

0001010101010100001

01010101010

0101010

01010

1001

101

011100

1110

011100011100

011100010101010

0101010101010000101110011100

010101010010

101010100101

0101000101110

001001101010

01010101010101011010101010000101010101000101010

010101010101010101101010101000101010101010000101010

01110000101010101010001

0101

110

010101010101010101

01110001010

00101

00

01110000101010101010

01010101010101010101

011100001010101010001

01010101010101010101

01

01110000101010101010

01110000101010

Inria



Modeling global plankton communities via multiomics and ML approaches

Luis Valenzuela
Inria Chile Research Center, Las Condes, Chile.

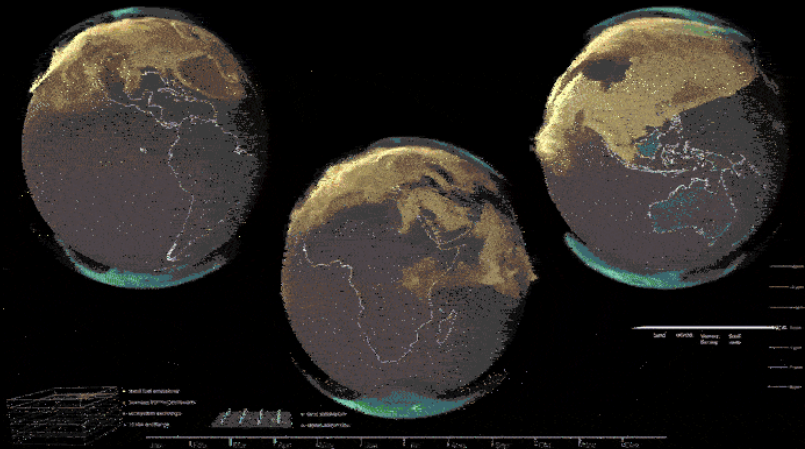
`luis.valenzuela@inria.cl`

<https://inria.cl>

Réunion Annuel OcéanIA
23 February 2023

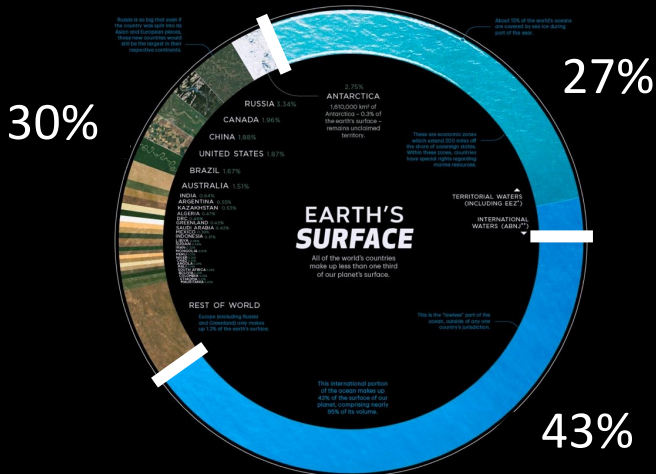
The Inria logo is written in a red, cursive script font.

Climate Change Driver: CO₂ emissions

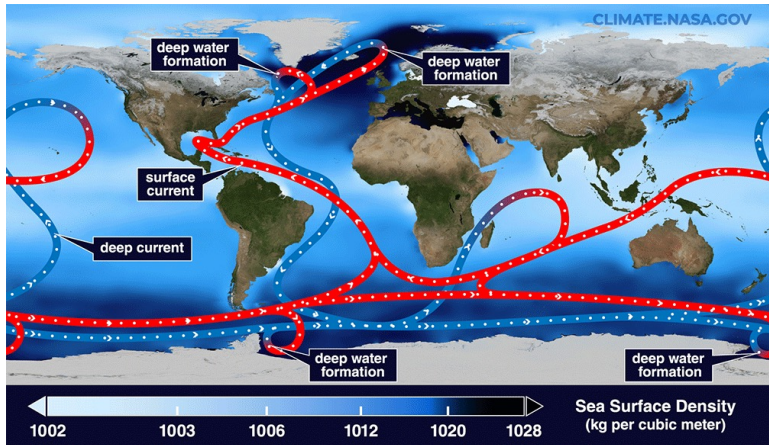


<https://svs.gsfc.nasa.gov/5110>

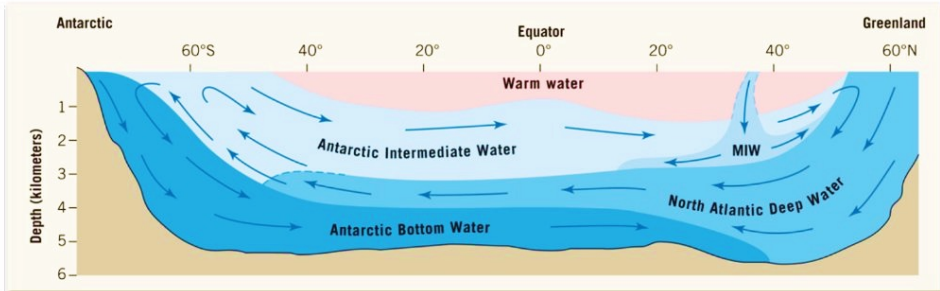
Climate Change Mitigator: Planet Ocean



Thermohaline circulation: great ocean conveyor belt



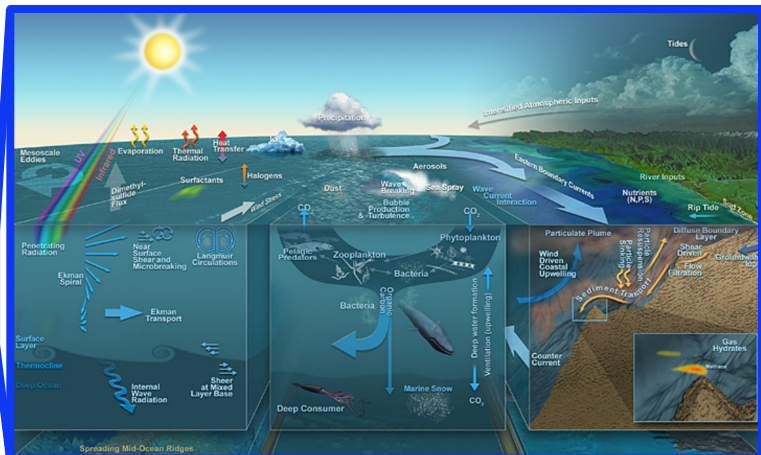
Cross section of the deep circulation in the Atlantic Ocean



Characterization of these water masses:

