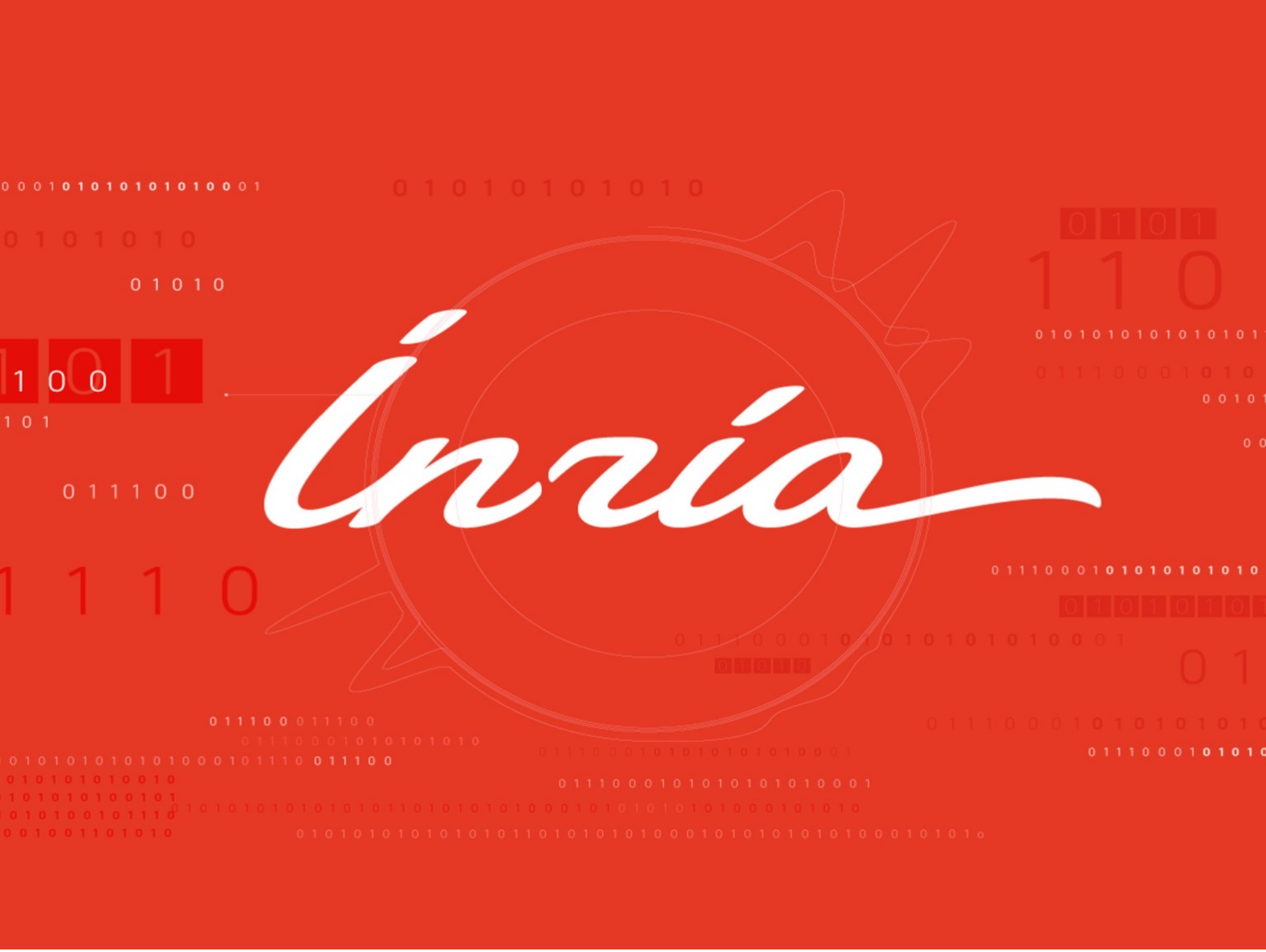

inria

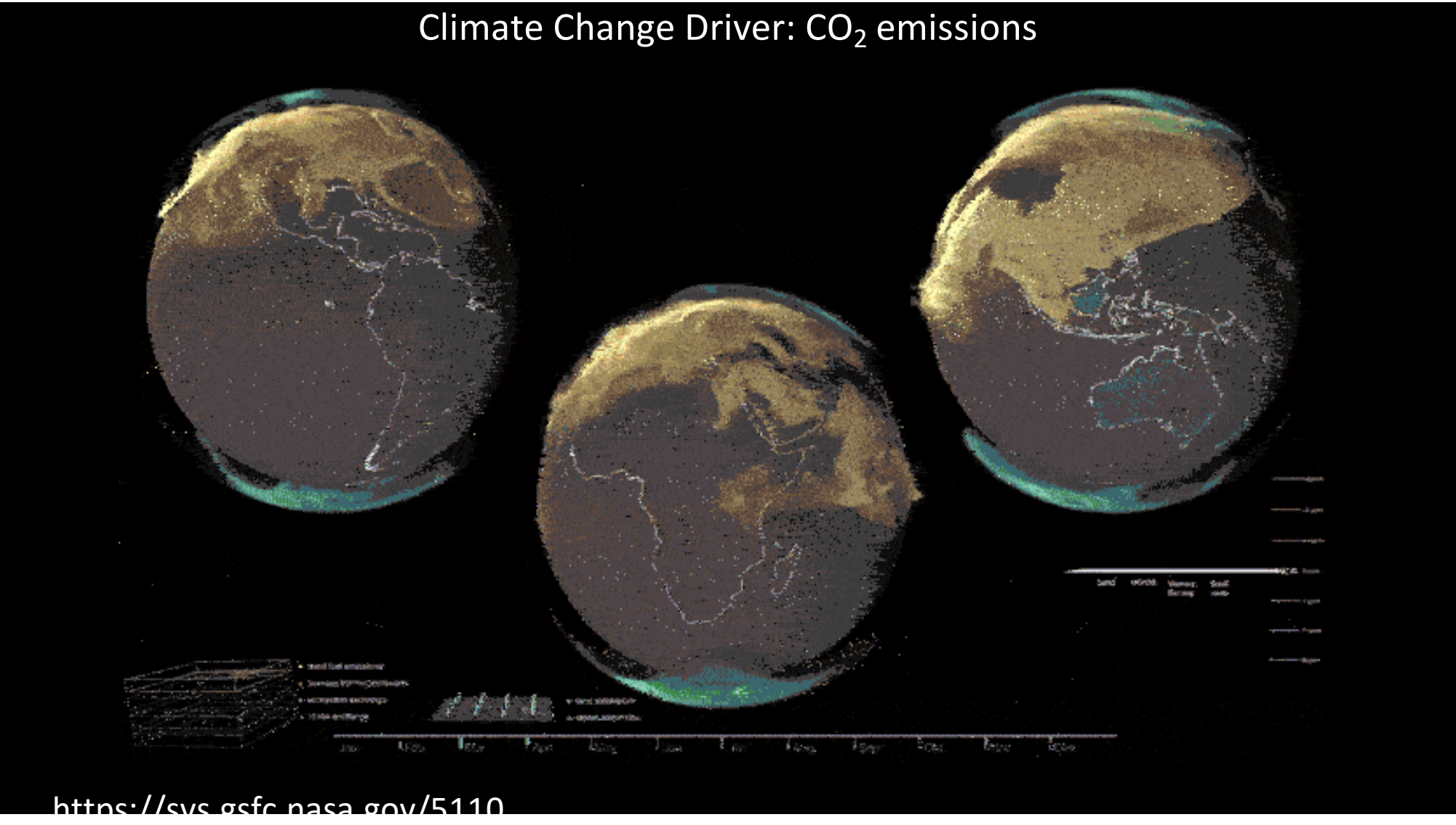
Modeling global plankton communities via multiomics and ML approaches

Luis Valenzuela
Inria Chile Research Center, Las Condes, Chile.
luis.valenzuela@inria.cl
https://inria.cl

Réunion Annuel OcéanIA
23 February 2023

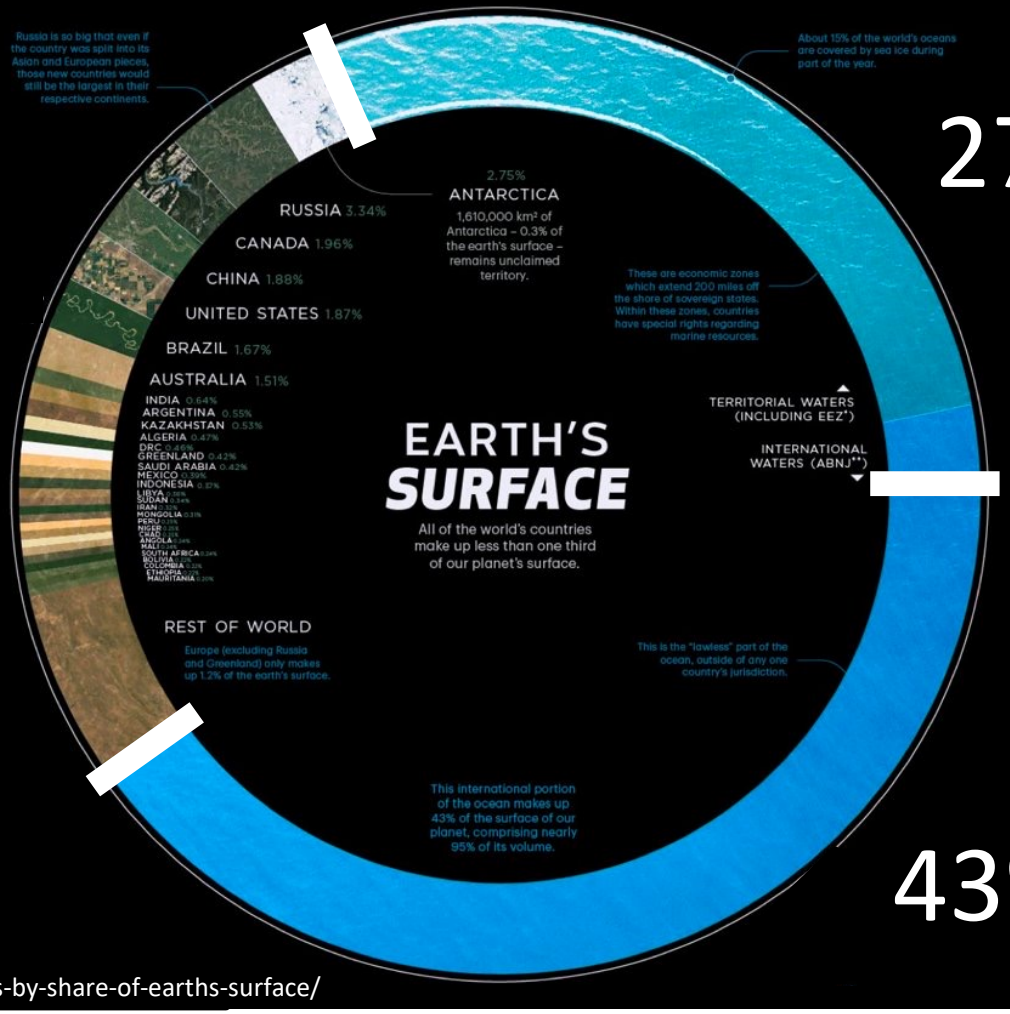
[1] Joint work *many others*.

