O. Jaillon, S. G. Acinas, H. Ogata, E. Pelletier, L. Stemmann, M. B. Sullivan, S. Sunagawa, L. Bopp, C. de Vargas, L. Karp-Boss, P. Wincker, F. Lombard, C. Bowler, and L. Zinger, "Global trends in marine plankton diversity across kingdoms of life," *Cell*, vol. 179, no. 5, pp. 1084–1097.e21, 2019.

[12] S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. d'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. coordinators, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, P. Bork, E. Boss, C. Bowler, M. Follows, L. Karp-Boss, U. Krzic, E. G. Reynaud, C. Sardet, M. Sieracki, and D. Velayoudon, "Structure and function of the global ocean microbiome," *Science*, vol. 348, no. 6237, p. 1261359, 2015.

[13] P. Frémont, M. Gehlen, and O. Jaillon, "Plankton biogeography in the 21st century and impacts of climate change: advances through genomics," *Comptes Rendus. Biologies*, vol. 346, pp. 13–24, 2023.

[14] G. Salazar, L. Paoli, A. Alberti, J. Huerta-Cepas, H.-J. Ruscheweyh, M. Cuenca, C. M. Field, L. P. Coelho, C. Cruaud, S. Engelen, A. C. Gregory, K. Labadie, C. Marec, E. Pelletier, M. Royo-Llonch, S. Roux, P. Sánchez, H. Uehara, A. A. Zayed, G. Zeller, M. Carmichael, C. Dimier, J. Ferland, S. Kandels, M. Picheral, S. Pisarev, J. Poulain, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, S. G. Acinas, M. Babin, P. Bork, C. Bowler, C. de Vargas, L. Guidi, P. Hingamp, D. Iudicone, L. Karp-Boss, E. Karsenti, H. Ogata, S. Pesant, S. Speich, M. B. Sullivan, P. Wincker, and S. Sunagawa, "Gene expression changes and community turnover differentially shape the global ocean metatranscriptome," *Cell*, vol. 179, no. 5, pp. 1068–1083.e21, 2019.

[15] D. Sigman and G. Haug, "6.18 - the biological pump in the past," in *Treatise on Geo-*

chemistry (H. D. Holland and K. K. Turekian, eds.), pp. 491–528, Oxford: Pergamon, 2003.

[16] J. R. Brum, J. C. Ignacio-Espinoza, S. Roux, G. Doulcier, S. G. Acinas, A. Alberti, S. Chaffron, C. Cruaud, C. de Vargas, J. M. Gasol, G. Gorsky, A. C. Gregory, L. Guidi, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, B. T. Poulos, S. M. Schwenck, S. Speich, C. Dimier, S. Kandels-Lewis, M. Picheral, S. Searson, T. O. Coordinators, P. Bork, C. Bowler, S. Sunagawa, P. Wincker, E. Karsenti, and M. B. Sullivan, "Patterns and ecological drivers of ocean viral communities," *Science*, vol. 348, no. 6237, p. 1261498, 2015.

[17] C. A. Suttle, "Marine viruses — major players in the global ecosystem," *Nature Reviews Microbiology*, vol. 5, no. 10, pp. 801–812, 2007.

[18] S. Minobe, A. Kuwano-Yoshida, N. Komori, S.-P. Xie, and R. J. Small, "Influence of the Gulf Stream on the troposphere," *Nature*, vol. 452, pp. 206–209, 03 2008.

[19] J. B. Palter, "The role of the Gulf Stream in european climate," *Annual Review of Marine Science*, vol. 7, pp. 113–137, 01 2015.

[20] V. Smetacek and S. Nicol, "Polar ocean ecosystems in a changing world," *Nature*, vol. 437, pp. 362–368, 09 2005.

[21] S. Levitus, M. E. Conkright, J. L. Reid, R. G. Najjar, and A. Mantyla, "Distribution of nitrate, phosphate and silicate in the world oceans," *Progress in Oceanography*, vol. 31, no. 3, pp. 245–273, 1993.

[22] Z. S. Kolber, R. T. Barber, K. H. Coale, S. E. Fitzwateri, R. M. Greene, K. S. Johnson, S. Lindley, and P. G. Falkowski, "Iron limitation of phytoplankton photosynthesis in the equatorial pacific ocean," *Nature*, vol. 371, no. 6493, pp. 145–149, 1994.

[23] H. Schoffman, H. Lis, Y. Shaked, and N. Keren, "Iron–nutrient interactions within phytoplankton," *Frontiers in Plant Science*, vol. 7, 2016.

[24] J. Wu, W. Sunda, E. A. Boyle, and D. M. Karl, "Phosphate depletion in the western north atlantic ocean," *Science*, vol. 289, no. 5480, pp. 759–762, 2000.