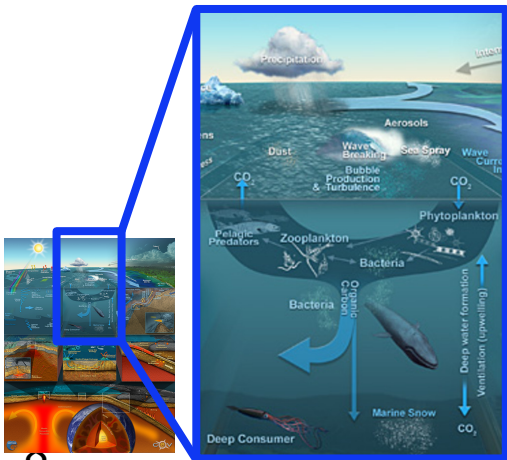
- Temperature-Salinity diagrams
- Isotopes (H, C, O)
- Bioinformatics

Bio-Geo-Physicochemical Oceanography: Marine Ecosystems



http://www.cev.washington.edu/file/Earth_and_Ocean_Processes.html

Key Invisible Microbiome World & Biological Carbon Pump



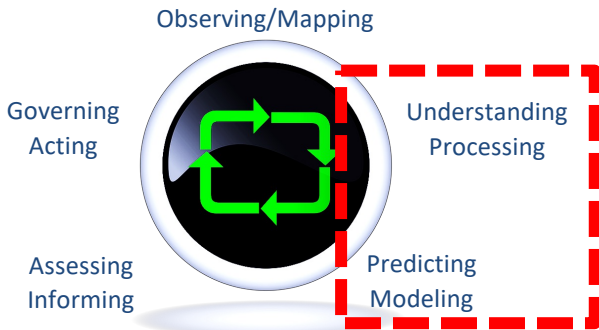
80% of marine life is made up of microorganisms

50% of the oxygen produced each day is provided by marine microorganisms

30% of the CO₂ emitted each day is captured by the ocean and its biodiversity.

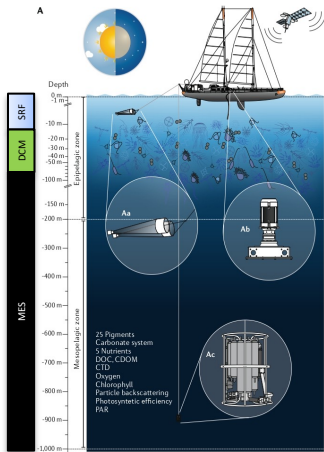
Management of marine ecosystems

Policy

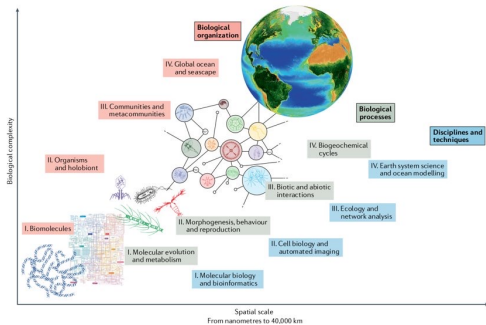
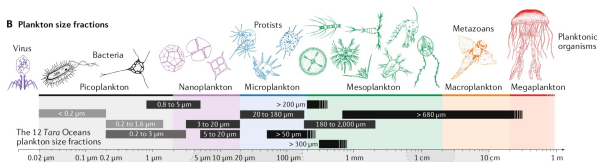


Science

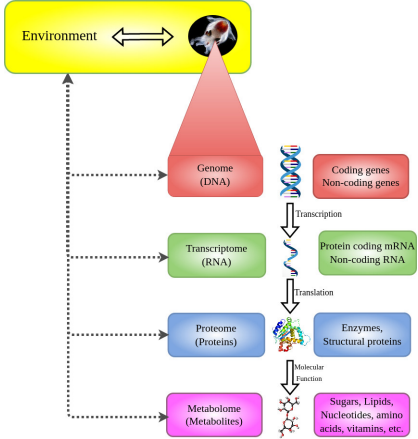
Ambitious goal: towards global ocean ecosystems biology



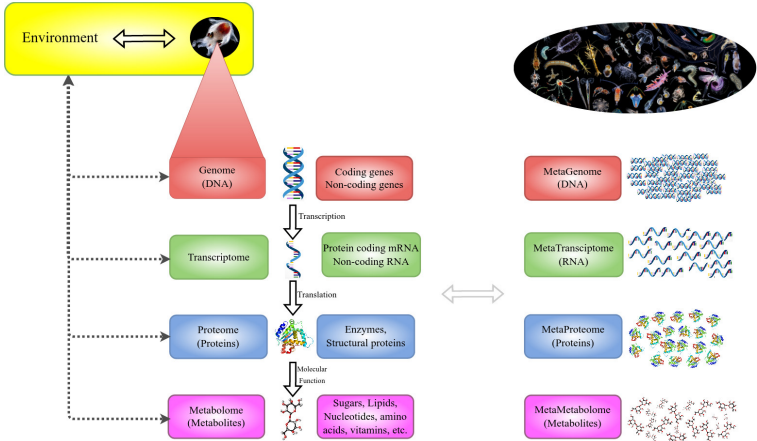
Sunagawa et al. (2020)



Global plankton communities: multi-omics data



Global plankton communities: multi-omics data



Ocean Microbial Reference Gene Catalog v2

Metagenomic dataset only:

~ 57.000 million reads (90 pb \pm 2.6pb)

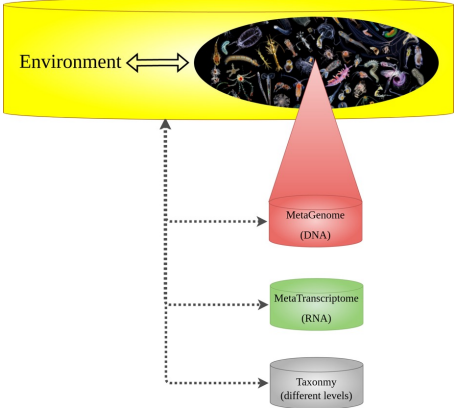
```
ATGC...CTGGG90  ...  TTCA...TTTCC90  
GCC...AAAAG90  ...  GGGGA...AGCTA90
```