Climate Change Driver: CO_2 emissions



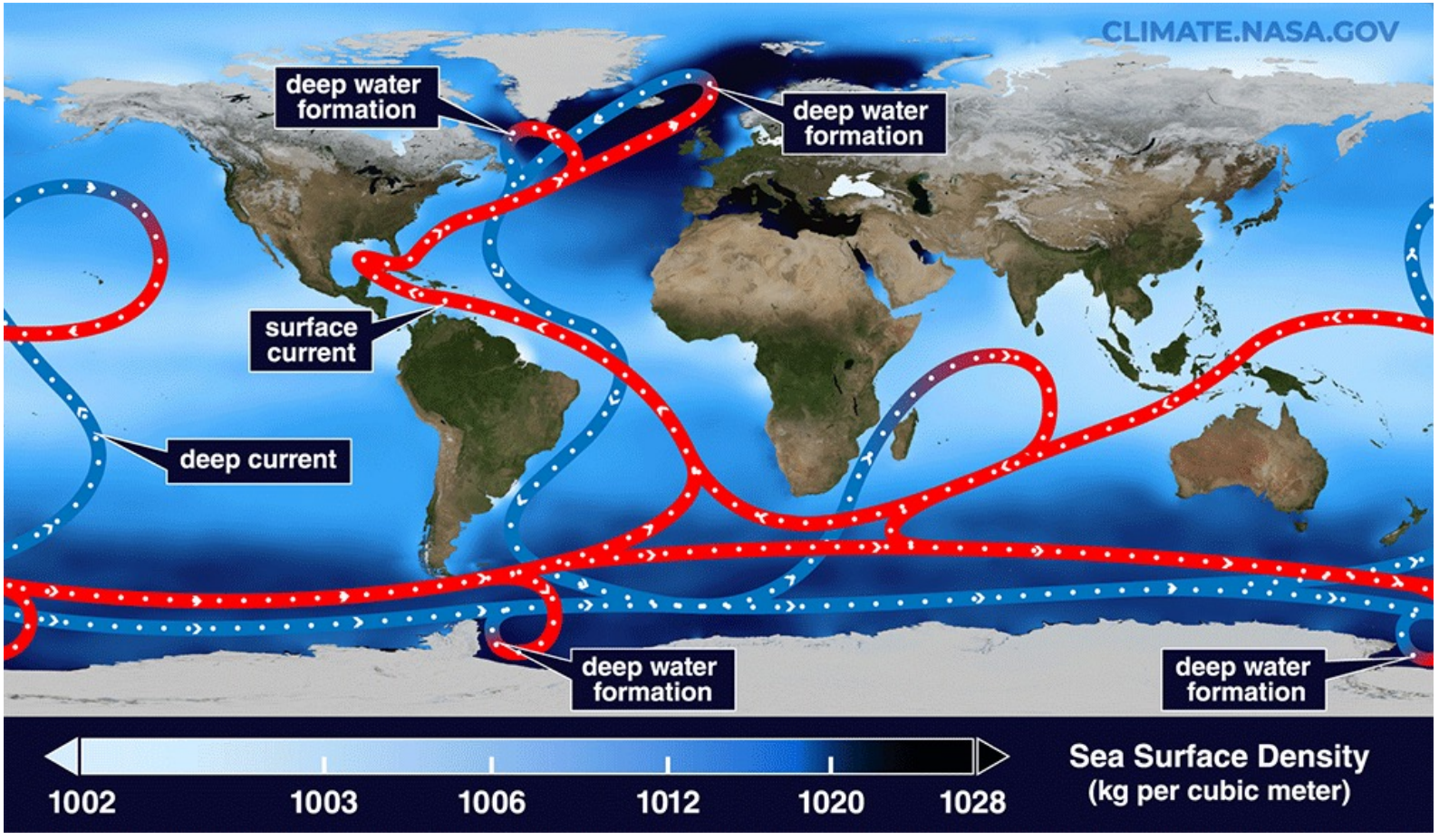
https://svs.gsfc.nasa.gov/5110

Climate Change Mitigator: Planet Ocean

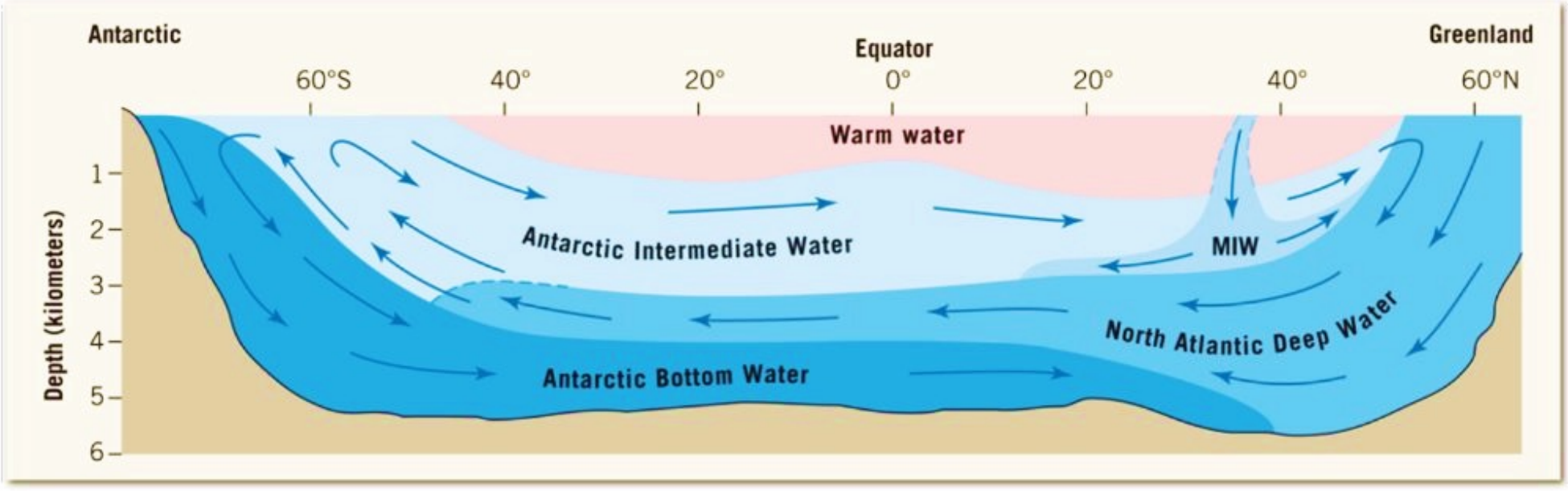


https://www.visualcapitalist.com/countries-by-share-of-earths-surface/

Thermohaline circulation: great ocean conveyor belt



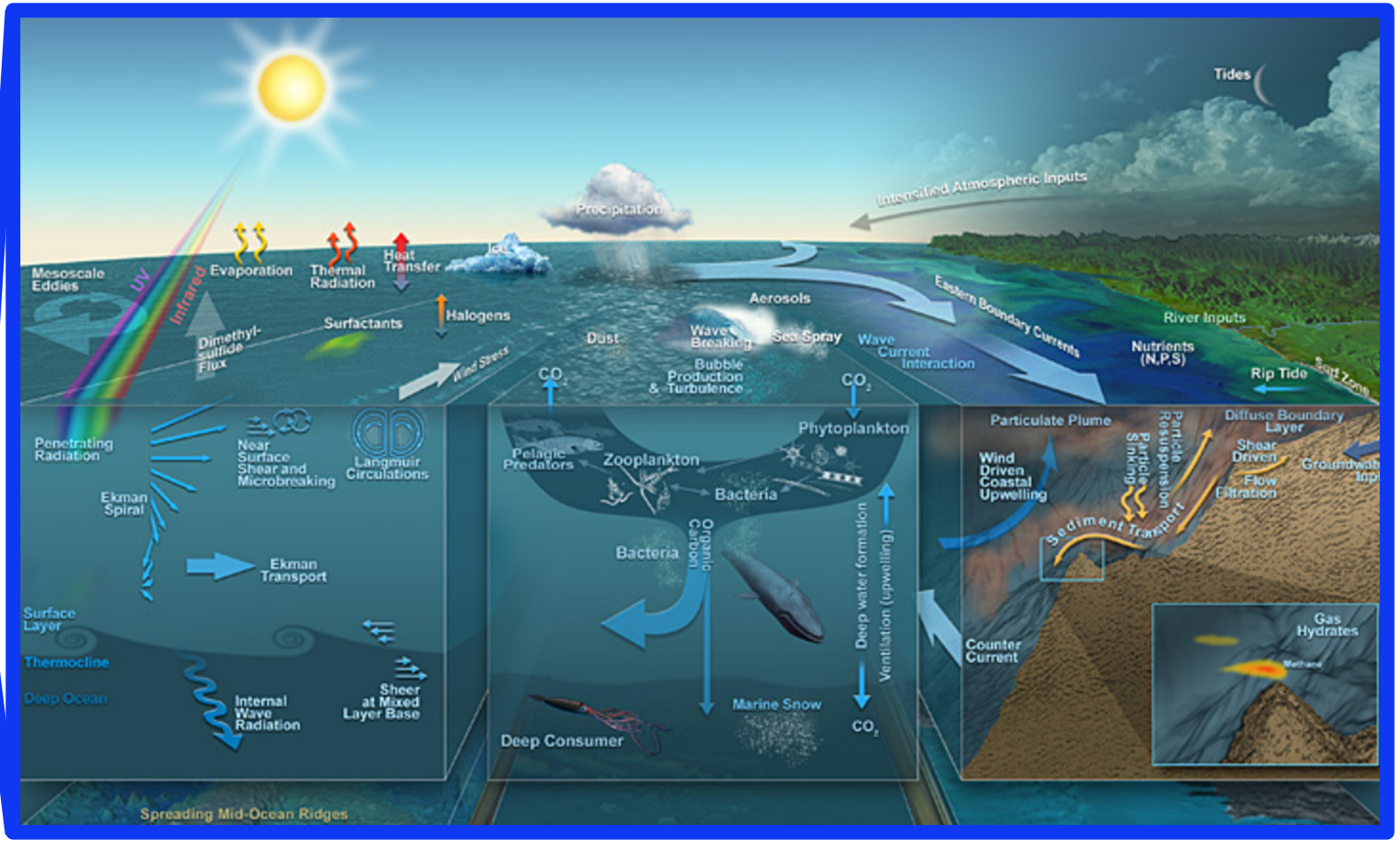
Cross section of the deep circulation in the Atlantic Ocean



Characterization of these water masses:

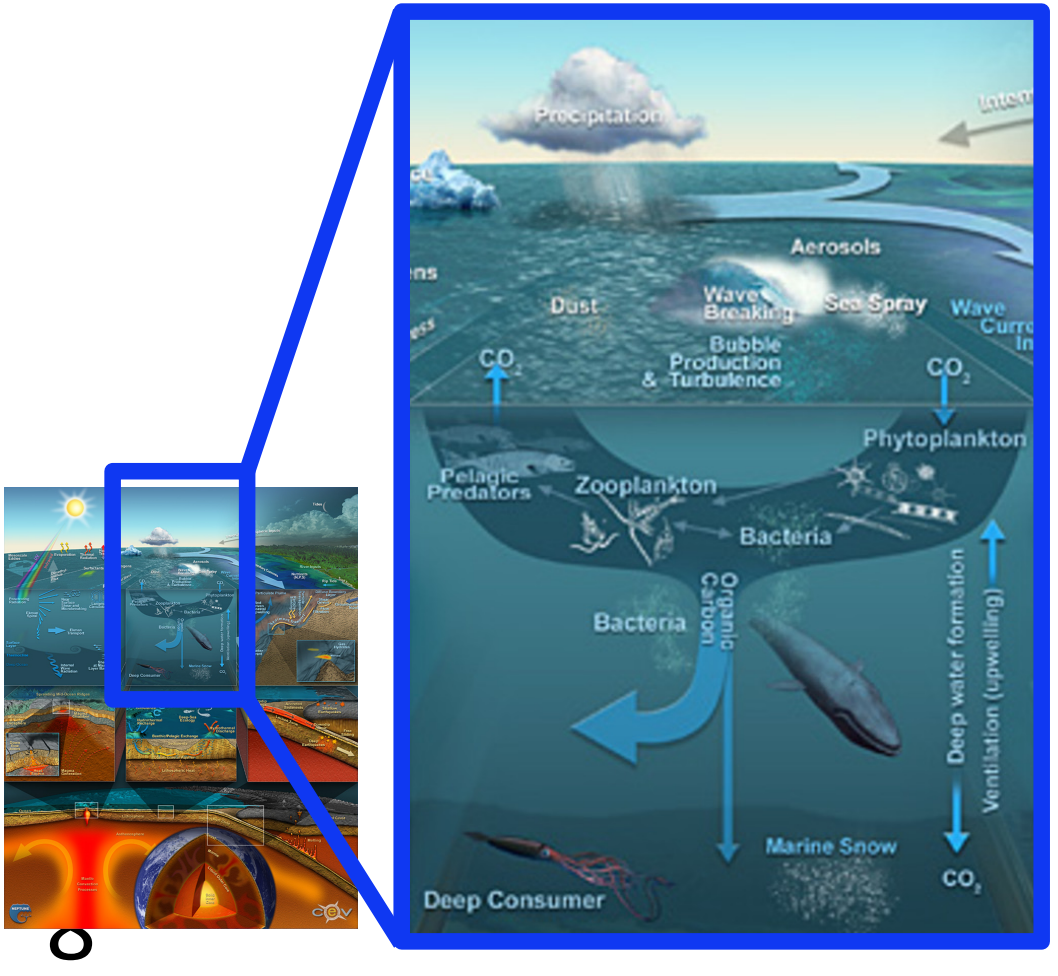
- Temperature-Salinity diagrams
- Isotopes (H, C, O)
- Bioinformatics

Bio-Geo-Physicochemical Oceanography: Marine Ecosystems



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

Key Invisible Microbiome World & Biological Carbon Pump

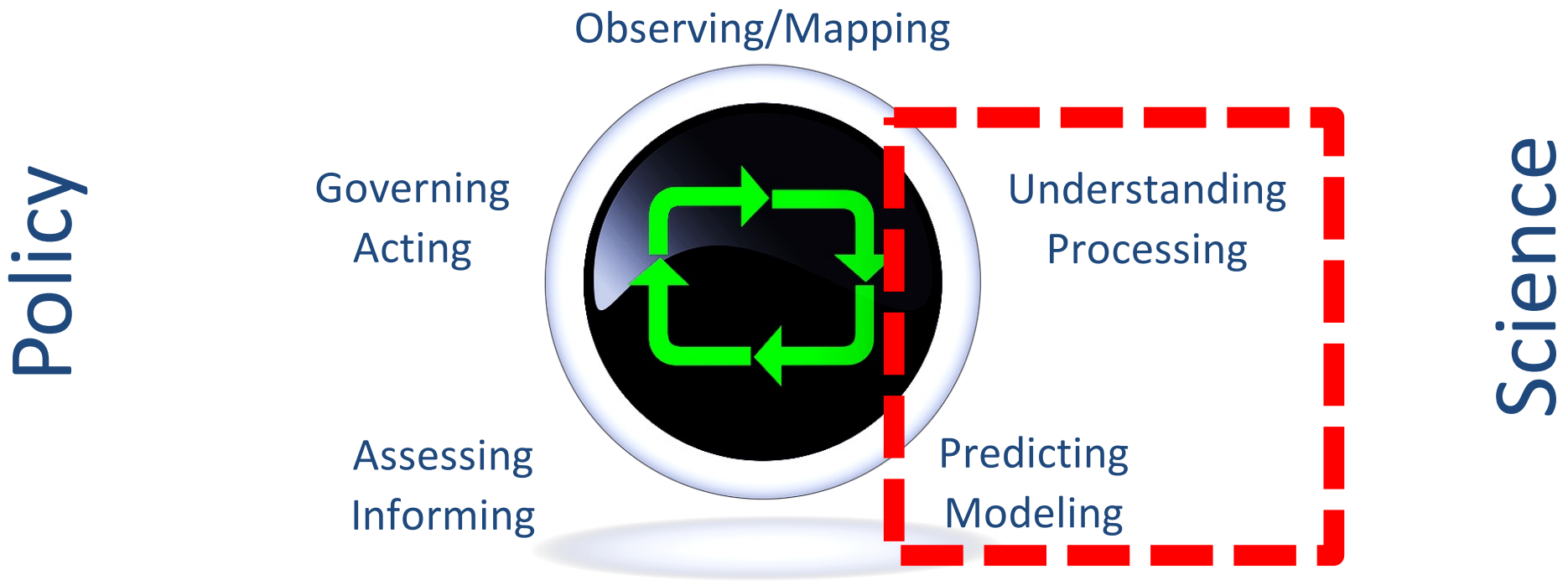


80% of marine life is made up of microorganisms

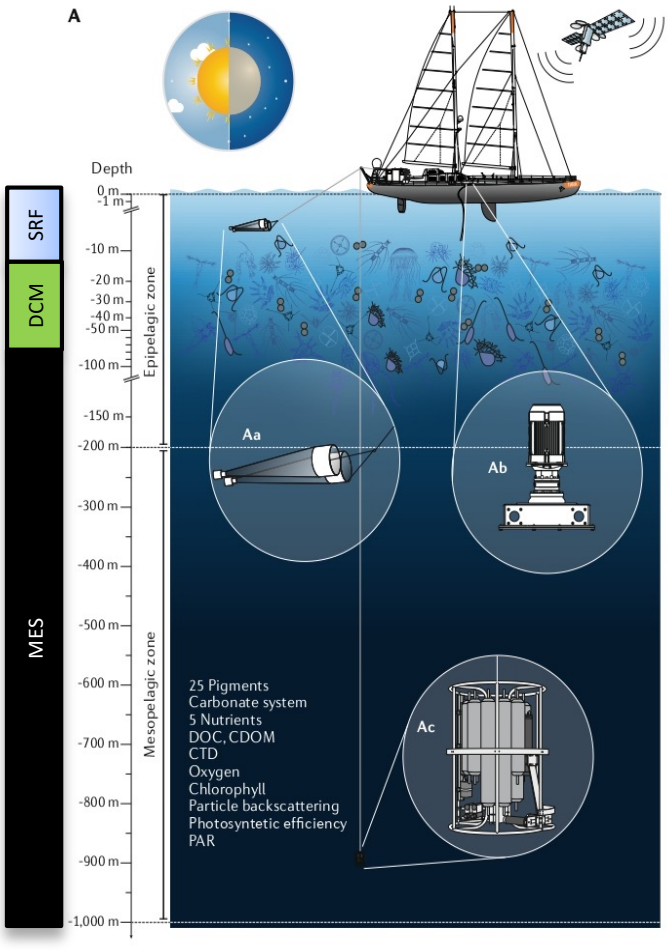
50% of the oxygen produced each day is provided by marine microorganisms