[25] A. Redfield, *On the Proportions of Organic Derivatives in Sea Water and Their Relation to the Composition of Plankton*, pp. 176–192. University Press of Liverpool, 1934.

[26] M. Winder and U. Sommer, "Phytoplankton response to a changing climate," *Hydrobiologia*, vol. 698, no. 1, pp. 5–16, 2012.

[27] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome datasets are compositional: And this is not optional," *Frontiers in Microbiology*, vol. 8, pp. 1–6, 2017.

[28] M. Waskom, "Seaborn: statistical data visualization." https://github.com/mwaskom/seaborn, 2021.

[29] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python." https://github.com/statsmodels/statsmodels, 2021.

[30] W. McKinney, *pandas: A Foundational Python Library for Data Analysis and Statistics*, 2010.

[31] S. Emerson and J. Hedges, *Chemical Oceanography and the Marine Carbon Cycle*. Cambridge University Press, 2008.

[32] A. Hagberg, P. Swart, and D. Schult, "Exploring network structure, dynamics, and function using networkx," 2008.

[33] M. Tumminello, T. Aste, T. Di Matteo, and R. Mantegna, "A tool for filtering information in complex systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 10421–6, 08 2005.

[34] G. Li, L. Cheng, J. Zhu, K. E. Trenberth, M. E. Mann, and J. P. Abraham, "Increasing ocean stratification over the past half-century," *Nature Climate Change*, vol. 10, no. 12, pp. 1116–1123, 2020.

[35] H. Kaneko, R. Blanc-Mathieu, H. Endo, S. Chaffron, T. O. Delmont, M. Gaia, N. Henry, R. Hernández-Velázquez, C. H. Nguyen, H. Mamitsuka, P. Forterre, O. Jaillon, C. de Vargas, M. B. Sullivan, C. A. Suttle, L. Guidi, and H. Ogata, "Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean," *iScience*, vol. 24, no. 1, p. 102002, 2021.

[36] C. Robinson, D. K. Steinberg, T. R. Anderson, J. Arístegui, C. A. Carlson, J. R. Frost, J.-F. Ghiglione, S. Hernández-León, G. A. Jackson, R. Koppelmann, B. Quéguiner, O. Ragueneau, F. Rassoulzadegan, B. H. Robison, C. Tamburini, T. Tanaka, K. F. Wishner, and J. Zhang, "Mesopelagic zone ecology and biogeochemistry – a synthesis," *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 57, no. 16, pp. 1504–1518, 2010. Ecological and Biogeochemical Interactions in the Dark Ocean.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[38] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[39] L. McInnes, J. Healy, and S. Astels, "Accelerated hierarchical density clustering," *arXiv preprint arXiv:1705.07321*, 2017.

[40] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[41] C. Bentéjac, A. Csörgo, and G. Martínez-Muñoz, "A comparative analysis of xgboost," *CoRR*, vol. abs/1911.01914, 2019.

[42] T. Chen *et al.*, "Xgboost: A scalable tree boosting system - python implementation." https://github.com/dmlc/xgboost, 2023. Version 1.7.6.

[43] R. Team, "Regprecise database," 2023. Accessed on: 2023-08-10.

[44] S.-H. Lee, M.-C. Lee, J. Puthumana, J. C. Park, S. Kang, J. Han, K.-H. Shin, H. G. Park, A.-S. Om, and J.-S. Lee, "Effects of temperature on growth and fatty acid synthesis in the cyclopoid copepod paracyclopina nana," *Fisheries Science*, vol. 83, no. 5, pp. 725–734, 2017.

[45] G. Olson and L. Ingram, "Effects of temperature and nutritional changes on the fatty acids of agmenellum quadruplicatum," *Journal of Bacteriology*, vol. 124, pp. 373–379, October 1975.

[46] D. C. Winkelman and B. J. Nikolau, "The effects of carbon source and growth temperature on the fatty acid profiles of thermobifida fusca," *Frontiers in Molecular Biosciences*, vol. 9, 2022.

[47] H. Frigstad, T. Andersen, R. G. Bellerby, A. Silyakova, and D. O. Hessen, "Variation in the seston c:n ratio of the arctic ocean and pan-arctic shelves," *Journal of Marine Systems*, vol. 129, pp. 214–223, 2014.