(meta)genome assembly

~ 200 metagenomes:
~ 230000 scaffolds, N50: 1300pb (average)
~ 42 million genes

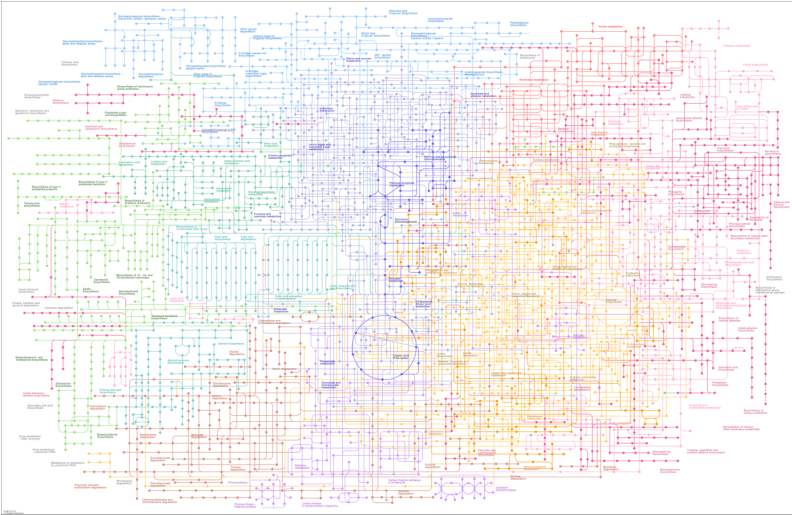
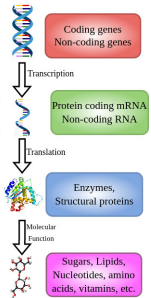
Functional annotation
Quantification
Normalization

	Gene1	Gene2	...	Gene42M
metaG1				
metaG2				
...				
metaG200				



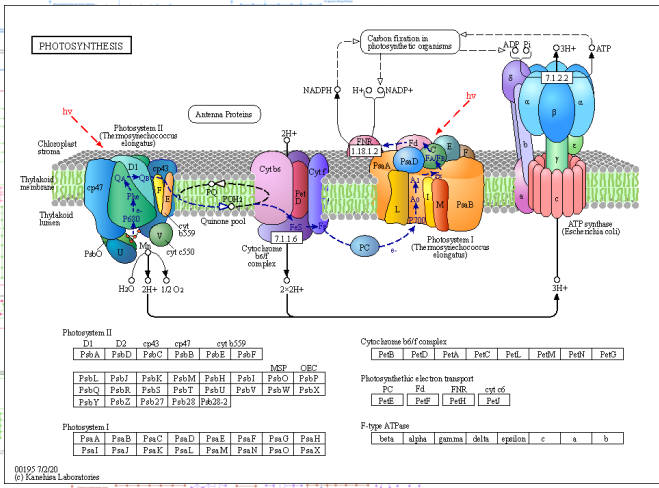
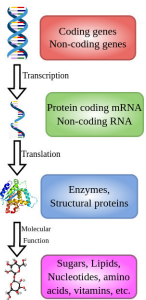
Metabolic pathways – Reference pathway

Edges: enzymes
Dots: metabolites



Metabolic pathways – Reference pathway

KEGG pathway example: Photosynthesis



Photosystem II

D1	D2	cp43	cp47	cyt b559
PsbA	PsbD	PsbC	PsbB	PsbE

Photosystem I

PsaA	PsaB	PsaC	PsaD	PsaE	PsaF	PsaG	PsaH
PsaI	PsaJ	PsaK	PsaL	PsaM	PsaN	PsaO	PsaX

Cytochrome b₆/f complex

PetB	PetD	PetA	PetC	PetL	PetM	PetN	PetG
------	------	------	------	------	------	------	------

Photosynthetic electron transport

PC	Fd	FNR	cyt c6
PetE	PetF	PetH	PetI

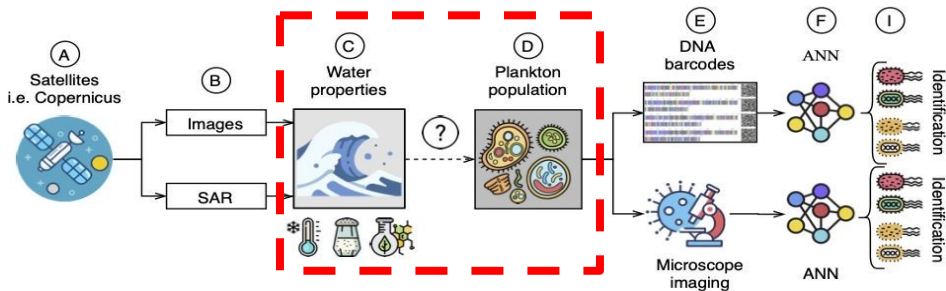
F-type ATPase

beta	alpha	gamma	delta	epsilon	c	a	b
------	-------	-------	-------	---------	---	---	---

Understanding plankton communities using AI & ML

Can the properties of water be inferred from the taxonomic and functional composition of plankton communities?

Is it feasible to infer the composition of plankton communities from the properties of water?



Goals:

Mapping D -> C

Mapping C -> D

Overview: Global patterns of Plankton Communities

Metagenomic composition

172 samples x 46.7M genes

	Gene1	Gene2	...	Gene47M
metaG1				
metaG2				
...				
metaG200				

Genes with known
molecular function
(KEGG): 11 M genes

Group abundances
of genes with equal
molecular function:

172 samples x 9024 molecular functions

	KO1	KO2	...	KO9024
metaG1				
metaG2				
...				
metaG200				

Metatranscriptomic composition

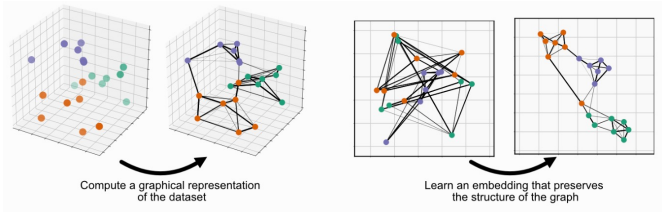
178 samples with similar number of columns.

Overview: Global patterns of Plankton Communities

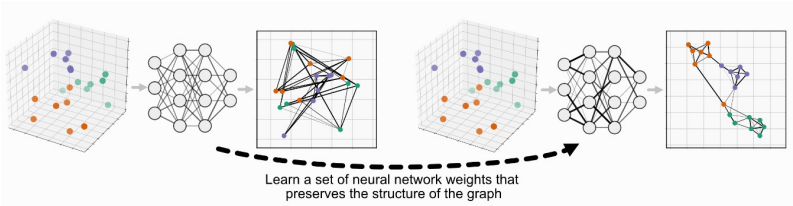
Dataset	Cardinality	Description
Environment	30	Measurements of environmental parameters, encompassing both physical and chemical attributes of water samples.
Genomic	9024	Abundances of molecular functions inferred from metagenomic assays employing the KEGG annotation database.
Transcriptomic	8935	Abundances of molecular functions identified from metatranscriptomic assays via the KEGG annotation database.
Domain	3	Relative abundances of taxonomic compositions at the Archaea, Bacteria, and Eukaryota level.
Phylum	170	Relative abundances of taxonomic compositions at the phylum level.
Class	379	Relative abundances of taxonomic compositions at the class level.
Order	534	Relative abundances of taxonomic compositions at the order level.
Family	587	Relative abundances of taxonomic compositions at the family level.
Genus	2134	Relative abundances of taxonomic compositions at the genus level.