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

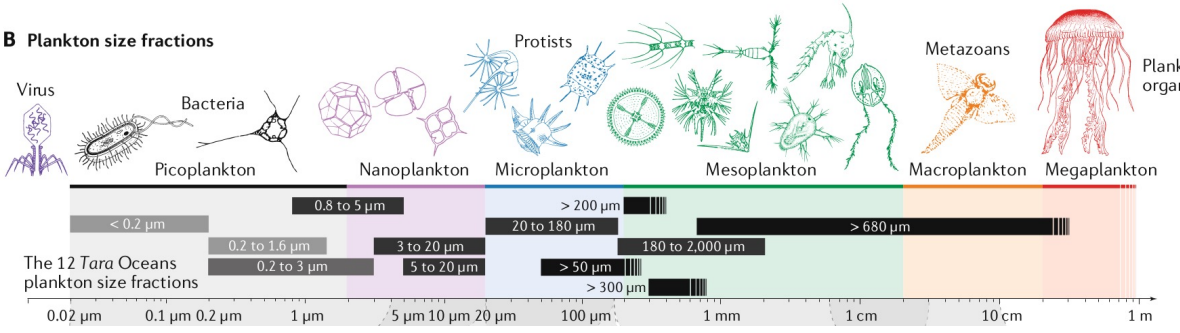
Management of marine ecosystems

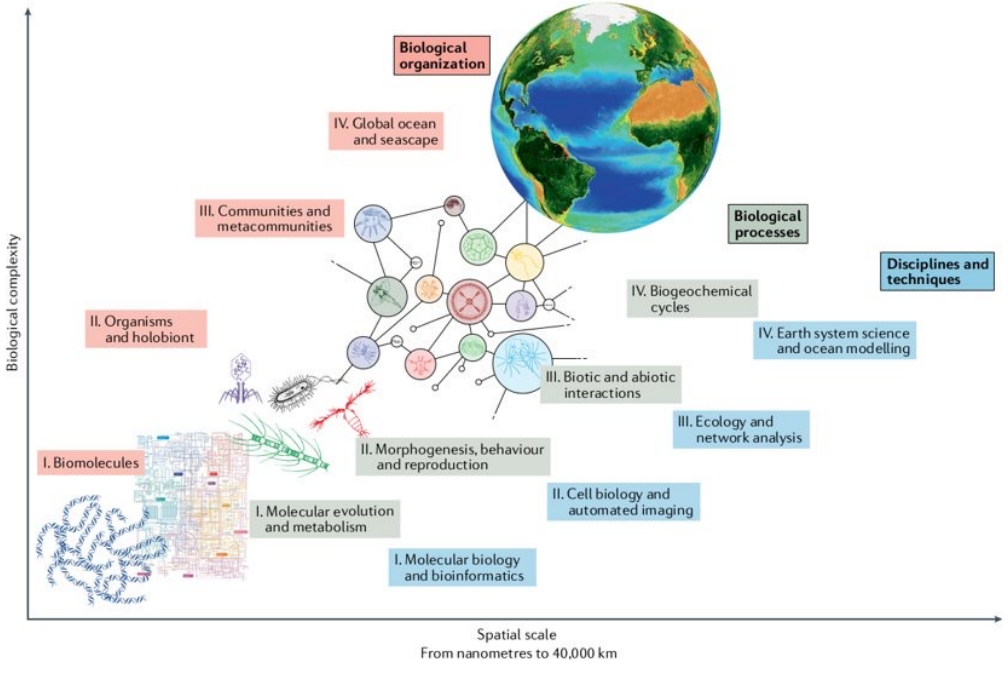


Ambitious goal: towards global ocean ecosystems biology

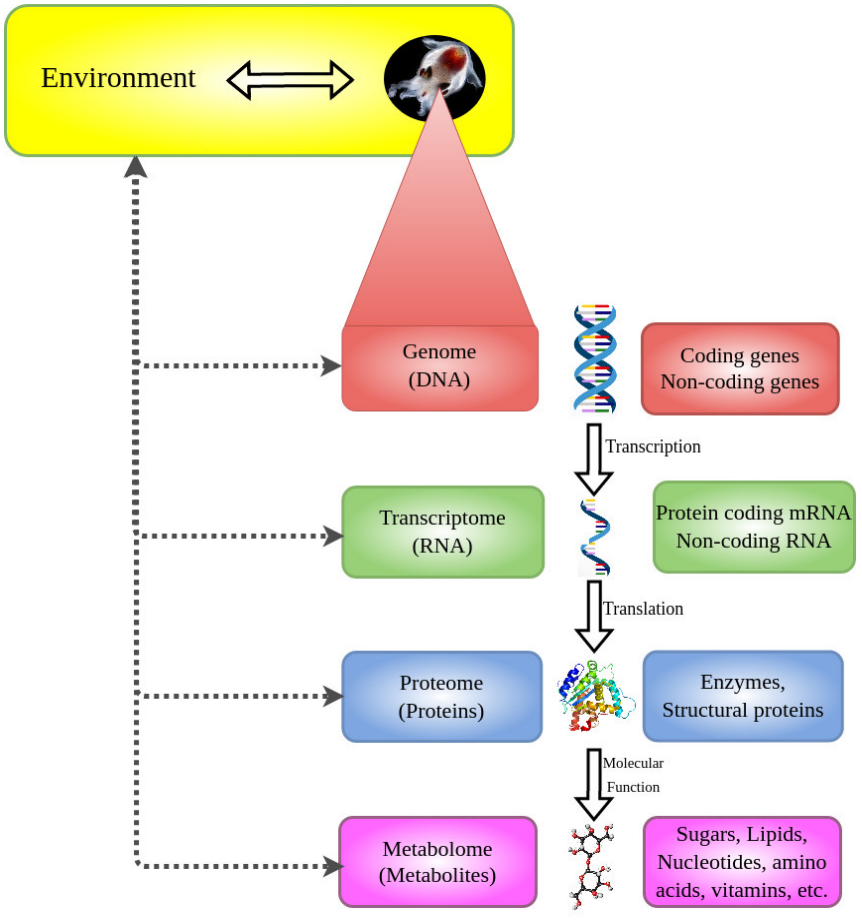


Sunagawa et al. (2020)

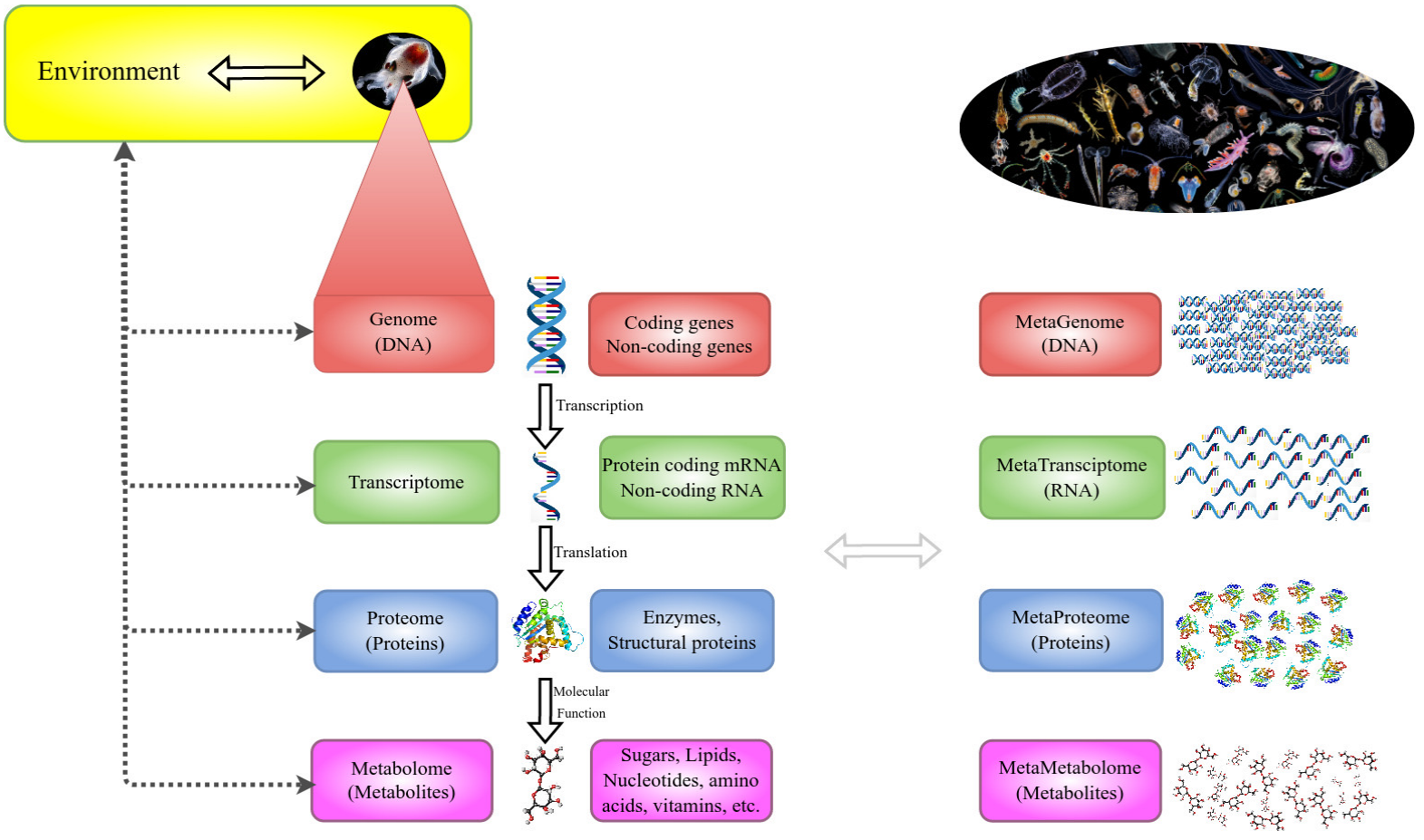




Global plankton communities: multi-omics data



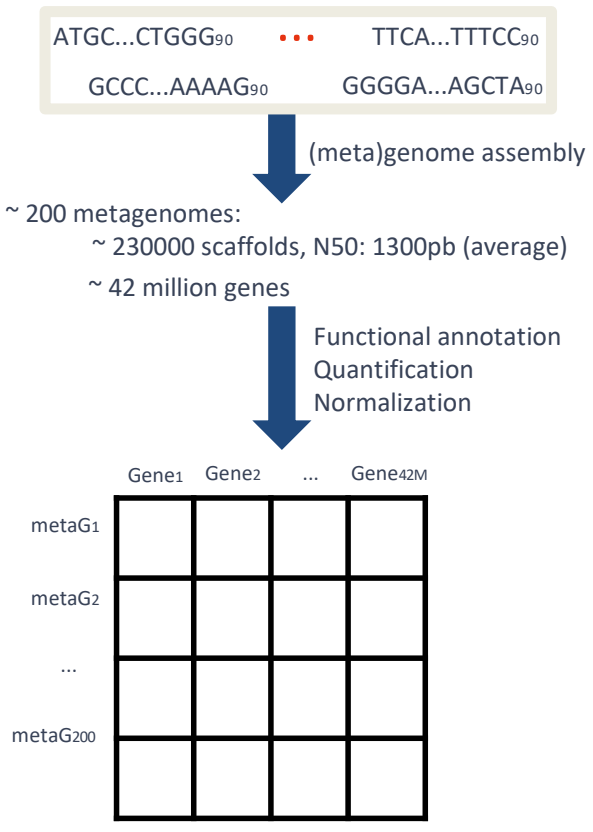
Global plankton communities: multi-omics data

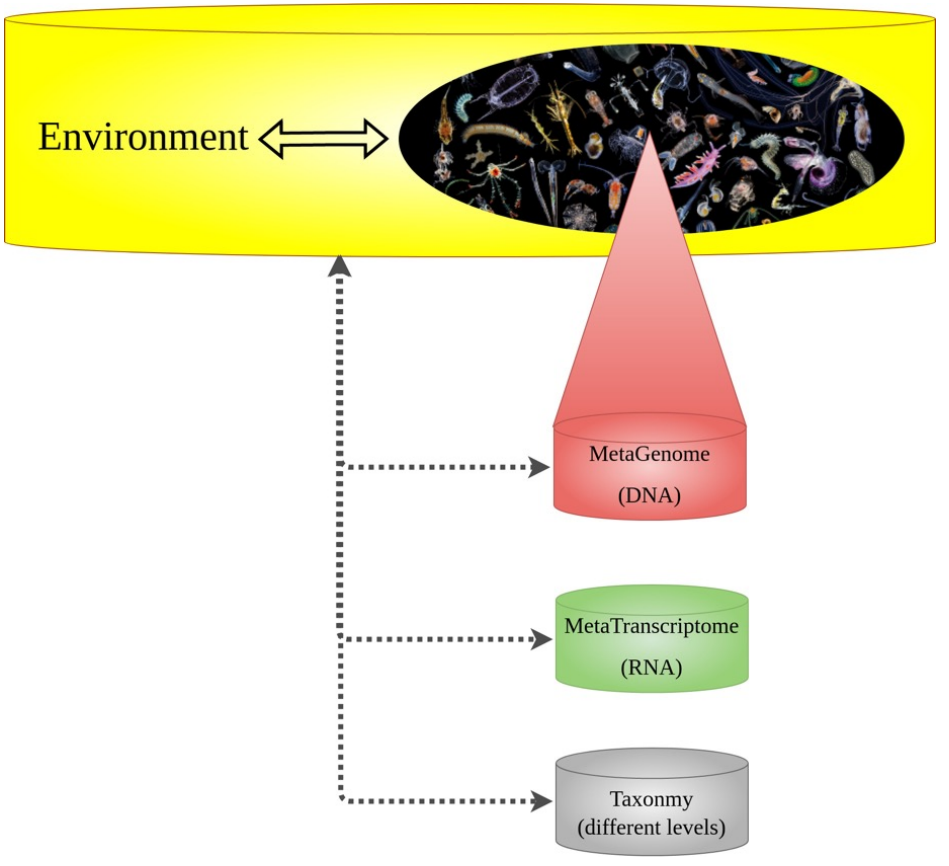


Ocean Microbial Reference Gene Catalog v2

Metagenomic dataset only:

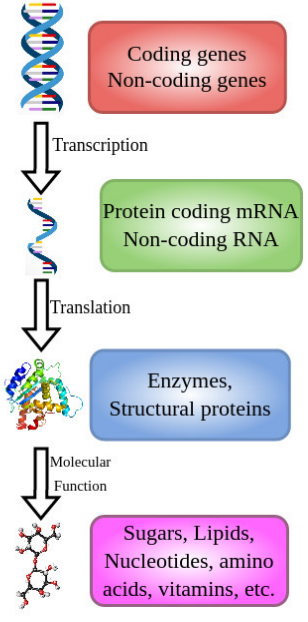
~ 57.000 million reads (90 pb ±2.6pb)