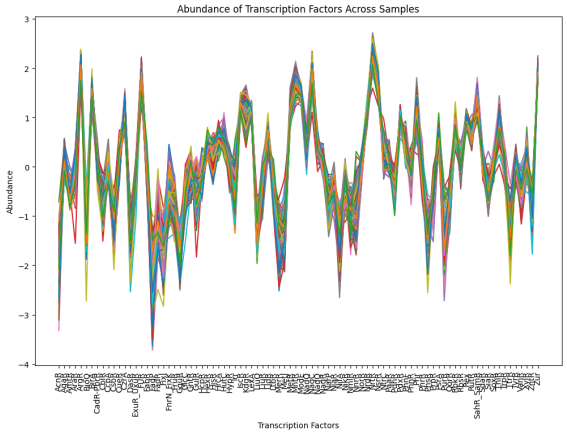
[48] E. Bertrand and A. Allen, "Influence of vitamin b auxotrophy on nitrogen metabolism in eukaryotic phytoplankton," *Frontiers in Microbiology*, vol. 3, 2012.

[49] R. Lange, H. Staaland, and A. Mostad, "The effect of salinity and temperature on solubility of oxygen and respiratory rate in oxygen-dependent marine invertebrates," *Journal of Experimental Marine Biology and Ecology*, vol. 9, no. 3, pp. 217–229, 1972.

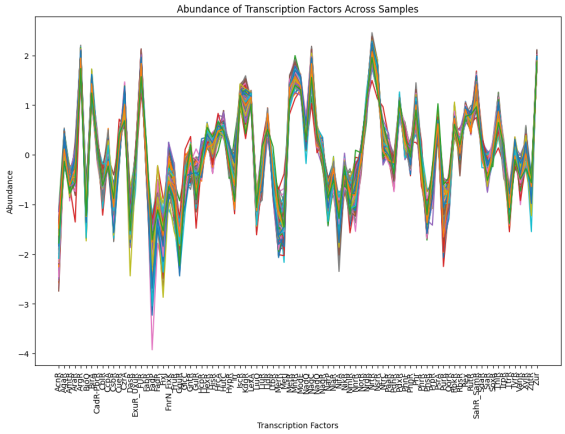
Annexes

Annex A. Temporal sequencing of Transcription Factor abundance comparison

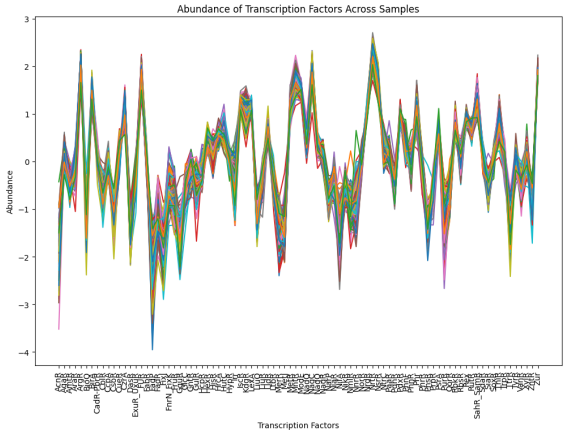
From Figure 3.32, one might question whether this behavior is exclusive to the class matrix M0. However, it is not. The results for the class matrices M1 and M2, for the same PRR are presented in Figure A.1