Overview: Global patterns of Plankton Communities

UMAP model

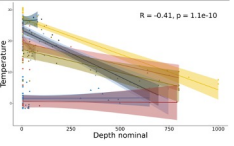
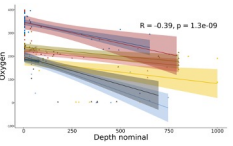
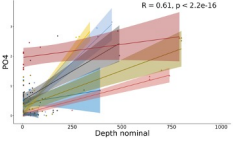
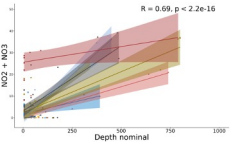
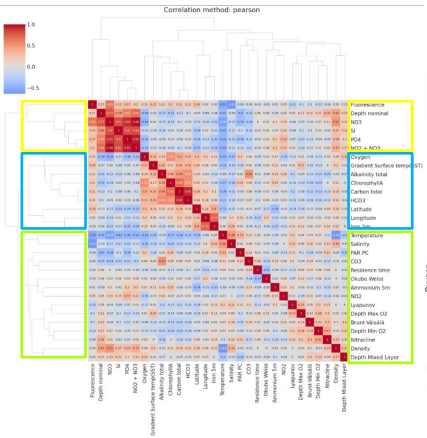


UMAP + Autoencoder model



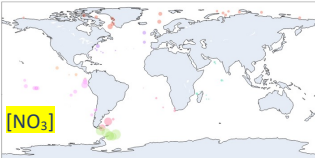
Overview: Global patterns of Plankton Communities

Environmental data exploration

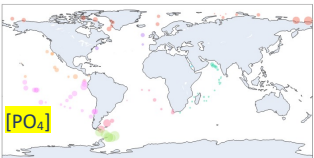


Overview: Global patterns of Plankton Communities

Environmental data exploration



Surface



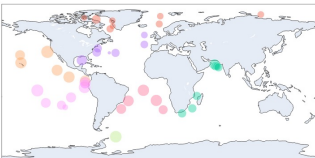
Surface



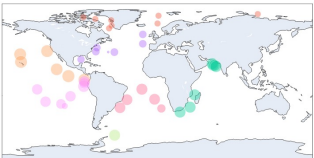
Deep Chlorophyll A Maximum



Deep Chlorophyll A Maximum



Mesopelagic

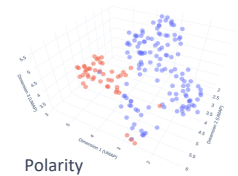


Mesopelagic

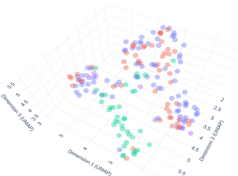
- Ocean Name
- Mediterranean Sea
 - Arctic Ocean
 - Indian Ocean
 - North Atlantic Ocean
 - North Pacific Ocean
 - Red Sea
 - South Atlantic Ocean
 - Southern Ocean
 - South Pacific Ocean

Overview: Global patterns of Plankton Communities

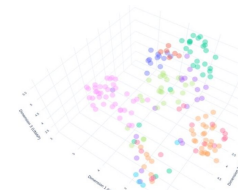
Environmental data exploration: potential niches



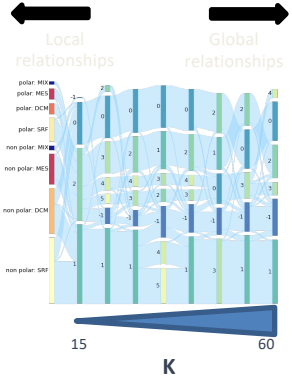
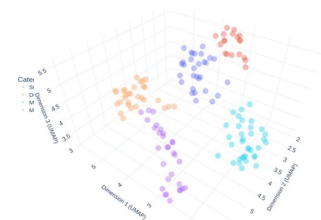
Polarity



Depth Layers

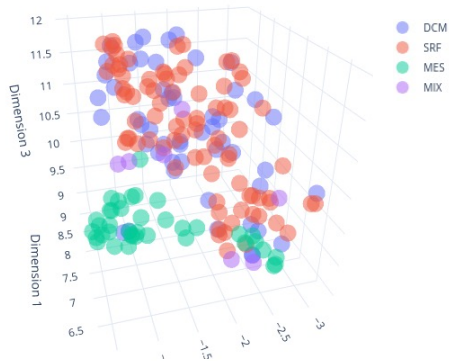
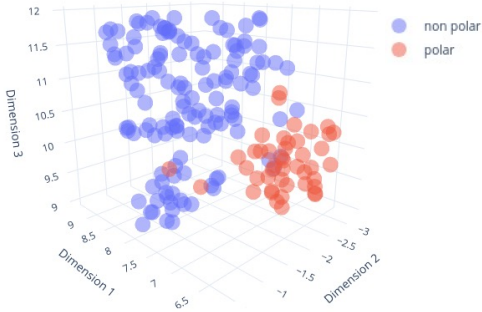


Oceans



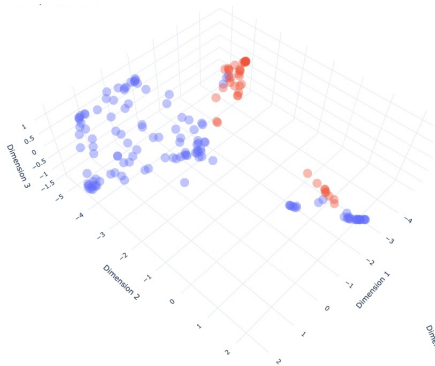
Overview: Global patterns of Plankton Communities

Input: metagenomic dataset (47M genes)



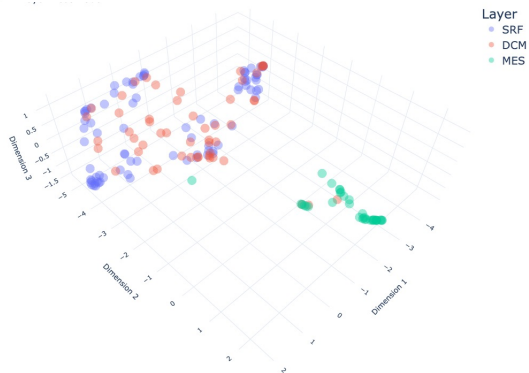
Overview: Global patterns of Plankton Communities

Input: metagenomic dataset (9024 Molecular Functions)