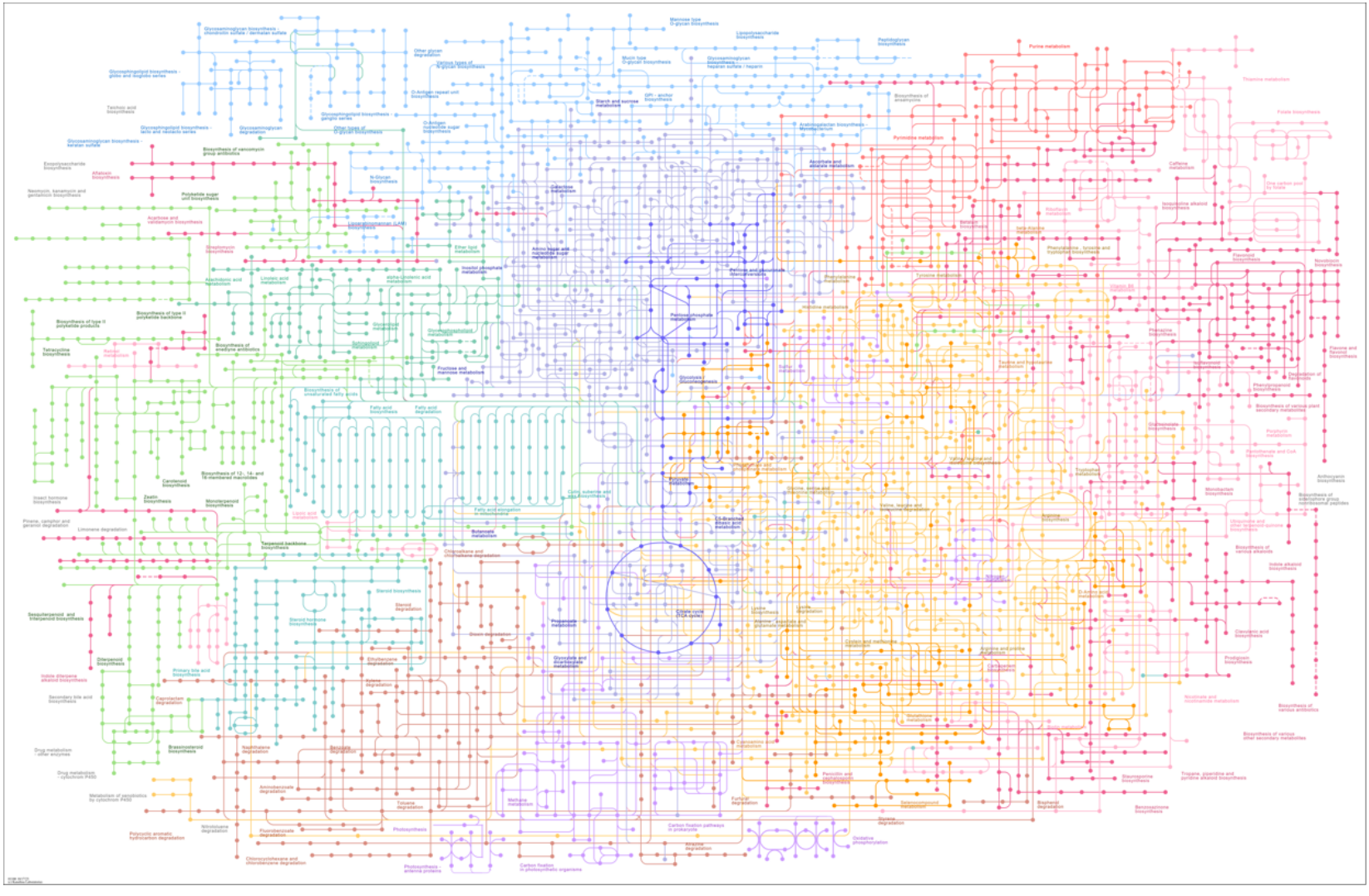




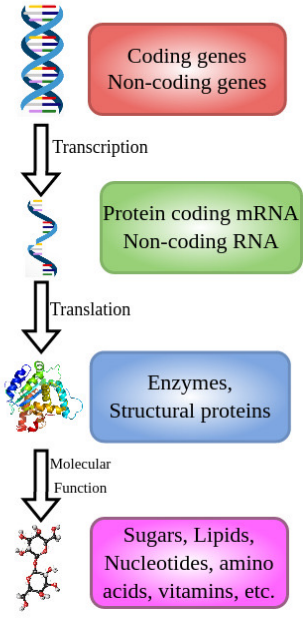
Metabolic pathways – Reference pathway

Edges: enzymes
Dots: metabolites

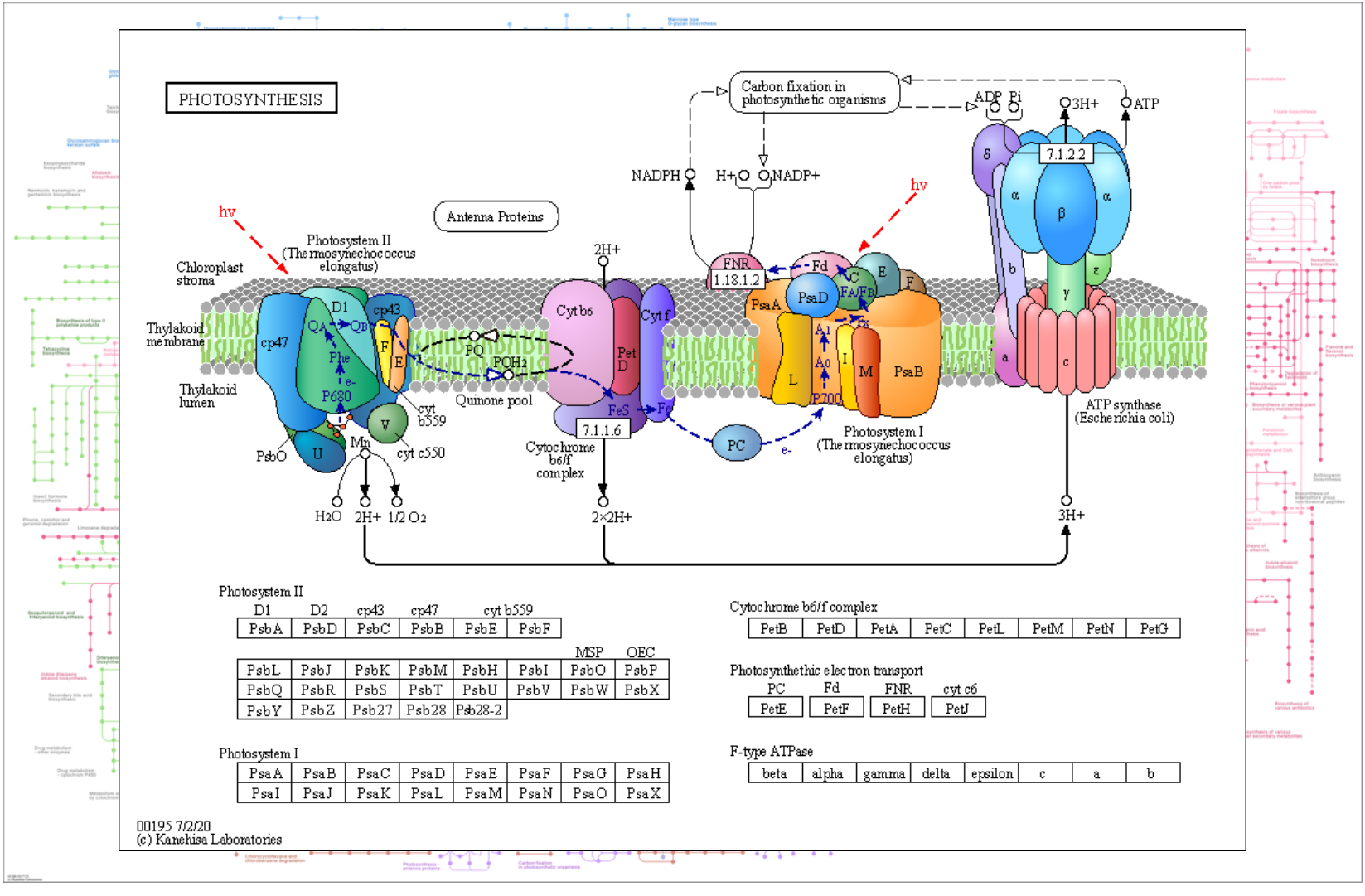




Metabolic pathways – Reference pathway



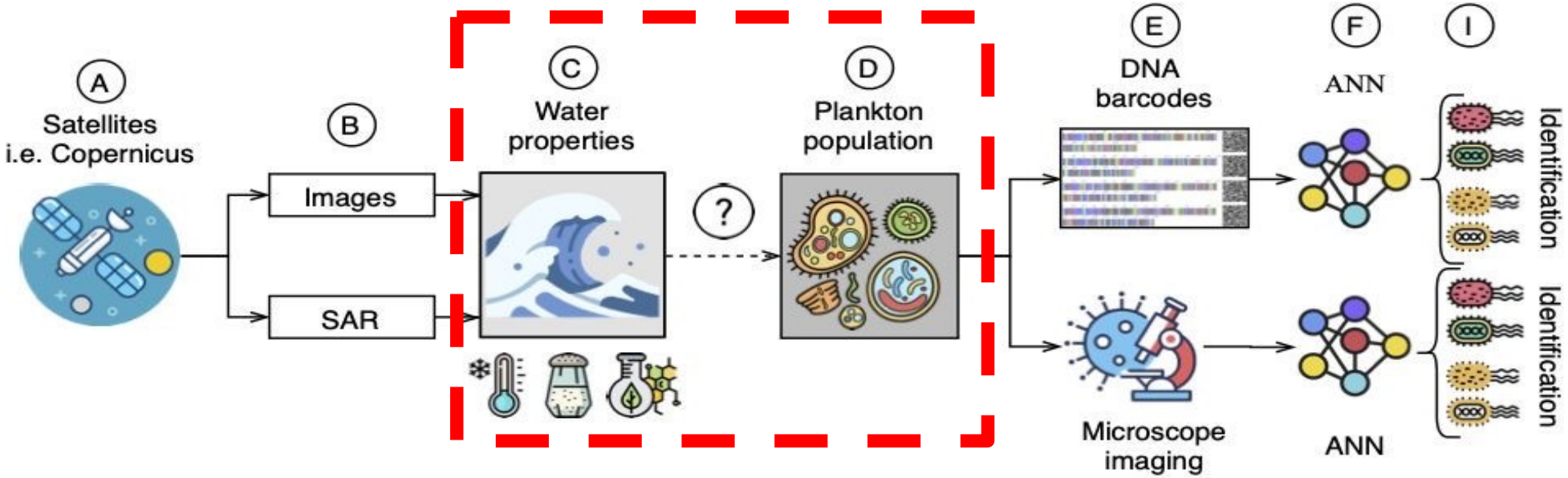
KEGG pathway example: Photosynthesis



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

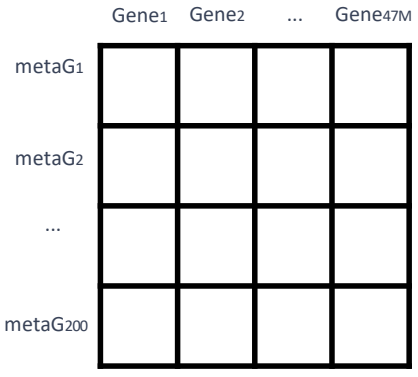
Mapping C -> D

Image by Luis Martí

Overview: Global patterns of Plankton Communities

Metagenomic composition

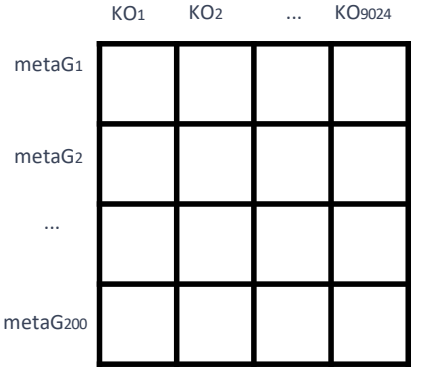
172 samples x 46.7M genes



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

Group abundances of genes with equal molecular function:

172 samples x 9024 molecular functions



Metatranscriptomic composition

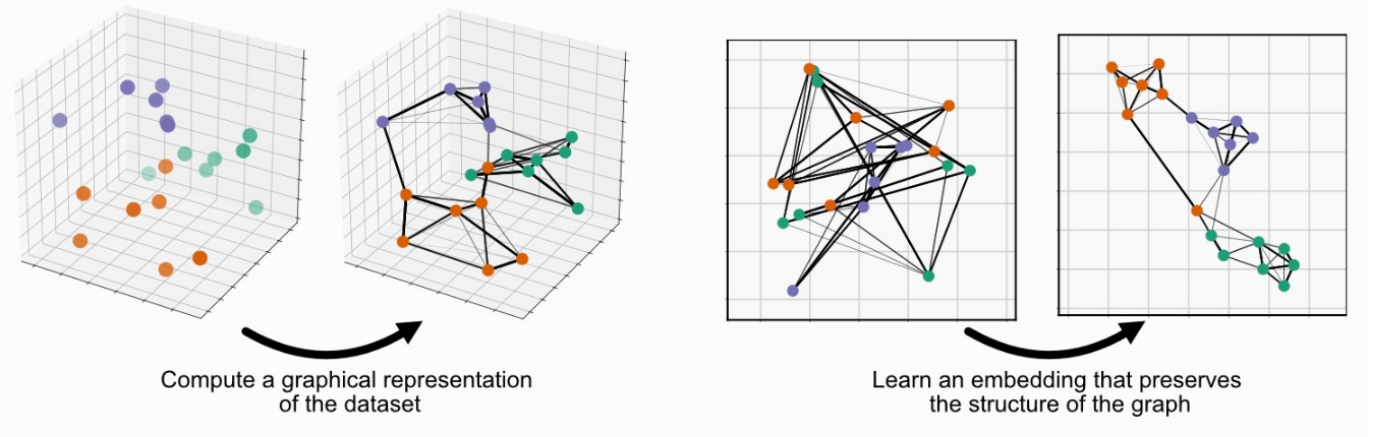
178 samples with similar number of columns.

Overview: Global patterns of Plankton Communities

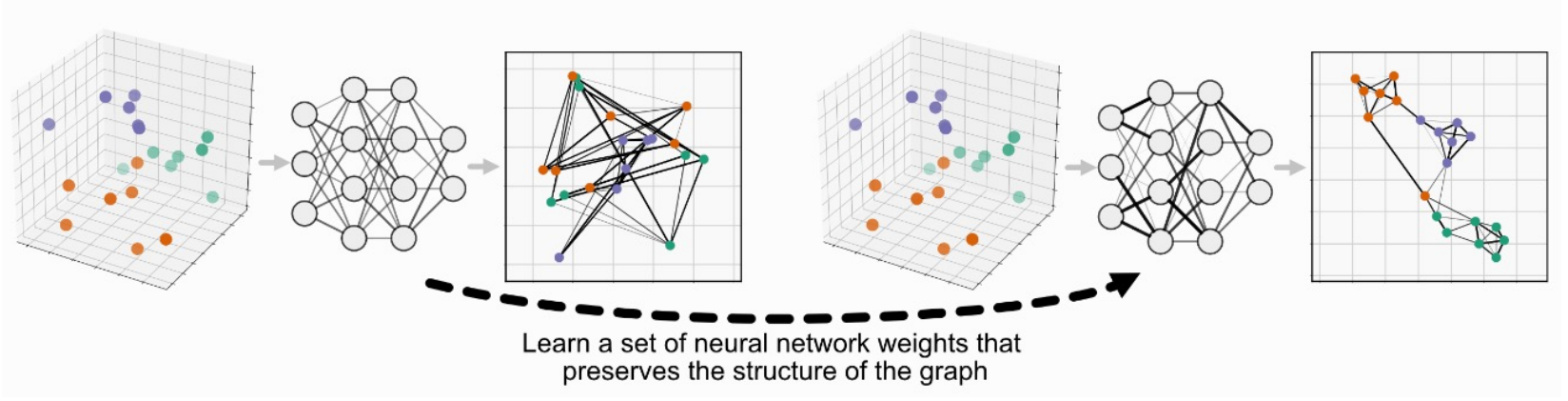
Dataset	Cardinality	Description
Environment	30	Measurements of environmental parameters, encompassing both physical and chemical attributes of water samples.
Genomic	9024	Abundances of molecular functions inferred from metagenomic assays employing the KEGG annotation database.
Transcriptomic	8935	Abundances 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

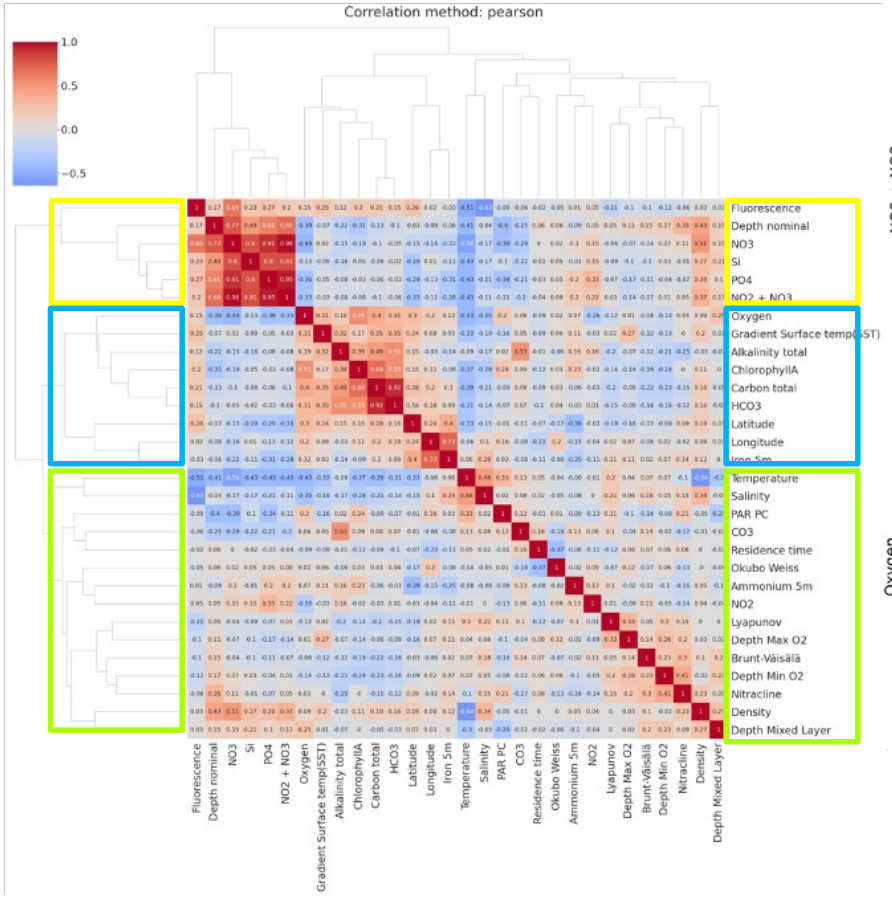


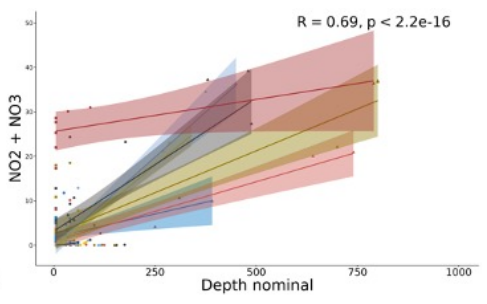
UMAP + Autoencoder model

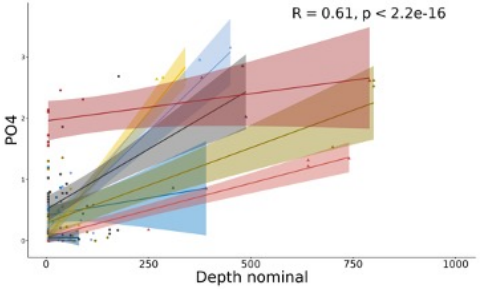


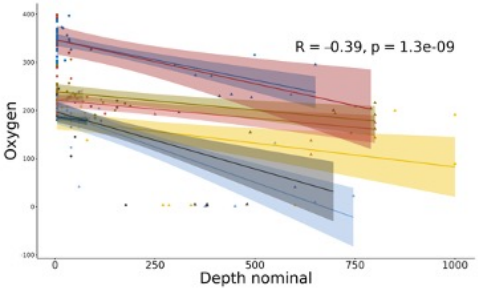
Overview: Global patterns of Plankton Communities