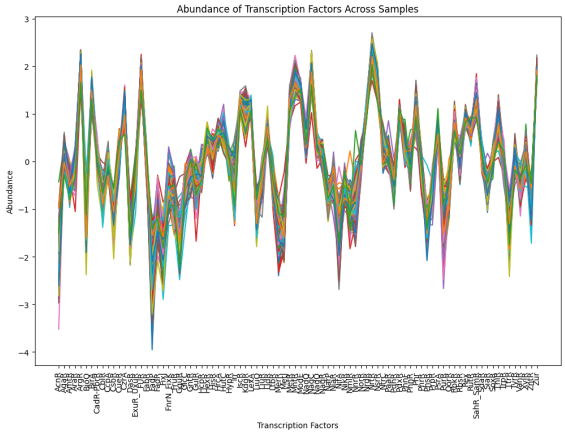
(a) M1, 150-10 PRR



(b) M1, 300-30 PRR



(c) M2, 150-10 PRR

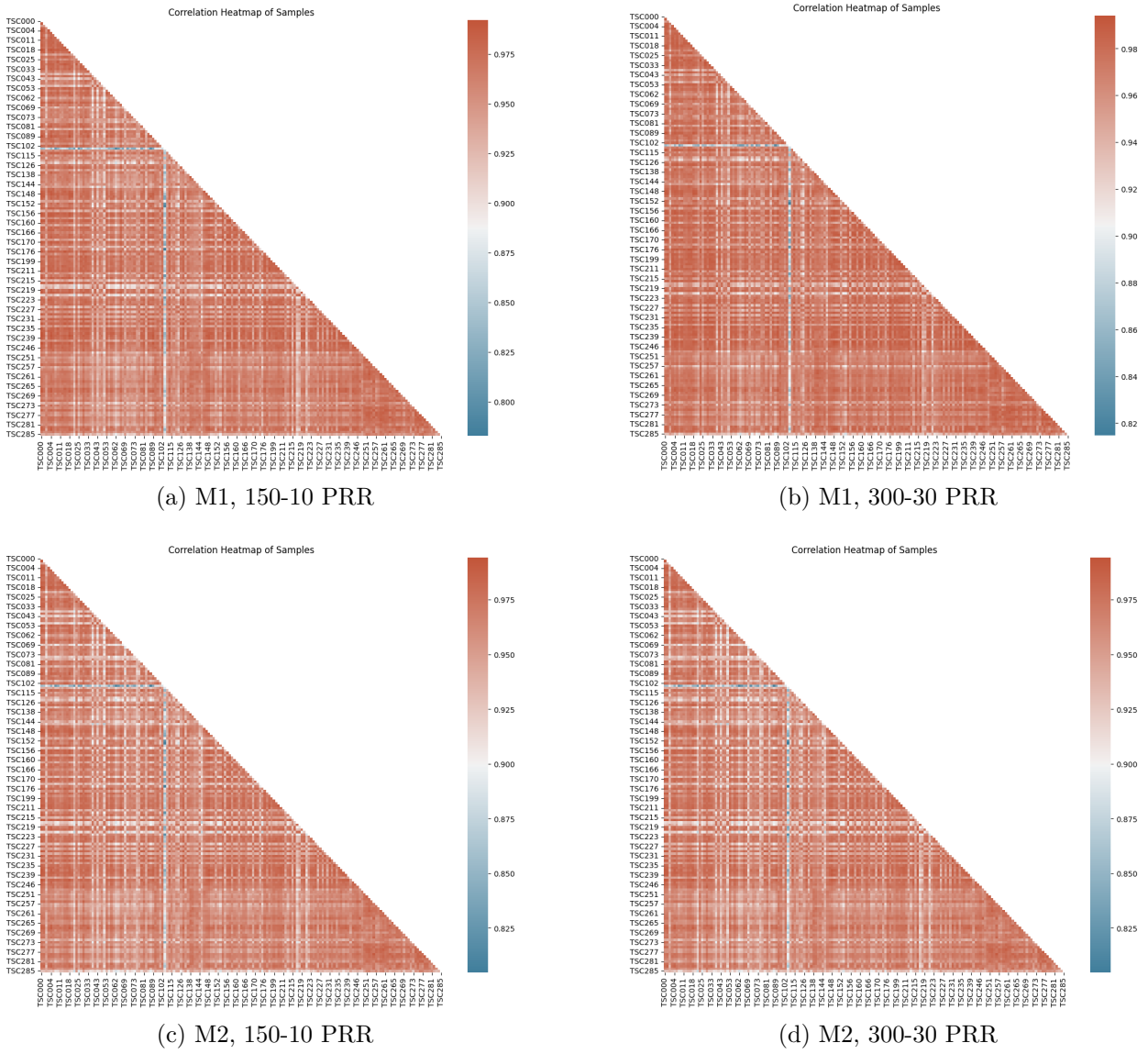


(d) M2, 300-30 PRR

Figure A.1: Temporal Sequencing of Transcription Factor Abundance: Comparative visualization for two distinct Potential Regulatory Regions (150-10 and 300-30) and two distinct Class Matrix (M1 and M2)

From the Figure A.1, we can observe that while there is a strong relationship between the samples, the series do not follow in the same way for the M2 matrix as for the M1 matrix (or the M0).

This can be better appreciated in the following correlation Figure A.2.



(a) M1, 150-10 PRR

(b) M1, 300-30 PRR

(c) M2, 150-10 PRR

(d) M2, 300-30 PRR

Figure A.2: Sample correlations for two distinct Potential Regulatory Regions (150-10 and 300-30) and two distinct Class Matrix (M1 and M2)

Annex B. Biotic data correlation hierarchy clustering

This annexed section provides a detailed view of transcription factors' (TFs) correlation patterns, revealed through hierarchical clustering of the biotic data (shown in Figure 4.6.b). To provide a more comprehensive visual overview, we established a threshold that resulted in the formation of four distinct clusters, as depicted in Figure B.1. The specific transcription factors that are categorized within these four clusters are detailed in Table B.1.

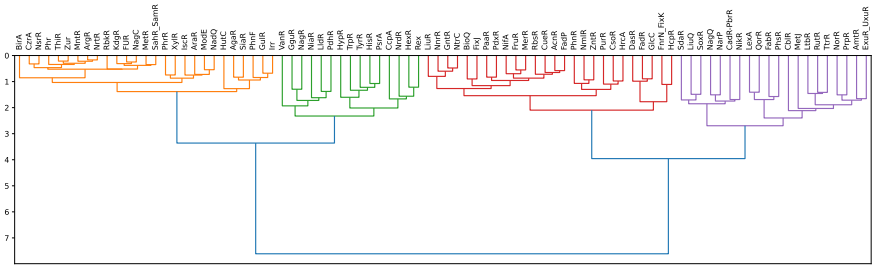


Figure B.1: Dendrogram where the distance threshold has been set to delineate four distinct clusters.