Polarity

- non polar
- polar

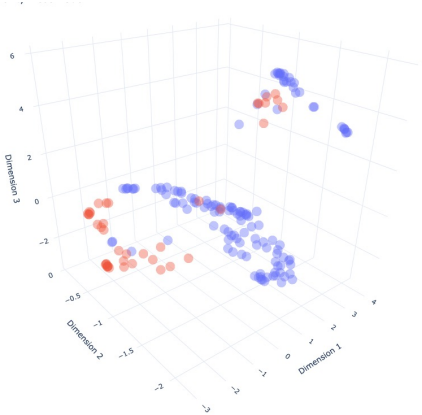


Layer

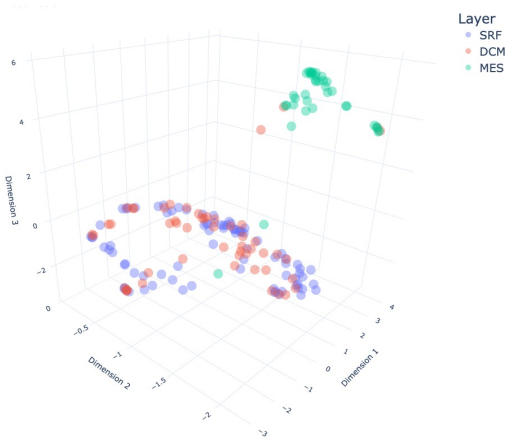
- SRF
- DCM
- MES

Overview: Global patterns of Plankton Communities

Input: metagenomic dataset (453 pathways)



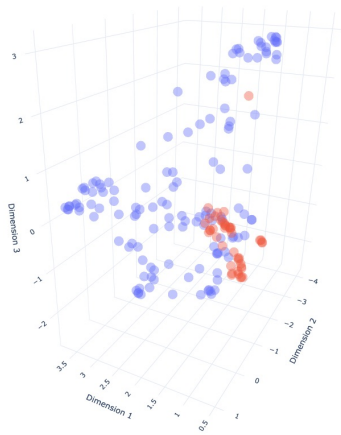
Polarity
● non polar
● polar



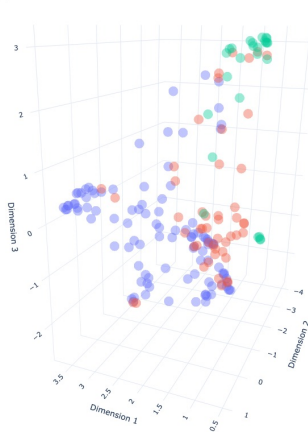
Layer
● SRF
● DCM
● MES

Overview: Global patterns of Plankton Communities

Input: metatranscriptomic dataset (8935 Molecular Functions)



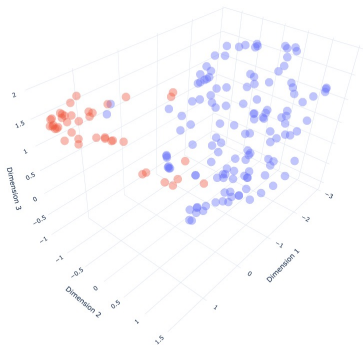
Polarity
● non polar
● polar



Layer
● SRF
● DCM
● MES

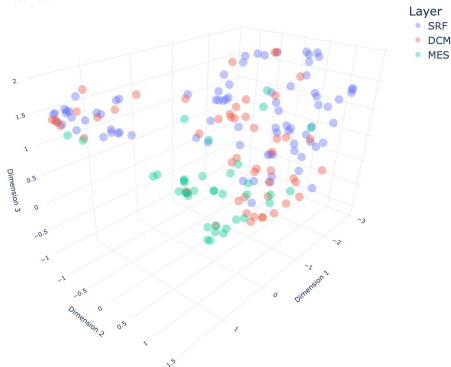
Overview: Global patterns of Plankton Communities

Input: metagenomic dataset (2124 Genus level)



Polarity

- non polar
- polar



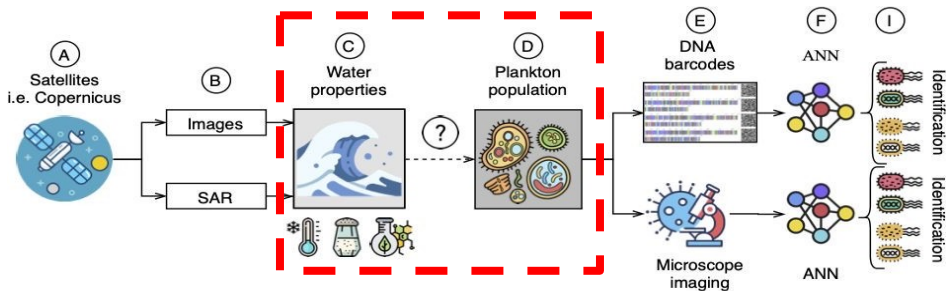
Layer

- SRF
- DCM
- MES

Understanding plankton communities using AI & ML

Can the properties of water be inferred from the taxonomic and functional composition of plankton communities?

Is it feasible to infer the composition of plankton communities from the properties of water?



Goals:

Mapping D -> C

Mapping C -> D

Predicting environmental conditions from plankton features

Symbolic regression (SR) consists in the inference of a free-form symbolic analytical function (f):

$$f: \mathbb{R}^n \rightarrow \mathbb{R}$$

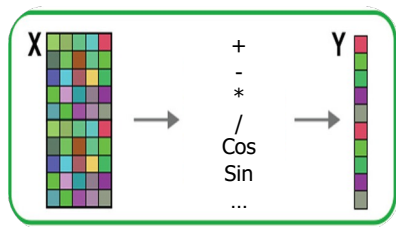
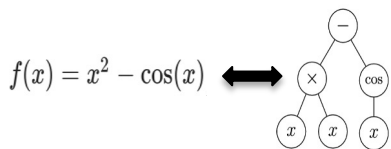
That fits

$$y = f(x_0, \dots, x_n)$$

given data

$$(x_0, \dots, x_n)$$

Predicting environmental conditions from plankton features



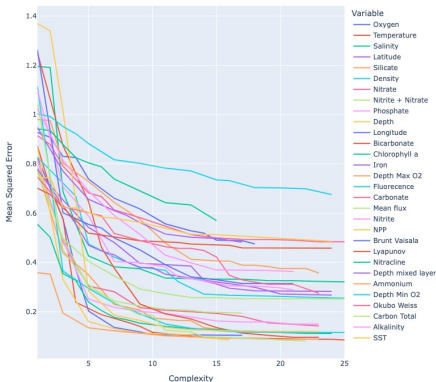
Predicting environmental conditions from plankton features

Given a set of environmental parameters $E = \{y_1, y_2, \dots, y_n\}$, and a set of input datasets $\{D_1, D_2, \dots, D_{11}\}$, the prediction process using symbolic regression can be described as follows.