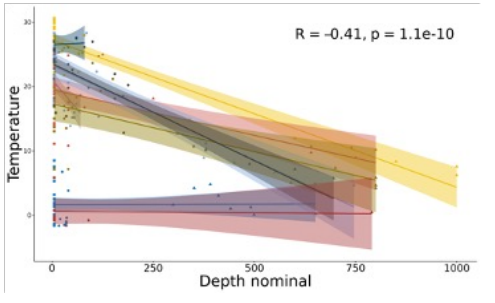
Environmental data exploration



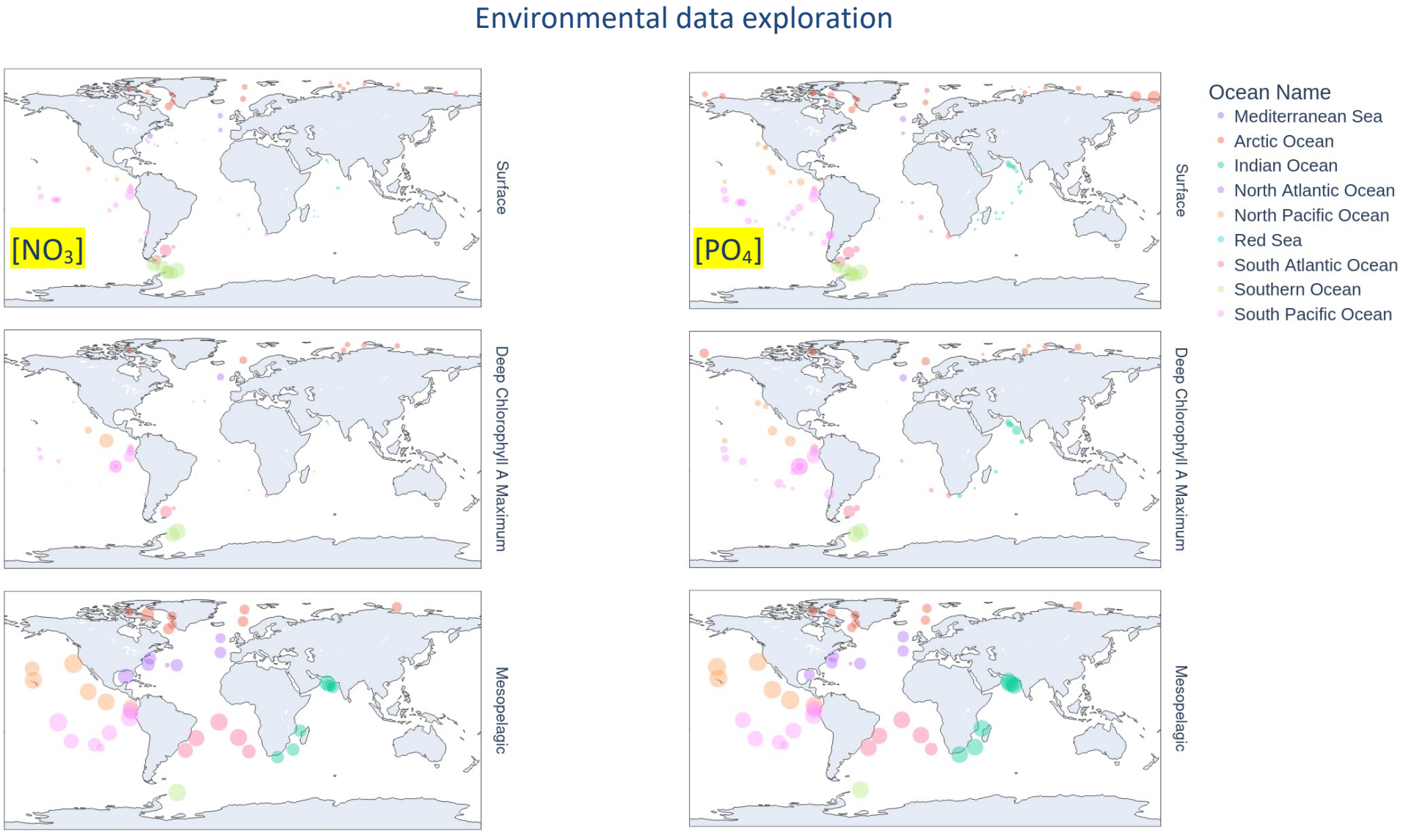




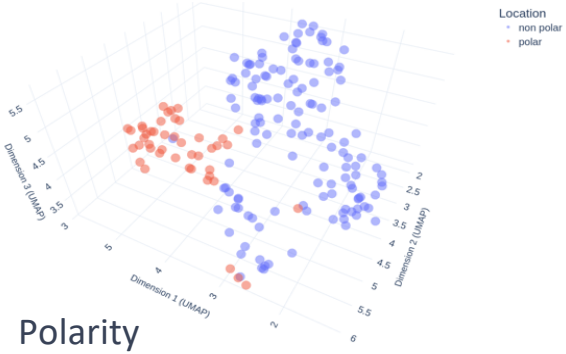




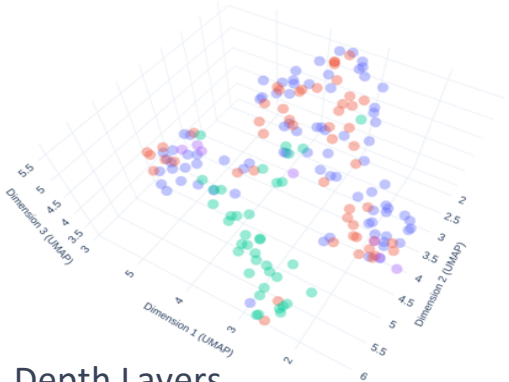
Overview: Global patterns of Plankton Communities



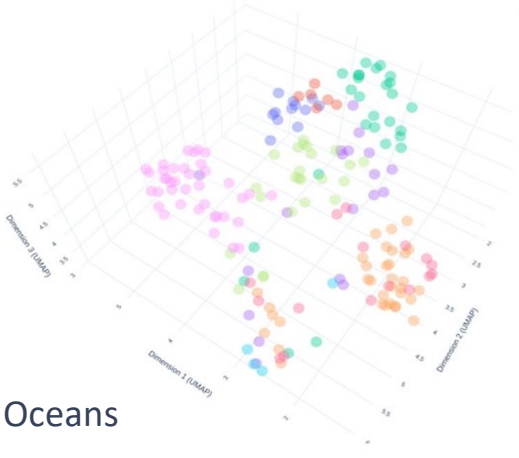
Overview: Global patterns of Plankton Communities



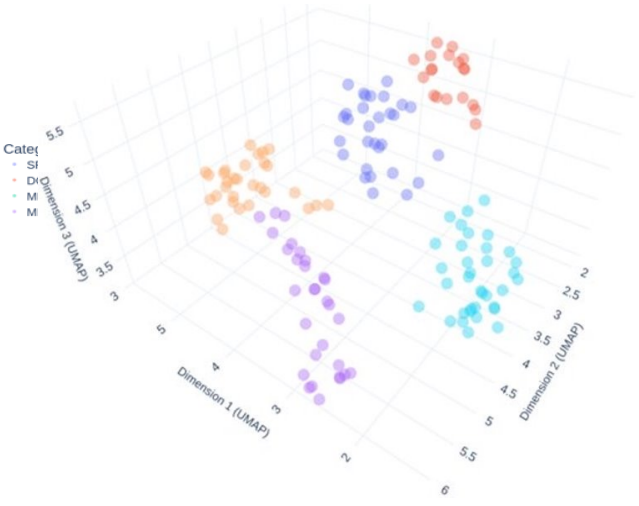
Polarity



Depth Layers



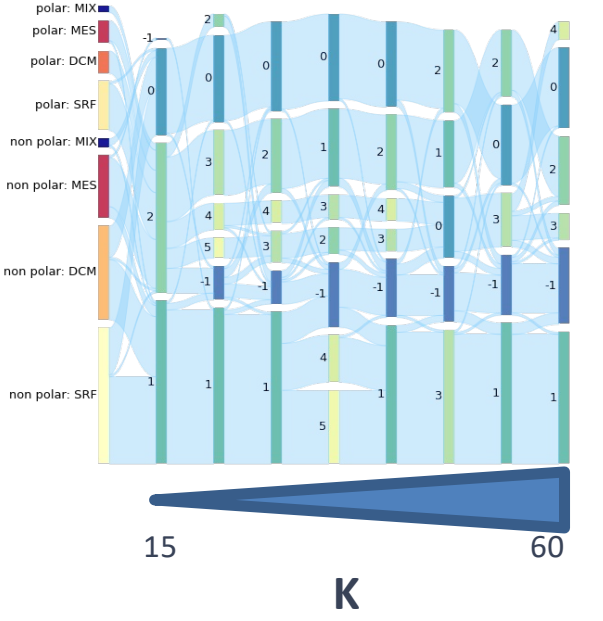
Oceans



Environmental data exploration: potential niches

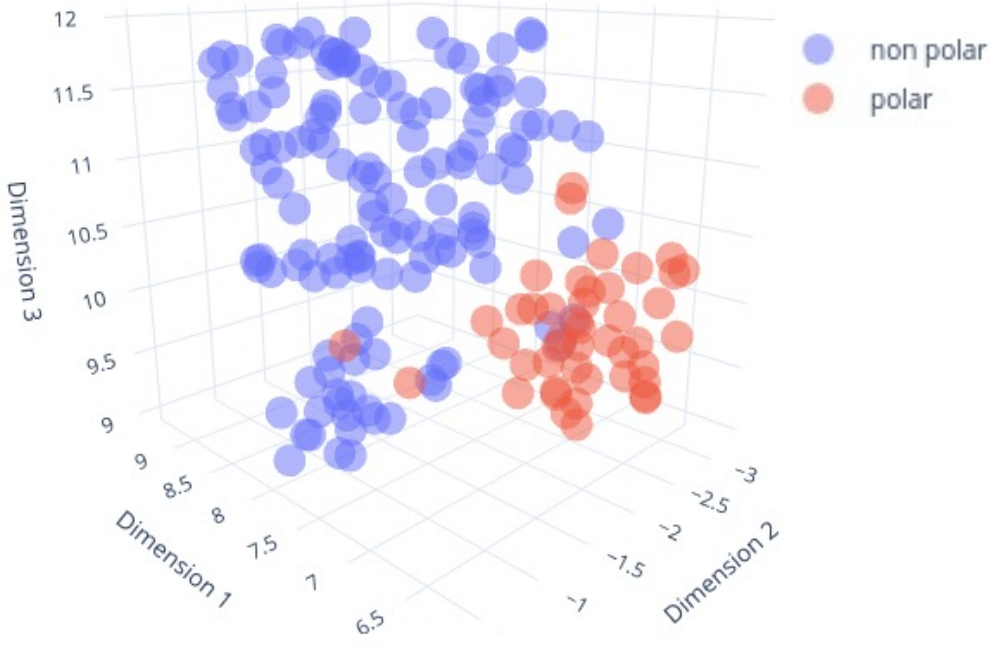
Local relationships

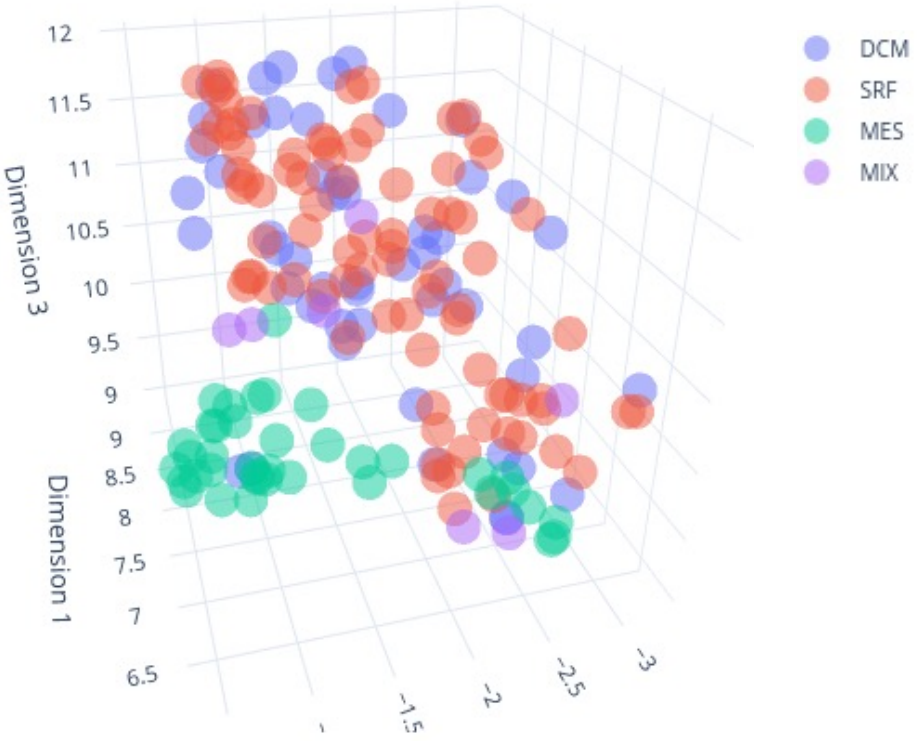
Global relationships



Overview: Global patterns of Plankton Communities

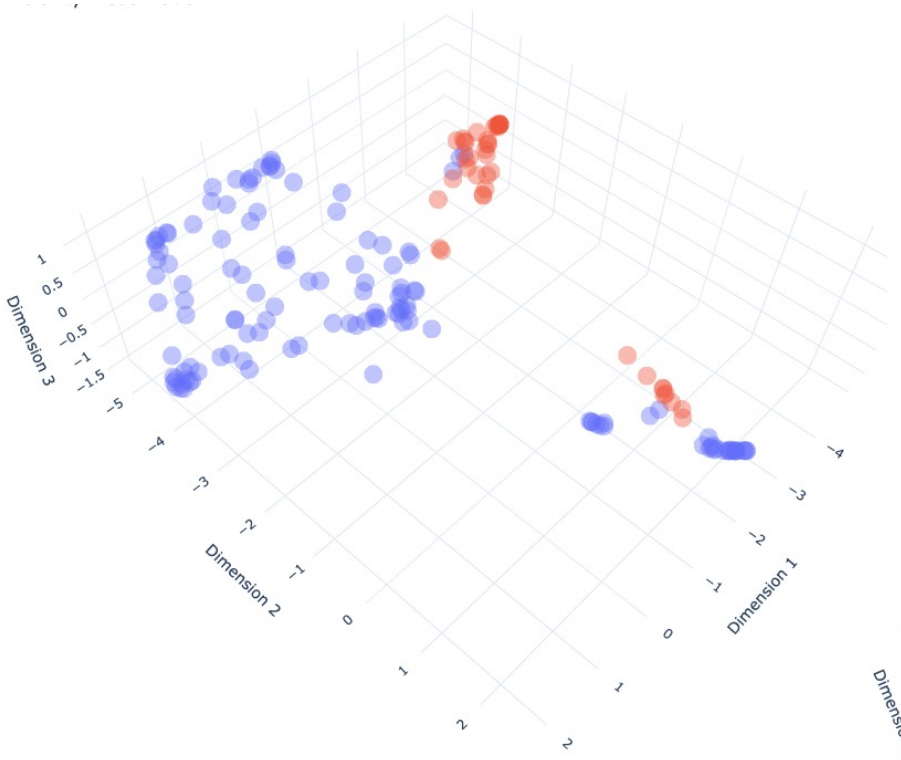
Input: metagenomic dataset (**47M genes**)