Table B.1: Clustered Transcription Factors from Correlation Matrix. This table lists all the transcription factors analyzed, organized based on the clustering from the hierarchical heatmap. The heatmap was thresholded to result in four distinct clusters. These clusters group transcription factors with similar correlation patterns, aiding in the identification of potential cooperative or antagonistic transcription factor interactions. This structure enhances the interpretability of the complex correlation patterns among the large number of transcription factors in our study.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
AgaR	CcpA	AcnR	AmtR
AraR	GguR	BioQ	CadR-PbrR
ArgR	HexR	CsoR	CblR
BirA	HisR	CueR	ExuR_UxuR
CzrA	HypR	DasR	FabR
FUR	LldR	FadP	LexA
GulR	NagR	FadR	LiuQ
HutC	NiaR	FixJ	LtbR
Irr	NrdR	FnrN_FixK	MetJ
IscR	PdhR	FruR	NagQ
KdgR	PsrA	GlcC	NarP
MetR	Rex	GntR	NikR
MntR	TrpR	HcpR	NorR
ModE	TyrR	HrcA	PhsR
NadQ	VanR	LiuR	PrpR
NagC		MerR	QorR
NrtR		NifA	RutR
NsrR		NmlR	SdaR
PhnF		NnrR	SoxR
Phr		NtrC	TtrR
PhrR		PaaR	
RbkR		PdxR	
SahR_SamR		PhnR	
SiaR		PurR	
ThiR		RbsR	
XylR		ZntR	
Zur			

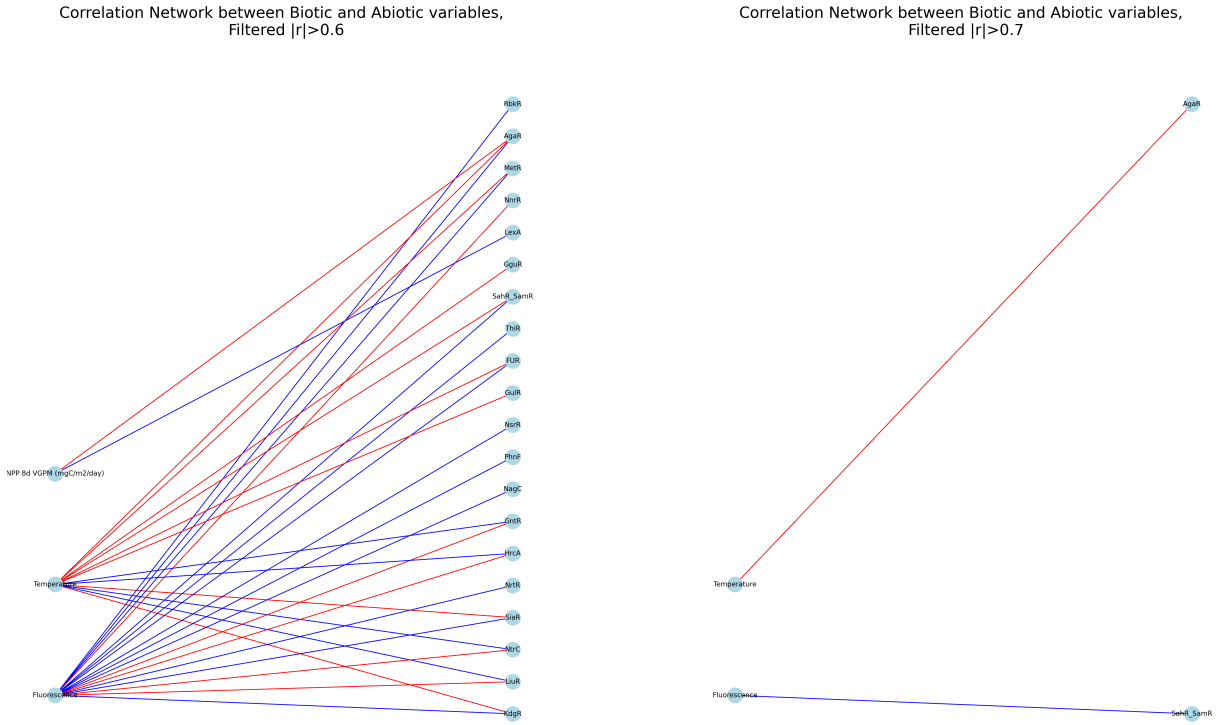
Annex C. Biotic and Abiotic Correlation

In this section, we delve deeper into the details of Figure 4.8. We specifically elucidate the transcription factors that exhibit the strongest correlations with their respective environmental variables. The specific details are exhibit in Table C.1

Table C.1: Biotic and Abiotic Correlated Variables with a 0.5 Threshold. This table presents the transcription factors that show a significant correlation (above 0.5) with each environmental variable. It offers a simplified yet substantial perspective on the interplay between environmental factors and genetic regulation, thereby focusing on the most meaningful relationships.