Algorithm 1 Prediction of Environmental Parameters Using SR and Multiple Datasets

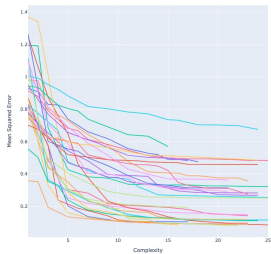
- 1: **for** each $y_i \in E$ **do**
- 2: Select a dataset D_j from $\{D_1, D_2, \dots, D_{11}\}$
- 3: Use D_j and PySR to predict y_i by inferring the function f_{ij} that best fits the available data.
- 4: **end for**

Metagenomic data predicting environmental target

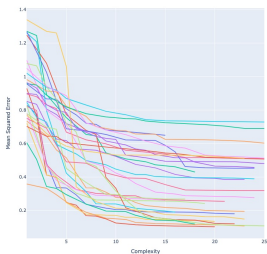


Predicting environmental conditions from plankton features

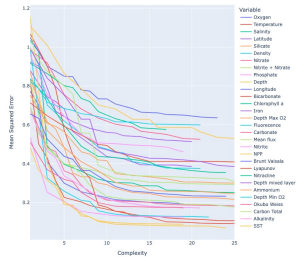
Environment vs Metagenomics



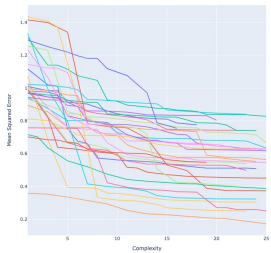
Environment vs Metagenomics (pathway level)



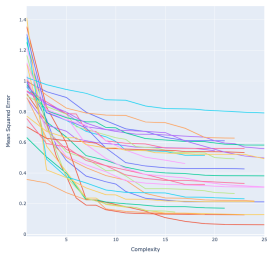
Environment vs Metatranscriptomics



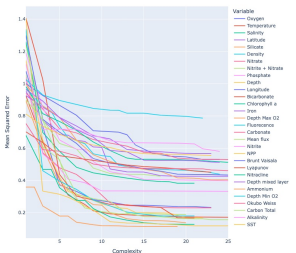
Environment vs Domain



Environment vs Family



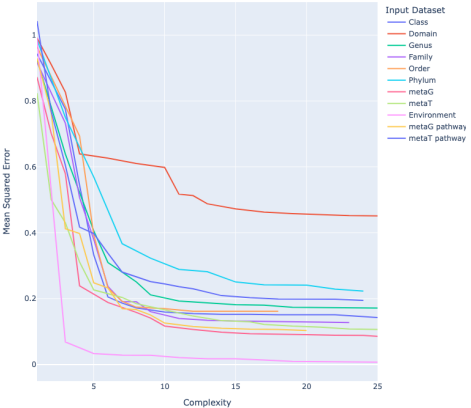
Environment vs Genus



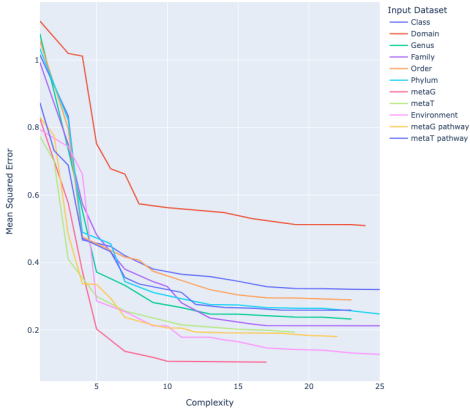
- Variable
- Oxygen
 - Temperature
 - Salinity
 - Latitude
 - Silicate
 - Density
 - Nitrate
 - Nitrite + Nitrate
 - Phosphate
 - Depth
 - Longitude
 - Bicarbonate
 - Chlorophyll a
 - Iron
 - Depth Max O2
 - Fluorescence
 - Carbonate
 - Mean flux
 - Nitrite
 - MP
 - Ernst Variable
 - Lysipinow
 - Nitroamine
 - Depth mixed layer
 - Americium
 - Depth Min O2
 - Diels-Wallis
 - Carbon Total
 - Alkalinity
 - SST

Predicting environmental conditions from plankton features

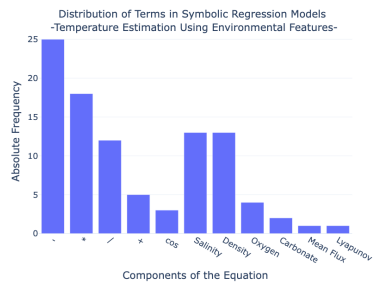
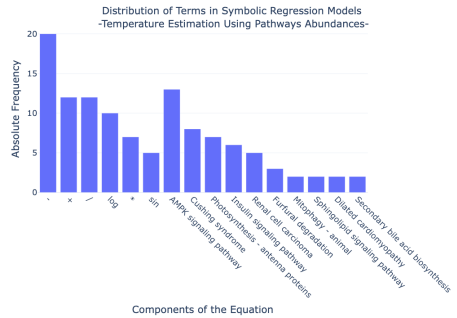
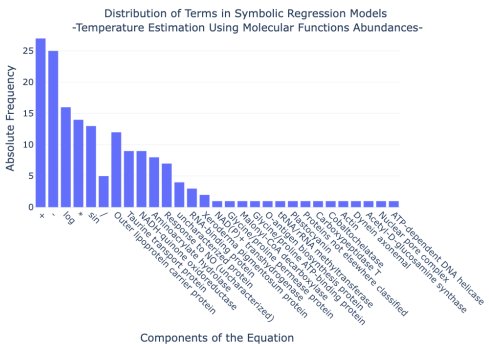
Predicting Temperature



Predicting Oxygen

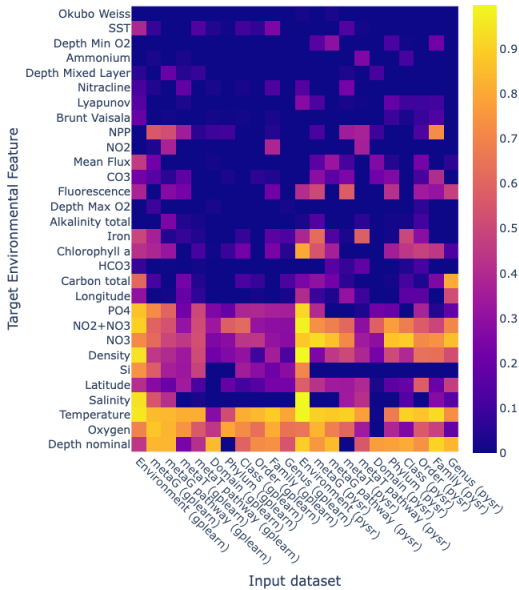


Predicting environmental conditions from plankton features



Predicting environmental conditions from plankton features

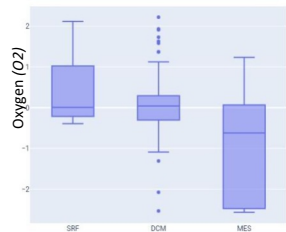
Symbolic Regressions Models performance
(R² Coefficients)