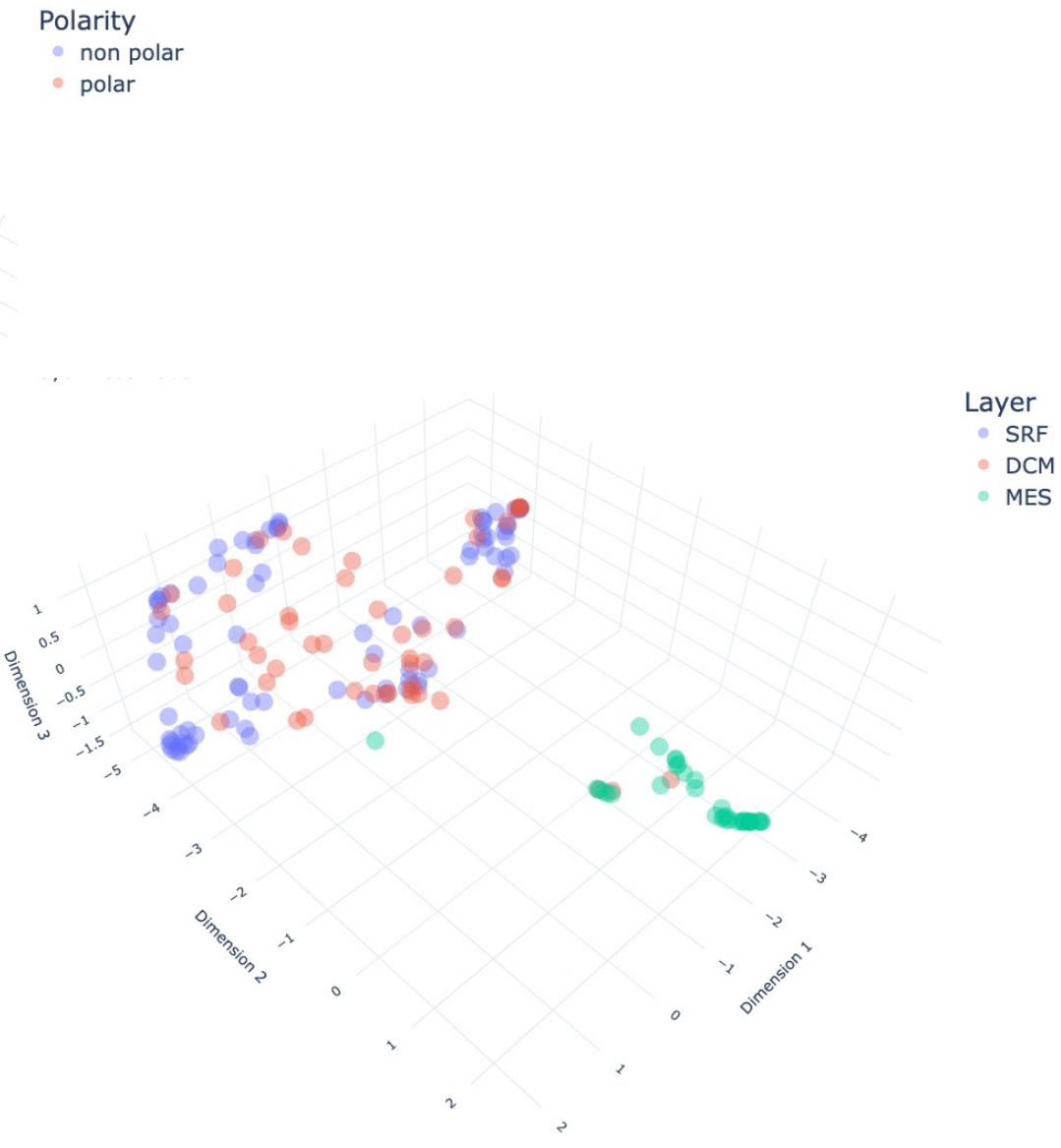




Overview: Global patterns of Plankton Communities

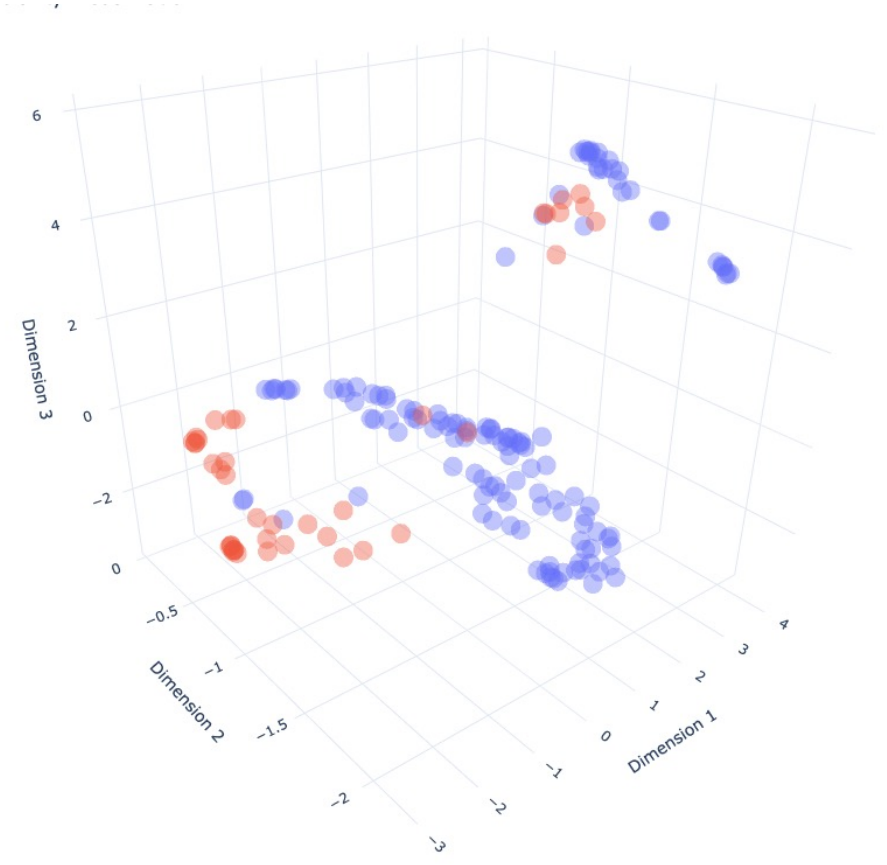
Input: metagenomic dataset (**9024 Molecular Functions**)

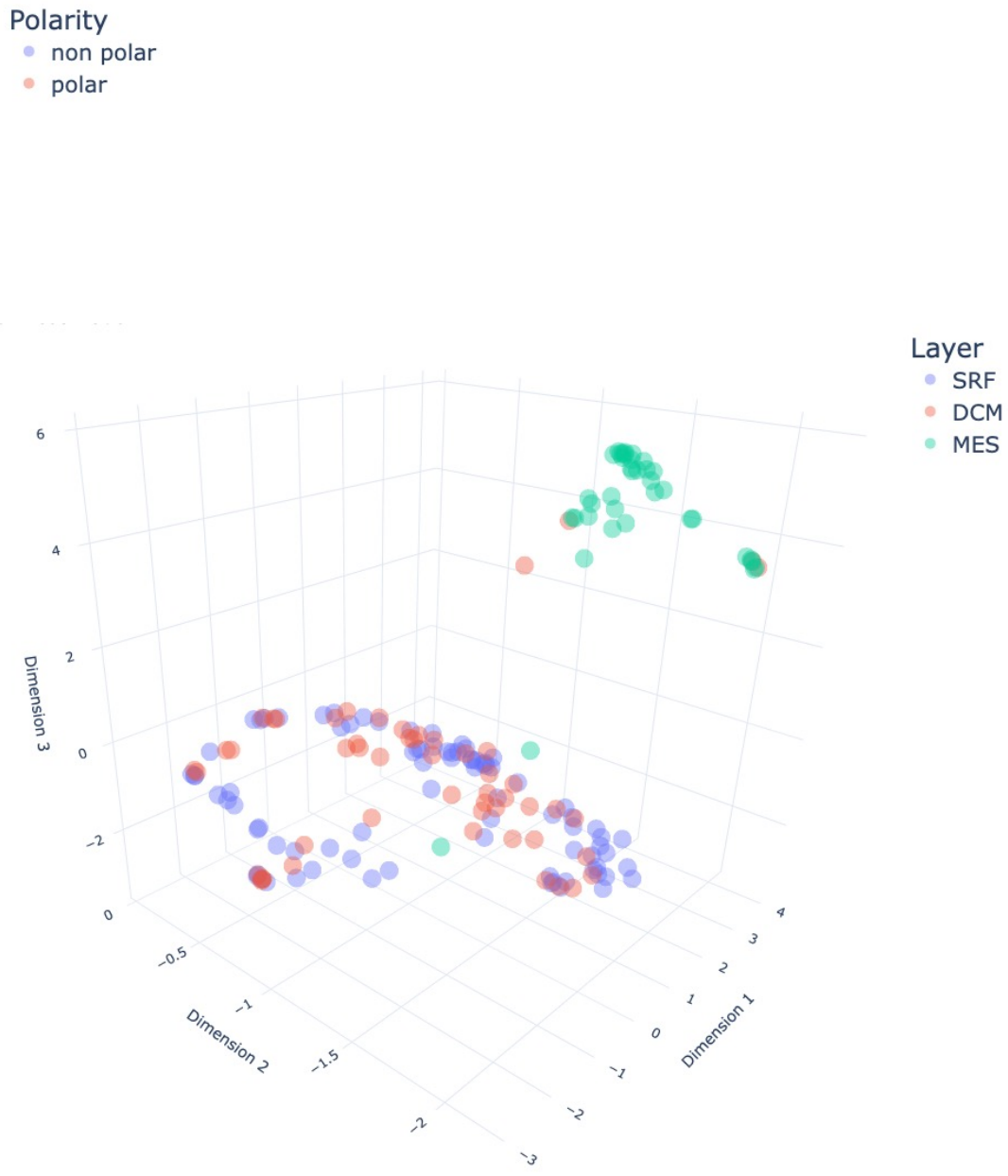




Overview: Global patterns of Plankton Communities

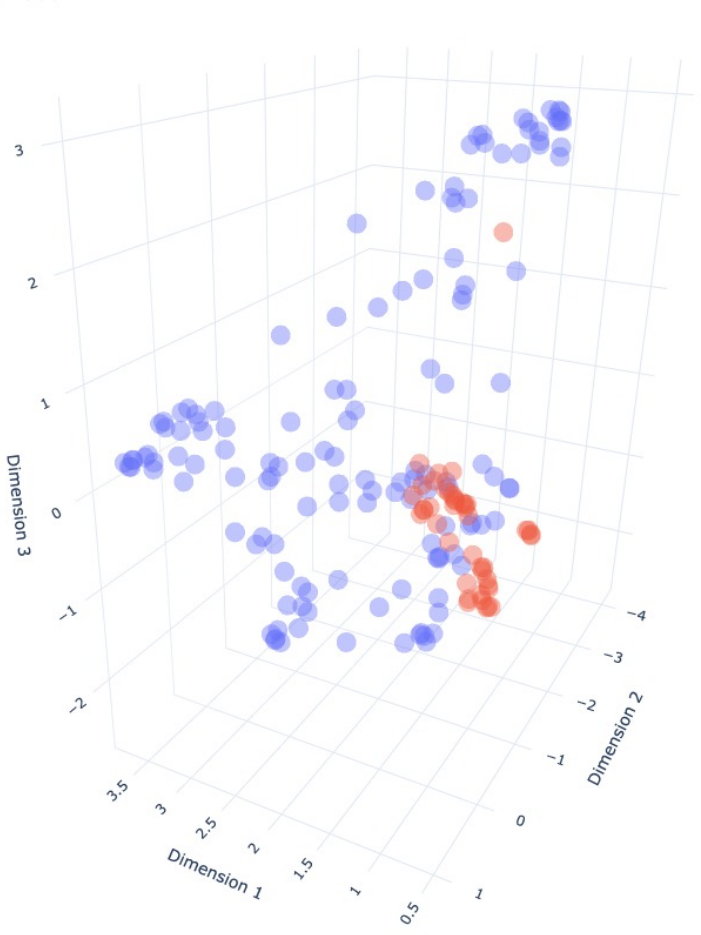
Input: metagenomic dataset (**453 pathways**)

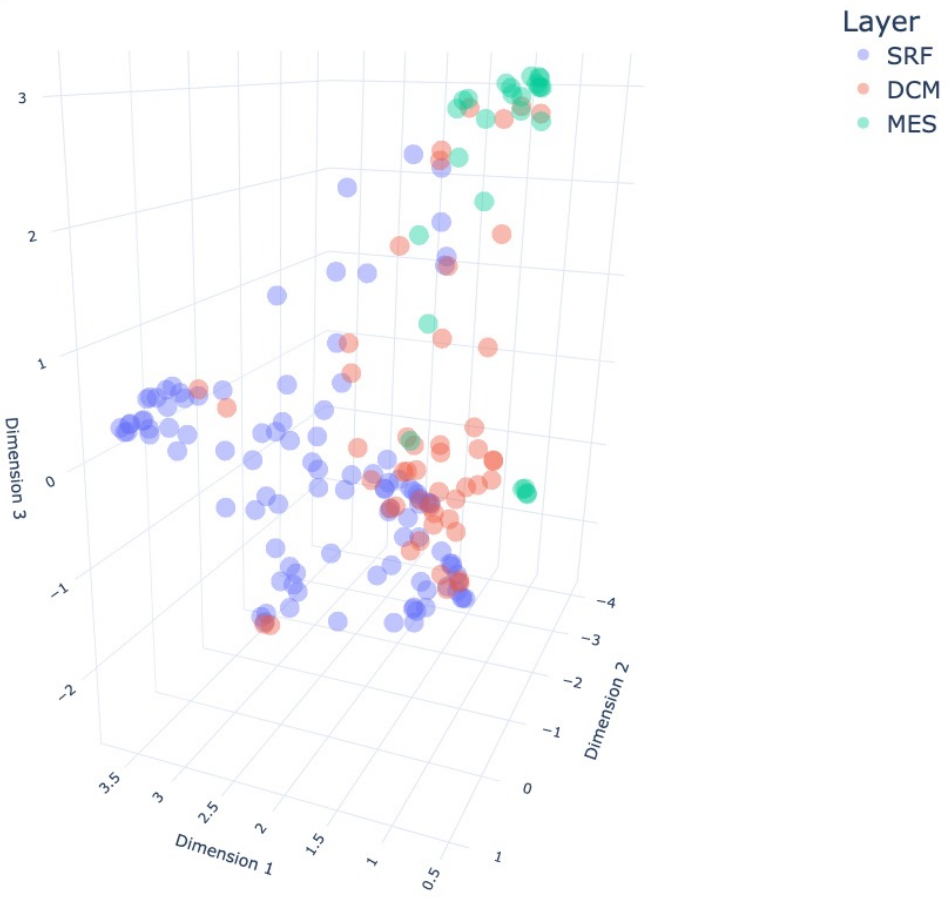




Overview: Global patterns of Plankton Communities

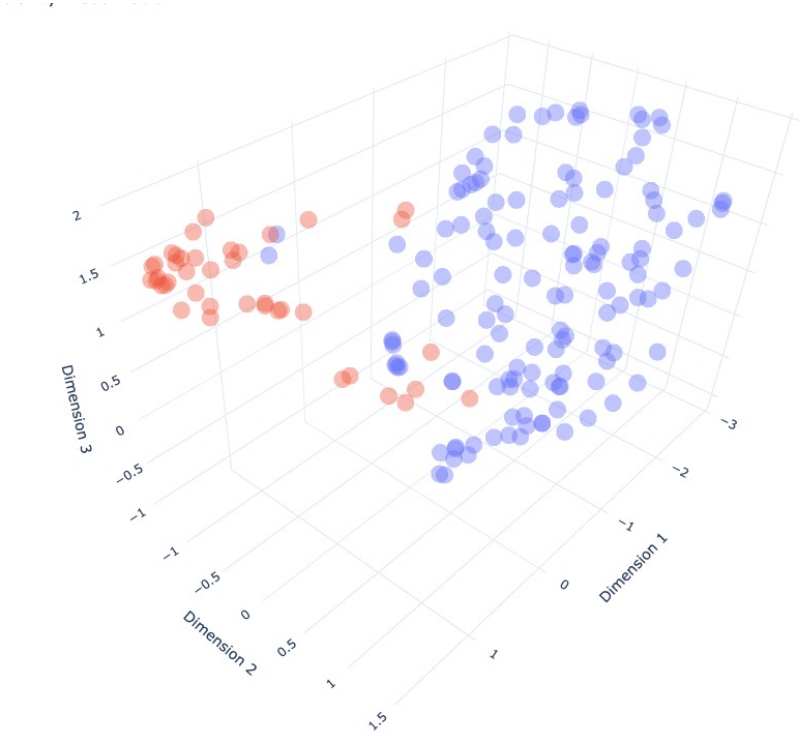
Input: metatranscriptomic dataset (**8935 Molecular Functions**)

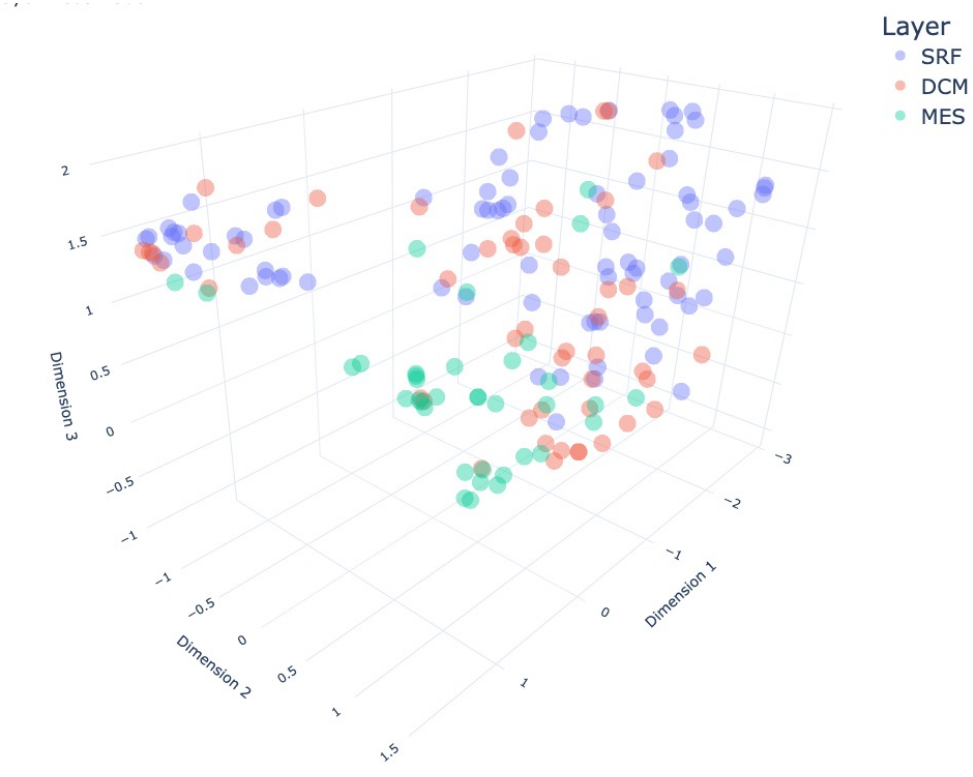




Overview: Global patterns of Plankton Communities

Input: metagenomic dataset (**2124 Genus level**)