Depth.nominal	Temperature	NO3	Fluorescence	Salinity	Carbon.total	NPP 8d VGPM (mgC/m2/day)	ChlorophyllA	PAR.PC
KdgR	KdgR	KdgR	KdgR	LiuR	LiuR	LiuR	FabR	CadR-PbrR
FixJ	LiuR	NrtR	AraR	GguR	NtrC	SiaR		
NrtR	NtrC	AcnR	LiuR		SahR_SamR	GntR		
ArgR	SiaR	GntR	NtrC		AgaR	GulR		
NagC	HrcA	NagC	AcnR		ThiR	FUR		
PaaR	GntR	NsrR	AgaR			PaaR		
BirA	Zur	PaaR	ArgR			GguR		
Phr	NagC	SahR_SamR	CsoR			LexA		
FadR	NsrR	FadP	CueR			Irr		
ThiR	GulR	MetR	CzrA			SahR_SamR		
MerR	FUR	Phr	FUR			AgaR		
GlcC	AgaR	FadR	FabR			CsoR		
MntR	CsoR		GguR			CzrA		
	FabR		GntR			PurR		
	GguR		GulR					
	Irr		HrcA					
	LexA		Irr					
	MetR		IscR					
	NnrR		MetR					
	PaaR		MntR					
	PurR		NadQ					
	RbkR		NagC					
	SahR_SamR		NnrR					
	ThiR		NrtR					
			NsrR					
			PaaR					
			PdxR					
			PhnF					
			Phr					
			PurR					
			RbkR					
			SahR_SamR					
			SiaR					
			ThiR					
			Zur					

For a more detailed description, we could increase the threshold to 0.6 and 0.7. The results are displayed in the following Figure:



(a) (b)

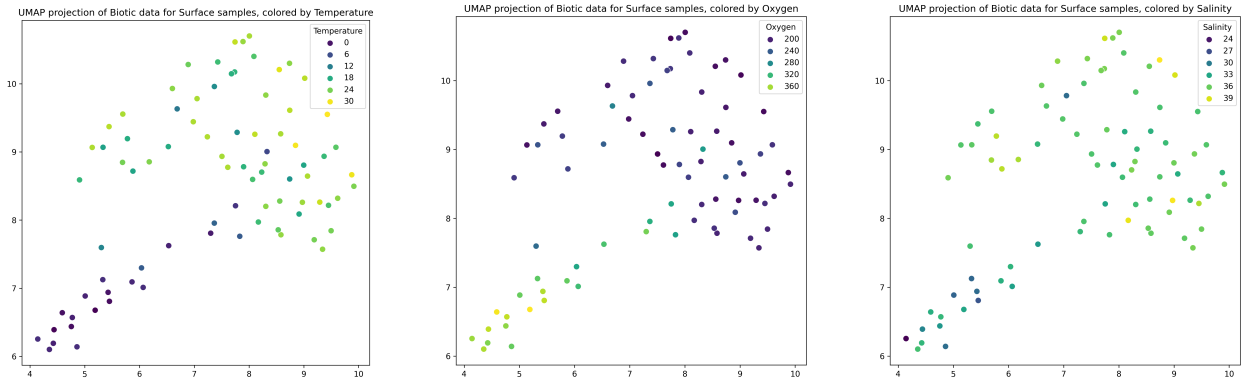
Figure C.1: Correlation networks of transcription factors and environmental variables at different thresholds. (a) Correlation network at a threshold of 0.6. This network retains relationships that have an absolute correlation of 0.6 or above, thereby revealing the stronger associations in the system. (b) Correlation network at a more stringent threshold of 0.7. Here, only the strongest relationships with an absolute correlation of 0.7 or above are maintained. The progressive increase in the threshold aids in highlighting the most crucial interactions in the marine ecosystem.

Annex D. UMAP Projection and its relation to environmental factors

In this section, we extend our exploration to observe how various abiotic variables manifest within the UMAP projections previously analyzed in the primary study. This deep dive seeks to illuminate the interplay between these environmental factors and the intrinsic structure unearthed through UMAP. By juxtaposing these projections against a diverse set of abiotic variables, we aim to uncover patterns, consistencies, or deviations that might offer a richer understanding of the ecosystem dynamics at play. We will separate out study by considering different subsamples accordingly.