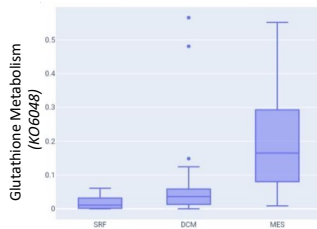
Predicting environmental conditions from plankton features

Target	Equation*	R ²
Depth	$\frac{K15635}{\left(\frac{K00524}{\sin(K15635 + \log(K07798))} + K16845\right)} - 0.50$	0.75
Oxygen	$7.13 \cdot K00856 - 7.13 \cdot K06048 + K11927$	0.83
Temperature	$K00339 - 4.51 \cdot K15551 + \sin(\log(K03634))$	0.90
Nitrate	$-K02037 + \frac{K04069}{\left(K00305 - \frac{K18477}{(K18459 - 0.14 \cdot K00392 + K05934)}\right)} + \frac{K10464}{K00055}$	0.93

K00856: adenosine kinase.
 K11927: ATP-dependent RNA
 helicase.
 K06048: glutamate-cysteine
 ligase carboxylate-amine
 ligase,



Layer



Layer



Predicting environmental conditions from plankton features

Target	Equation*	R^2
Depth	$(-\text{Chl} \cdot (\text{NO}_3 + 1.05) + \sin(\text{N})) \cdot \cos(\text{NPP} \cdot \text{O}_2 - 0.34)$	0.89
Oxygen	$-1.23 \cdot \text{NO}_3 \cdot \cos(0.80 \cdot \text{L}) - \text{T}$	0.66
Temperature	$-0.92 \cdot \text{D} + 0.81 \cdot \text{S} + 0.19 \cdot \cos(1.22 \cdot \text{D} - 1.88 \cdot \text{S})$	0.99
Nitrate	$0.08 \cdot \text{L} + \text{N} - 0.01 \cdot (0.74 - \text{I})/\text{Si}$	0.91

Chl: Chlorophyll A, NO₃: Nitrate, N: Nitrite plus nitrate, NPP: Net Primary Production, L: Latitude, T: Temperature, D: Density, S: Salinity, I: Iron, Si: Silicate.

Predicting environmental conditions from plankton features

Symbolic classifications performance (F1)

Oceanic Layers Comparison

Surface vs DCM

DCM vs Mesopelagic

Surface vs Mesopelagic

1.00

0.76

0.80

0.75

0.94

0.86

0.90

0.90

1.00

1.00

1.00

1.00

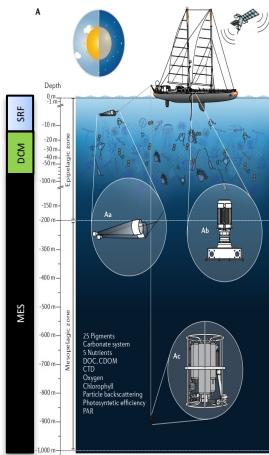
Environment

metaG

metaG pathway

Family

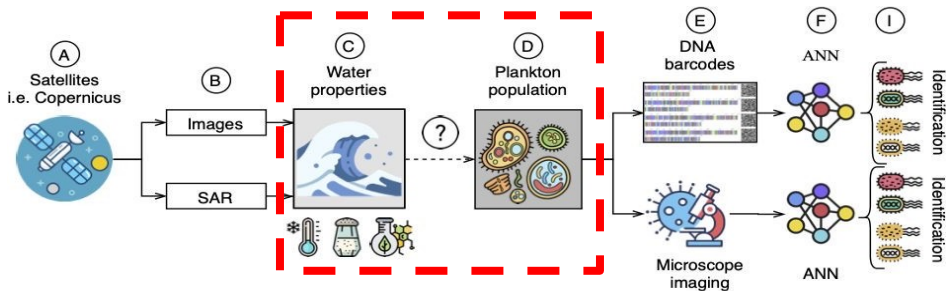
Input dataset



Understanding plankton communities using AI & ML

Can the properties of water be inferred from the taxonomic and functional composition of plankton communities?

Is it feasible to infer the composition of plankton communities from the properties of water?



Goals:

Mapping D -> C

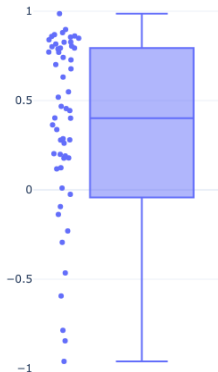
Mapping C -> D

Predicting plankton features from environmental conditions

Using environmental features to predict directly key metagenomic features

61 Molecular Functions

(from metagenomic dataset)



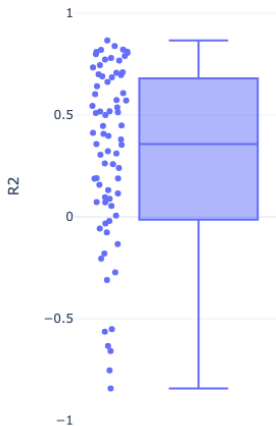
K05724	0.9868	$5.4993143e-6/(\text{Iron}5m - 0.94)$
K00275	0.8970	$-0.072*\text{DepthN} - 0.072*\text{PO4} + 0.072*\text{Temp} + 0.851$
K11927	0.8796	$0.173*\text{ChlorophyllA} - 0.173*\text{Temp} + 0.173*\sin(\text{O2}) + 0.307$
K15551	0.8693	$-0.021*\text{DepthN} - 0.071*\text{Temp} + 0.1127$
K03433	0.8610	$0.333*\log(\text{DepthN} + 1.612)$
K02037	0.8584	$(0.0457*\text{Iron}5m + 0.200)/(\text{PO4} + 1.156)$
K15635	0.8553	$0.202*\log(\text{DepthN} + 1.570)$
K03163	0.8489	$\text{DepthN}*(0.075 - 0.0186*\text{Longit}) + 0.0435$
K07574	0.8394	$-0.062*\text{Temp} + 0.062*\sin(\text{Latit}*(-\text{CarbonT} + \text{ChlorophyllA} - \text{Latit})) + 0.09$
K02533	0.8270	$0.093*\text{Temp} + 0.297$
K13525	0.8264	$0.470*\log(\text{DepthN} + 1.654)$
K05501	0.8166	$0.030*\text{ChlorophyllA} + 0.030*\text{O2} + 0.108$
K05934	0.8053	$0.061*\text{DepthN} + 0.061*\sin(\text{DepthN}) + 0.0619$
K00324	0.8007	$0.216*\text{Temp} + 1.201$