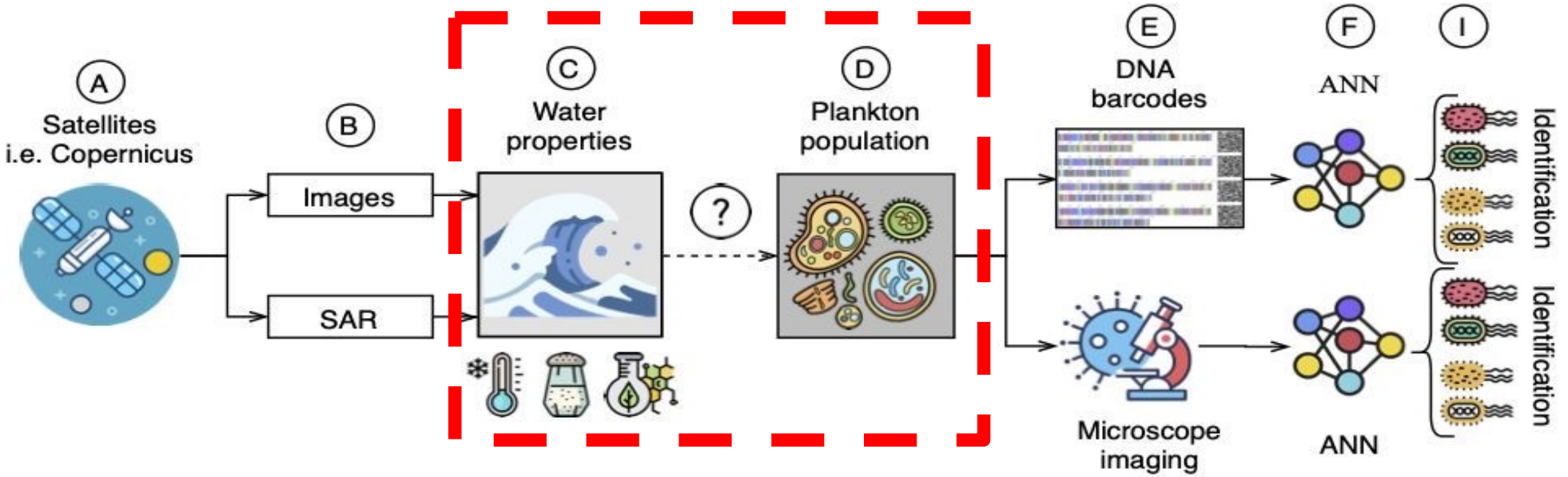




Understanding plankton communities using AI & ML

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

Image by Luis Martí

Predicting environmental conditions from plankton features

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

$$f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$$

That fits

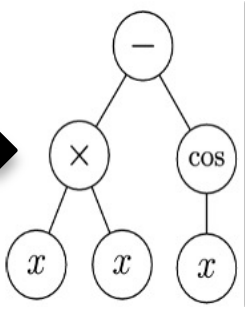
$$y = f(x_0, \ldots, x_n)$$

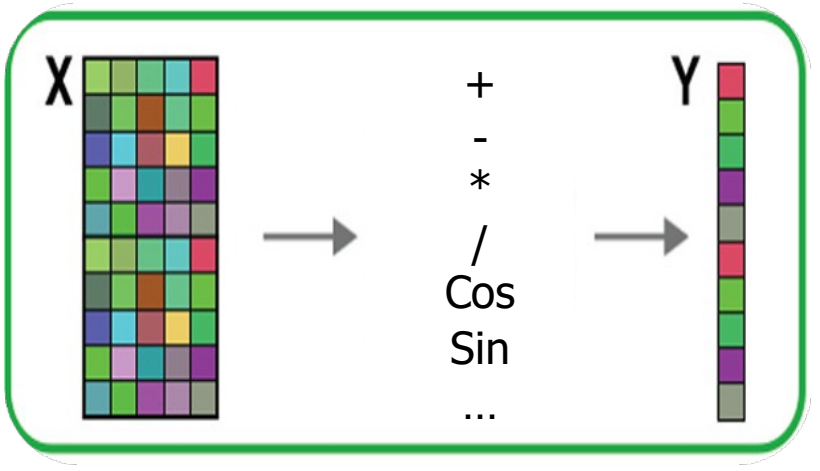
given data

$$(x_0, \ldots, x_n)$$

Predicting environmental conditions from plankton features

$$f(x) = x^2 - \cos(x)$$





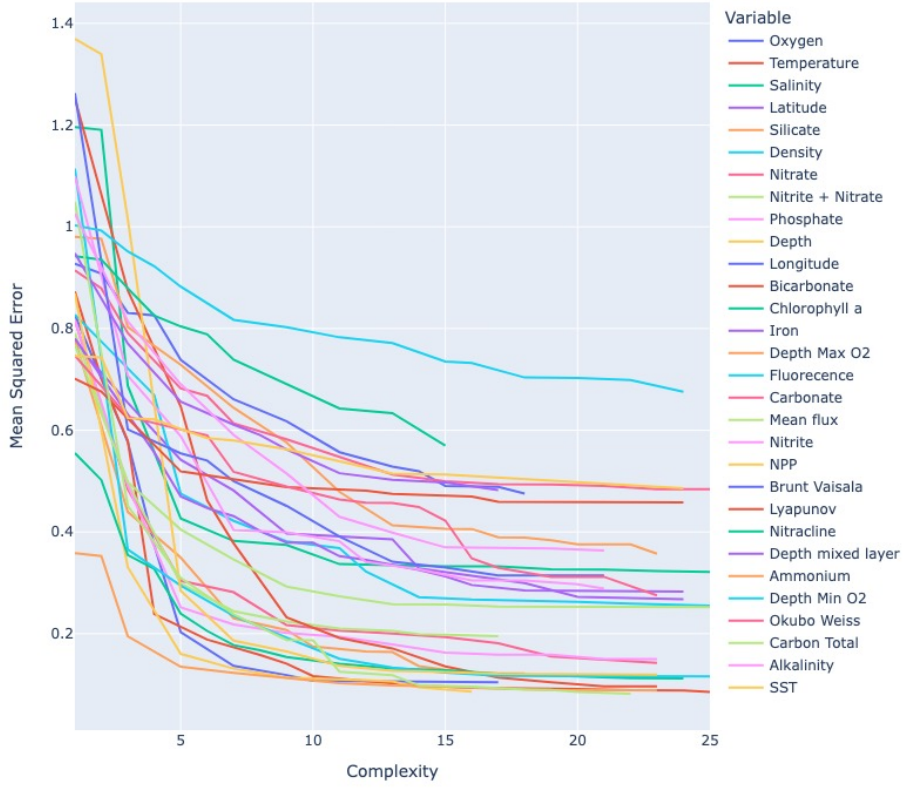
Predicting environmental conditions from plankton features

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

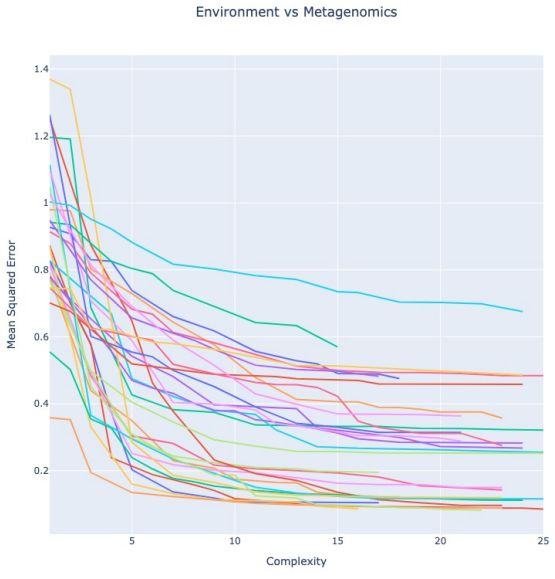
Algorithm 1 Prediction of Environmental Parameters Using SR and Multiple Datasets

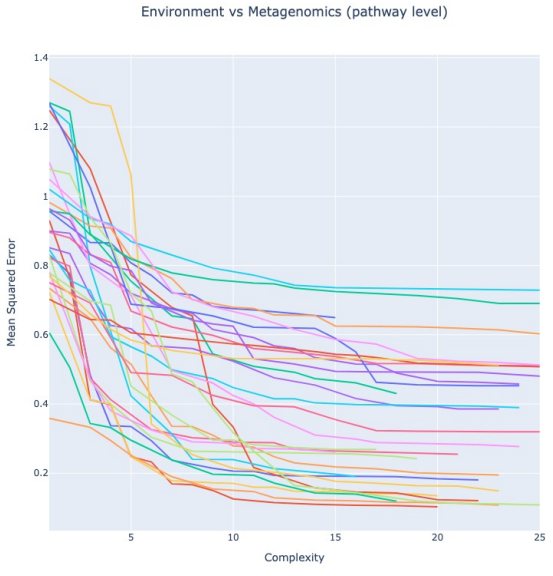
1: **for** each $y_i \in E$ **do**
2: Select a dataset D_j from $\{D_1, D_2, \ldots, D_{11}\}$
3: Use D_j and PySR to predict y_i by inferring the function f_{ij} that best fits the available data.
4: **end for**

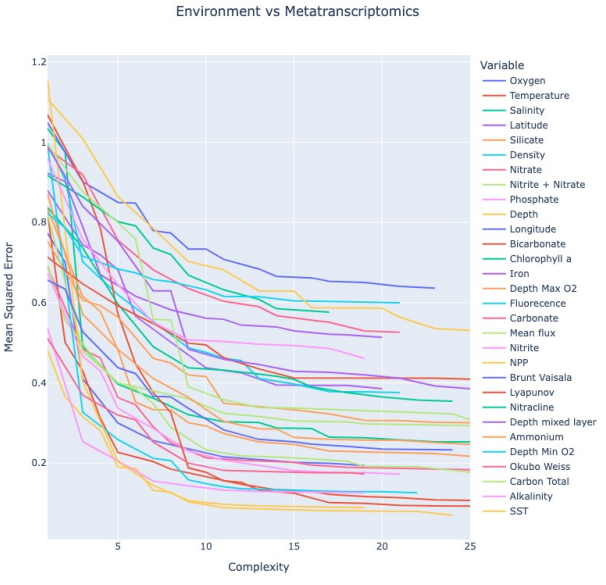
Metagenomic data predicting environmental target

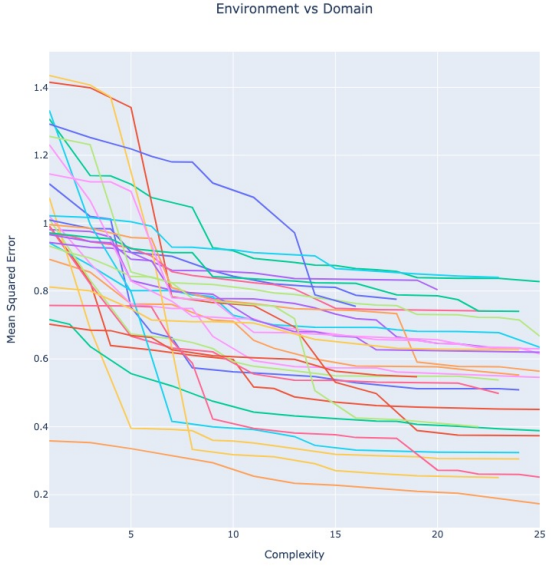


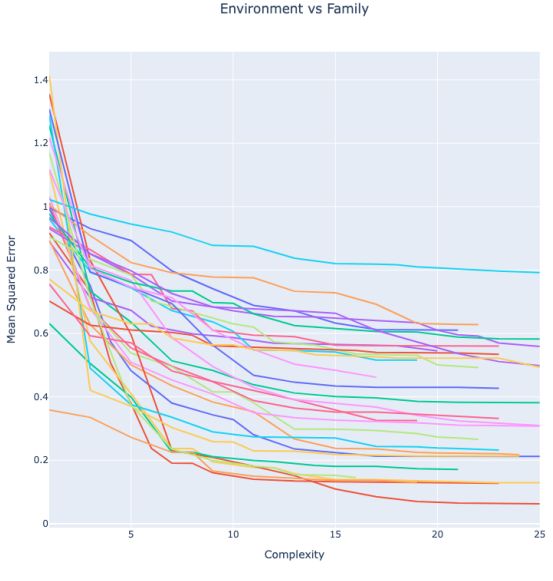
Predicting environmental conditions from plankton features