D.1. Projection of Surface Samples

A pertinent inquiry is the manifestation of environmental features within the 2D UMAP projection of Surface Samples. Upon exploration, the only attributes yielding significant insights were Temperature, Oxygen, and Salinity:



(a) (b) (c)

Figure D.1: 2D UMAP projection of Surface Samples highlighting the influence of distinct environmental variables. **(a)** Temperature, **(b)** Oxygen, and **(c)** Salinity. Each visualization underscores the distribution and clustering tendencies of samples based on the respective environmental parameters.

The discernible patterns among these variables align with prior expectations, especially given the robust correlation witnessed between them. This coherence is further anchored in theory as highlighted by [49].

Annex E. Selected Transcription Factors via RFECV

While Table 5.4 summarizes the performance metrics, it omits the actual transcription factors chosen. Below, we present these factors, organized as: [**Samples considered**] $\longrightarrow$ [**Target variable**].

All samples $\longrightarrow$ Polar

'AgaR', 'CcpA', 'DasR', 'ExuR_UxuR', 'FabR', 'FadR', 'GguR', 'GlcC', 'HexR', 'HisR', 'HrcA', 'NadQ', 'NarP', 'NifA', 'NorR', 'NtrC', 'PaaR', 'PhnR', 'PrpR', 'PsrA', 'PurR', 'SoxR', 'TrpR', 'TyrR'

All samples $\longrightarrow$ SRF/DCM/MES

'AcnR', 'AgaR', 'AmtR', 'AraR', 'ArgR', 'BioQ', 'BirA', 'CadR-PbrR', 'CblR', 'CcpA', 'CsoR', 'CueR', 'CzrA', 'DasR', 'ExuR_UxuR', 'FUR', 'FabR', 'FadP', 'FadR', 'FixJ', 'FnrN_FixK', 'FruR', 'GguR', 'GlcC', 'GntR', 'GulR', 'HcpR', 'HexR', 'HisR', 'HrcA', 'HutC', 'HypR', 'Irr', 'IscR', 'KdgR', 'LexA', 'LiuQ', 'LiuR', 'LldR', 'LtbR', 'MerR', 'MetJ', 'MetR', 'MntR', 'ModE', 'NadQ', 'NagC', 'NagQ', 'NagR', 'NarP', 'NiaR', 'NifA', 'NikR', 'NmlR', 'NnrR', 'NorR', 'NrdR', 'NrtR', 'NtrC', 'PaaR', 'PdhR', 'PdxR', 'PhnF', 'PhnR', 'PhrR', 'PhsR', 'PrpR', 'PsrA', 'PurR', 'QorR', 'RbkR', 'RbsR', 'Rex', 'RutR', 'SdaR', 'SoxR', 'ThiR', 'TrpR', 'TtrR', 'TyrR', 'VanR', 'XylR', 'ZntR', 'Zur'

All samples ⟶ EPI/MES

'BirA', 'CadR-PbrR', 'CsoR', 'FadP', 'FadR', 'GlcC', 'KdgR', 'LexA', 'LiuR', 'NikR', 'NnrR', 'NtrC', 'PaaR', 'PdxR', 'PhrR', 'PurR', 'RbsR', 'XylR'

All samples ⟶ Ocean

'AcnR', 'AgaR', 'AmtR', 'AraR', 'ArgR', 'BioQ', 'BirA', 'CadR-PbrR', 'CblR', 'CcpA', 'CsoR', 'CueR', 'DasR', 'ExuR_UxuR', 'FabR', 'FadP', 'FadR', 'FixJ', 'FnrN_FixK', 'FruR', 'GguR', 'GlcC', 'GntR', 'GulR', 'HcpR', 'HexR', 'HisR', 'HrcA', 'HutC', 'HypR', 'Irr', 'IscR', 'KdgR', 'LexA', 'LiuQ', 'LiuR', 'LldR', 'LtbR', 'MerR', 'MetR', 'MntR', 'ModE', 'NadQ', 'NagC', 'NagQ', 'NagR', 'NarP', 'NiaR', 'NifA', 'NikR', 'NmlR', 'NnrR', 'NorR', 'NrdR', 'NrtR', 'NsrR', 'NtrC', 'PaaR', 'PdhR', 'PdxR', 'PhnR', 'PhrR', 'PhsR', 'PrpR', 'PsrA', 'PurR', 'QorR', 'RbsR', 'Rex', 'RutR', 'SahR_SamR', 'SdaR', 'SiaR', 'SoxR', 'TrpR', 'TtrR', 'TyrR', 'VanR', 'XylR', 'ZntR'