Predicting plankton features from environmental conditions

Using environmental features to predict directly key metagenomic features

79 Pathways

(from metagenomic dataset)

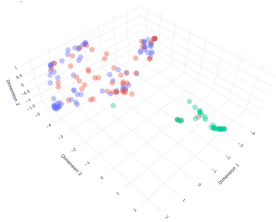


Glycine, serine and threonine metabolism	0.865	$\text{ChlorophyllA} - \text{Temp} + 25.171$
Retinol metabolism	0.838	$0.126 \cdot \text{DepthN} - 0.126 \cdot \text{Temp} + 0.470$
Leishmaniasis	0.821	$0.089 \cdot \text{DepthN} + 0.104$
Phenylalanine metabolism	0.820	$0.536 \cdot \text{ChlorophyllA} - 0.536 \cdot \text{Temp} + 6.649$
Fanconi anemia pathway	0.811	$0.181 \cdot \text{DepthN} + 0.195$
Basal transcription factors	0.810	$\log(\text{DepthN} + 1.800)$
Mitophagy	0.805	$0.145 \cdot \text{DepthN} - 0.145 \cdot \text{Temp} + 0.979$

Predicting plankton features from environmental conditions

metagenomic dataset
(9024 Molecular Functions)

UMAP
+
Autoencoder

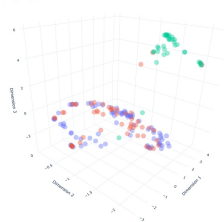


Symbolic
Regressions

Environmental features

metagenomic dataset
(453 pathways)

UMAP
+
Autoencoder

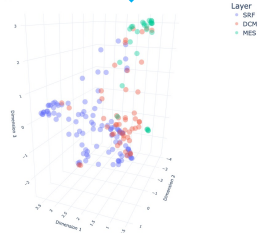


Symbolic
Regressions

Environmental features

metatranscriptomic dataset
(8935 Molecular Functions)

UMAP
+
Autoencoder

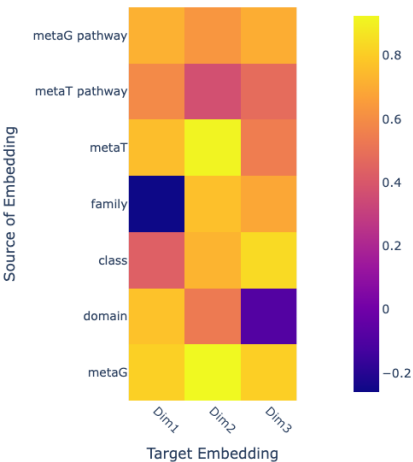


Symbolic
Regressions

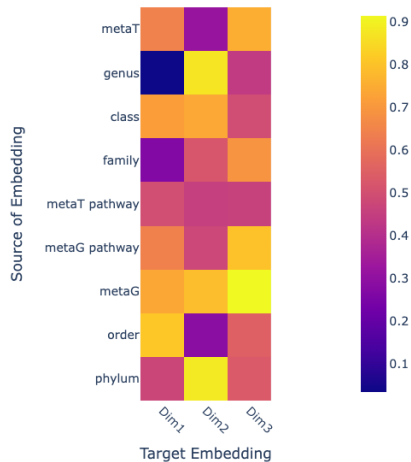
Environmental features

Predicting plankton features from environmental conditions

SR models predicting euclidean embedding



SR models predicting hiperbolic embedding



Conclusion

- After reducing the dimensionality of each of the datasets and visualizing general patterns, the influence of polarity and the depth at which the sample was taken stand out. Categories such as provinces or oceans only group samples when they come from the surface.
- There is a subset of environmental variables that are consistently predicted by omics data, reflecting their potential regulation.
- Similarly, there is a subset of omics-origin variables that are predicted with high performance, also indicating potential mechanisms of gene expression regulation based on environmental factors.
- Depending on the layers analyzed, differences may focus on the use of antibiotics in surface layers, or on processes related to protein folding or proteostasis when compared with deeper layers.

Acknowledgements

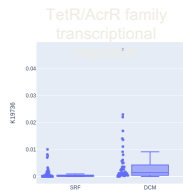
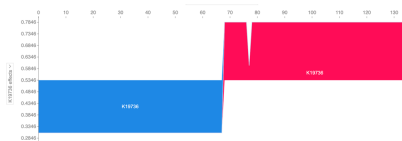
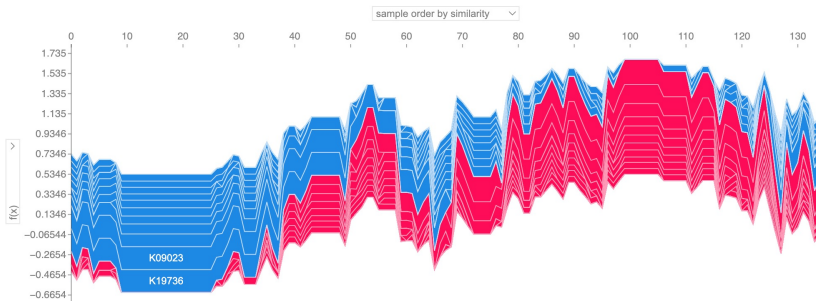
- Joint work with Nahuel Pilquinao, José Vásques, Luis Martí, Nayat Sánchez-Pi.

Inria

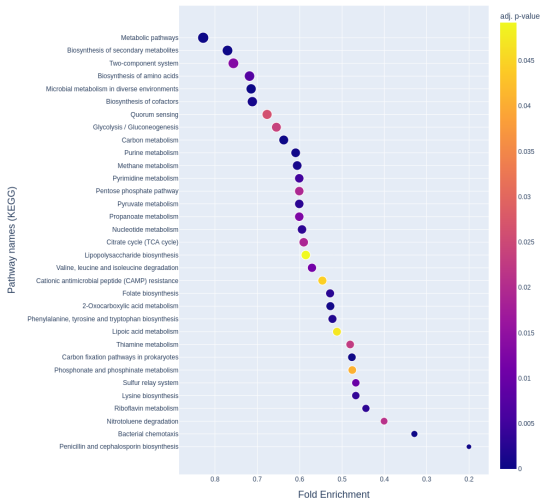
Thank you! Obrigado! Merci ! ¡Gracias!

<https://inria.cl>

(Metagenomic) Layers Characterization: SRF vs DCM (5 vs 55 m depth)



(Metagenomic) Layers Characterization: SRF vs DCM (5 vs 55 m depth)



(Metagenomic) Layers Characterization: SRF vs MES (5 vs 500 m depth)