All samples ⟶ Province

'AcnR', 'AgaR', 'AmtR', 'AraR', 'BirA', 'CadR-PbrR', 'CblR', 'CcpA', 'CueR', 'DasR', 'ExuR_UxuR', 'FabR', 'FadP', 'FadR', 'FixJ', 'FnrN_FixK', 'GguR', 'GlcC', 'GulR', 'HcpR', 'HexR', 'HisR', 'HrcA', 'HypR', 'IscR', 'KdgR', 'LiuQ', 'MerR', 'MetR', 'MntR', 'ModE', 'NadQ', 'NagQ', 'NagR', 'NarP', 'NorR', 'NrdR', 'NsrR', 'NtrC', 'PaaR', 'PhnF', 'PsrA', 'RbkR', 'RbsR', 'SdaR', 'SoxR', 'TrpR'

Surface samples ⟶ Polar

'AgaR', 'FadR', 'GguR', 'LexA', 'PurR', 'VanR'

Surface samples ⟶ Ocean

['AgaR', 'AraR', 'CadR-PbrR', 'CcpA', 'FabR', 'FadR', 'FnrN_FixK', 'GguR', 'GulR', 'HrcA', 'HypR', 'ModE', 'NagQ', 'NiaR', 'NnrR', 'NrdR', 'PrpR', 'SoxR', 'TyrR'

Surface samples ⟶ Province

'AcnR', 'AgaR', 'AmtR', 'BirA', 'CadR-PbrR', 'CblR', 'CcpA', 'CsoR', 'CueR', 'CzrA', 'DasR', 'ExuR_UxuR', 'FabR', 'FadP', 'FadR', 'FixJ', 'FnrN_FixK', 'FruR', 'GguR', 'HcpR', 'HrcA', 'HutC', 'HypR', 'Irr', 'IscR', 'LexA', 'LiuQ', 'LiuR', 'LldR', 'MerR', 'MetJ', 'MntR', 'ModE', 'NadQ', 'NagR', 'NikR', 'NmlR', 'NnrR', 'NorR', 'NrdR', 'NsrR', 'NtrC', 'PaaR', 'PdxR', 'PhnF', 'Phr', 'PhsR', 'PsrA', 'PurR', 'RbkR', 'SdaR', 'SoxR', 'TtrR', 'TyrR', 'VanR', 'ZntR'

Non Polar samples ⟶ SRF/DCM/MES

'BirA', 'CadR-PbrR', 'CueR', 'FabR', 'FadR', 'FixJ', 'GguR', 'HutC', 'LexA', 'LldR', 'MetR', 'NagQ', 'NikR', 'NrtR', 'PhnR', 'PhsR', 'QorR', 'TrpR', 'VanR', 'ZntR'

Non Polar samples ⟶ EPI/MES

'GguR', 'GlcC', 'GntR', 'GulR', 'IscR', 'MetR', 'NarP', 'NrtR', 'PdxR', 'PhnF', 'PrpR'

Non Polar samples ⟶ Ocean

'AcnR', 'AraR', 'ArgR', 'BirA', 'CadR-PbrR', 'CblR', 'CcpA', 'CueR', 'CzrA', 'FabR', 'FadR', 'FnrN_FixK', 'FruR', 'GguR', 'GlcC', 'GntR', 'GulR', 'HcpR', 'HexR', 'HrcA', 'HutC', 'HypR', 'IscR', 'KdgR', 'LiuR', 'LldR', 'LtbR', 'MerR', 'ModE', 'NadQ', 'NagR', 'NarP', 'NiaR', 'NifA', 'NorR', 'NrdR', 'NtrC', 'PaaR', 'PdhR', 'PdxR', 'PhrR', 'PhsR', 'PrpR', 'PsrA', 'RbsR', 'Rex', 'RutR', 'SahR_SamR', 'SdaR', 'SiaR', 'SoxR', 'TtrR', 'VanR', 'XylR', 'ZntR'

Non Polar samples ⟶ Province

'AgaR', 'AmtR', 'AraR', 'BioQ', 'CadR-PbrR', 'CblR', 'CcpA', 'CueR', 'CzrA', 'DasR', 'FabR', 'FadP', 'FadR', 'FixJ', 'FruR', 'GguR', 'GlcC', 'HcpR', 'HexR', 'HrcA', 'HypR', 'LiuQ', 'LiuR', 'NadQ', 'NagQ', 'NagR', 'NorR', 'NsrR', 'NtrC', 'PhnF', 'PsrA', 'RbkR', 'RbsR', 'Rex', 'SoxR', 'ThiR', 'TrpR', 'TtrR', 'ZntR'