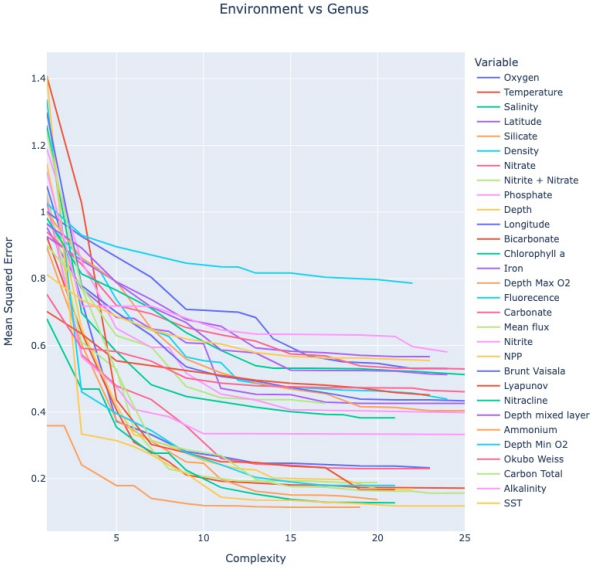




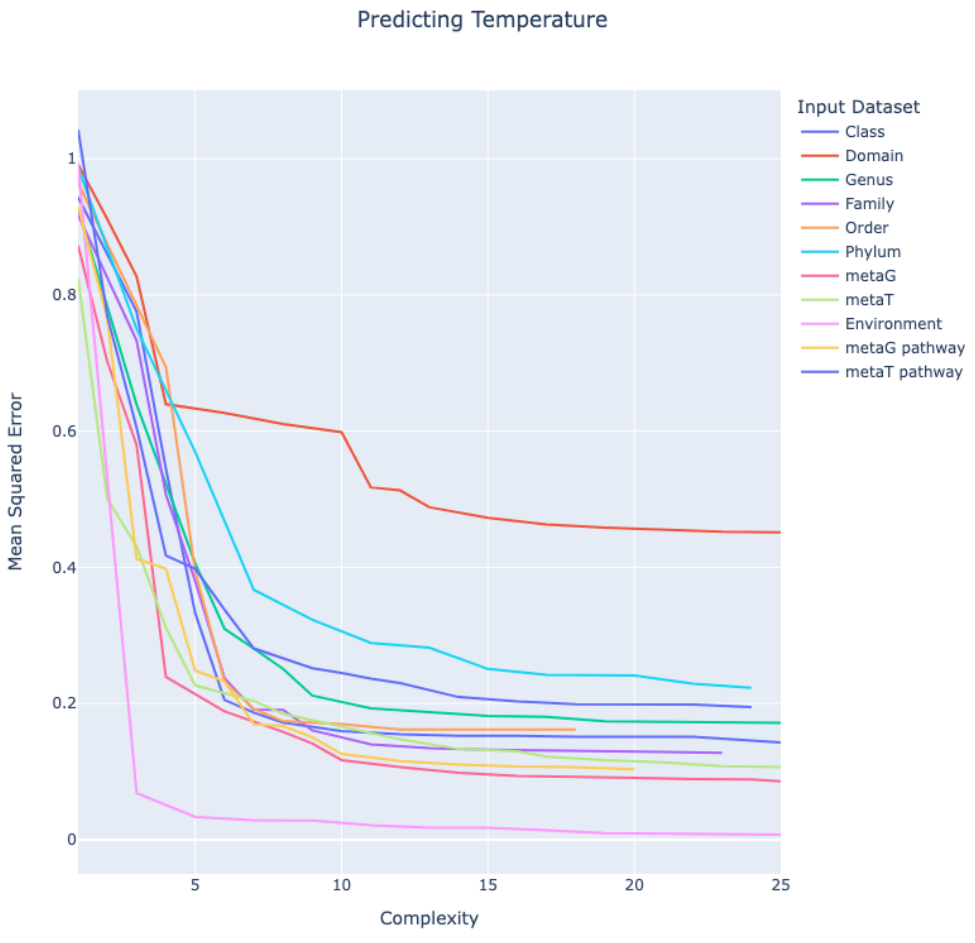


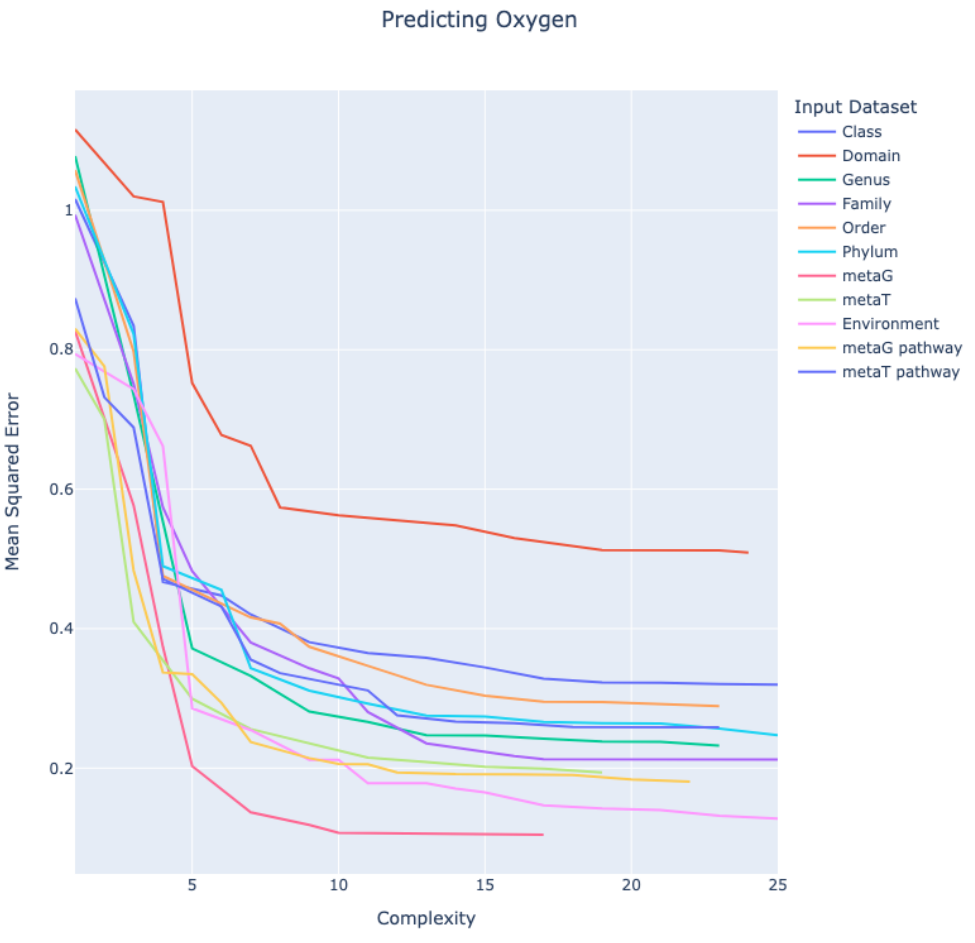




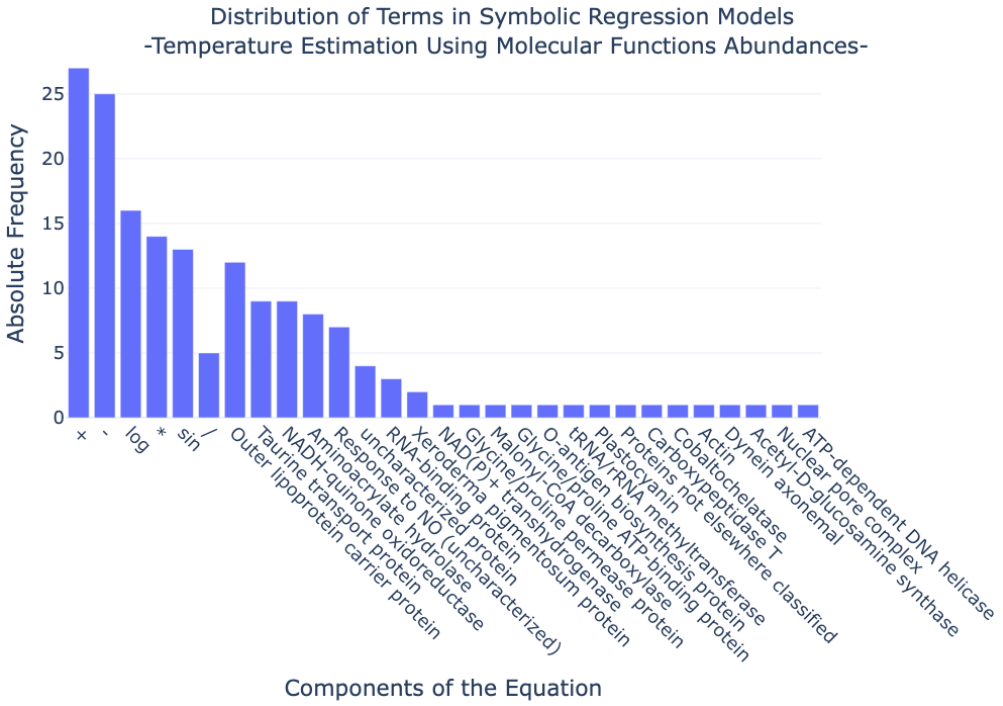


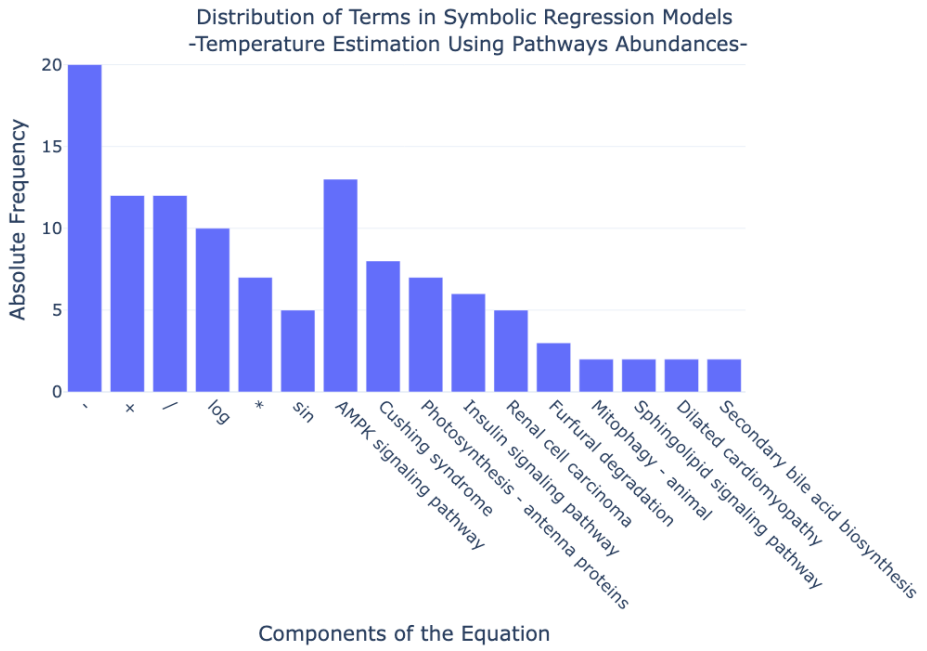
Predicting environmental conditions from plankton features

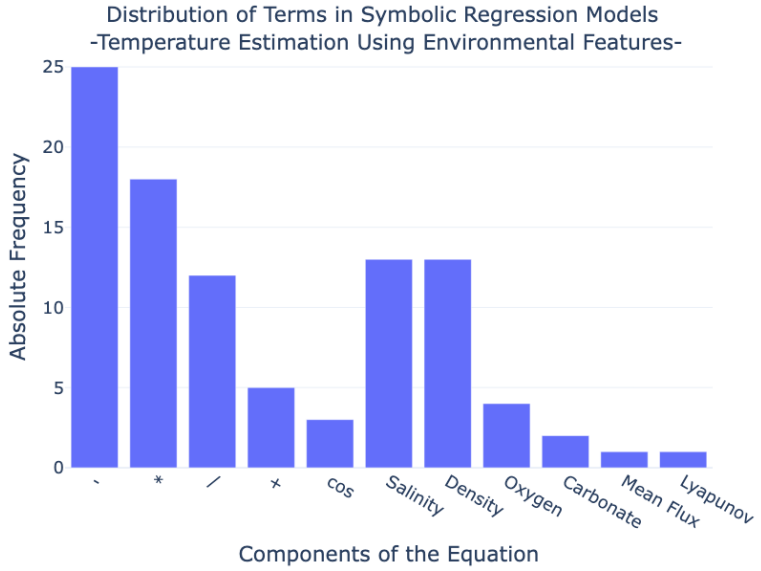




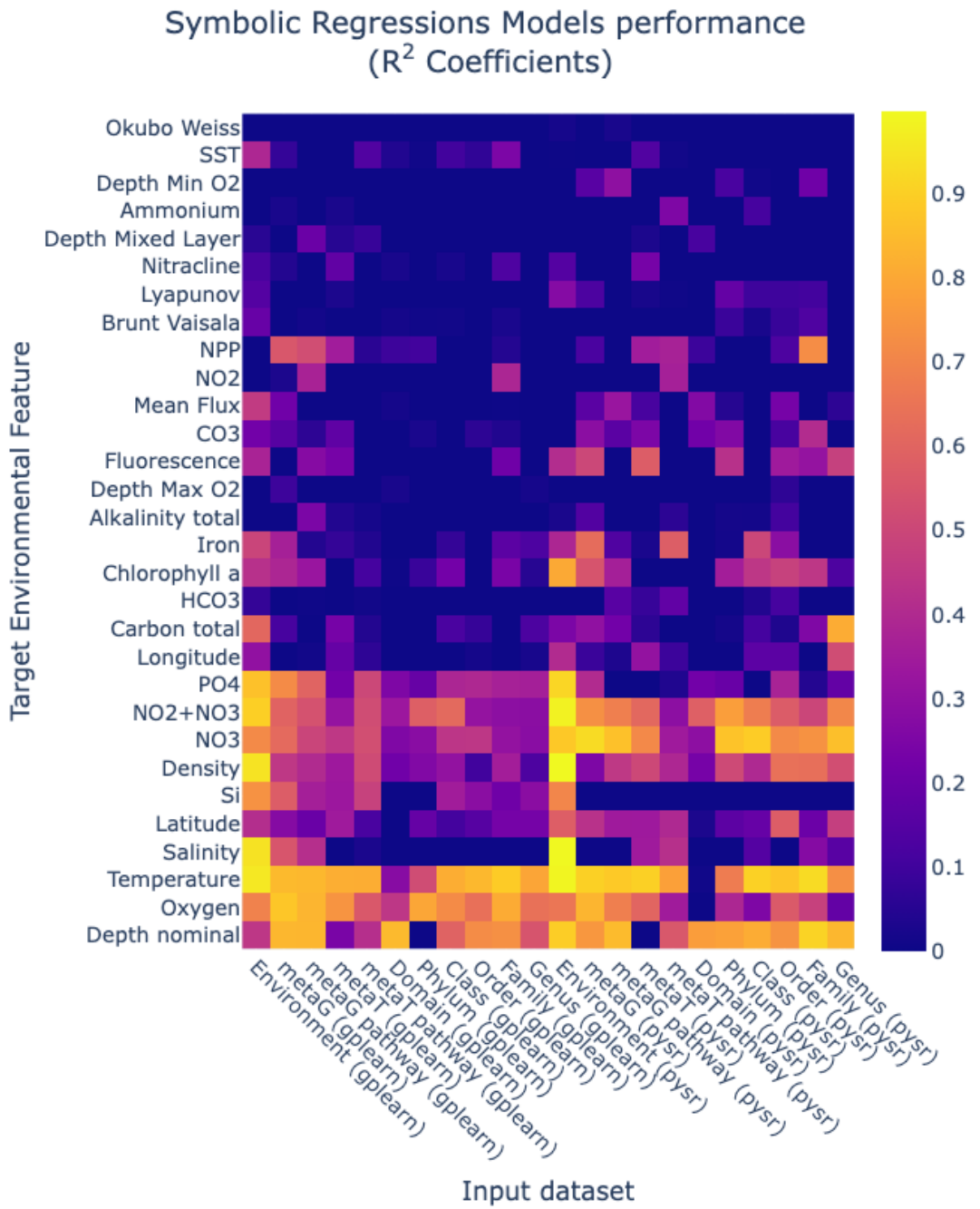
Predicting environmental conditions from plankton features







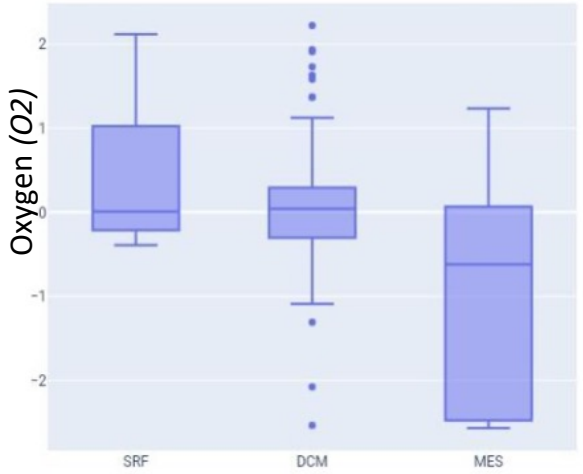
Predicting environmental conditions from plankton features



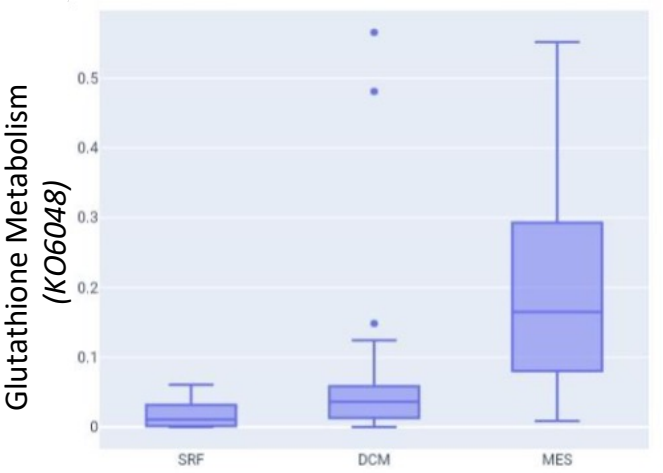
Predicting environmental conditions from plankton features

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

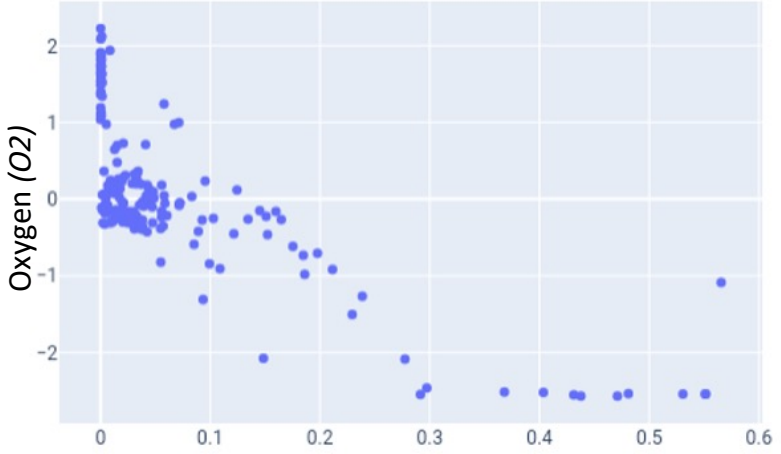
K00856: adenosine kinase.
K11927: ATP-dependent RNA helicase.
K06048: glutamate-cysteine ligase carboxylate-amine ligase,



Layer



Layer



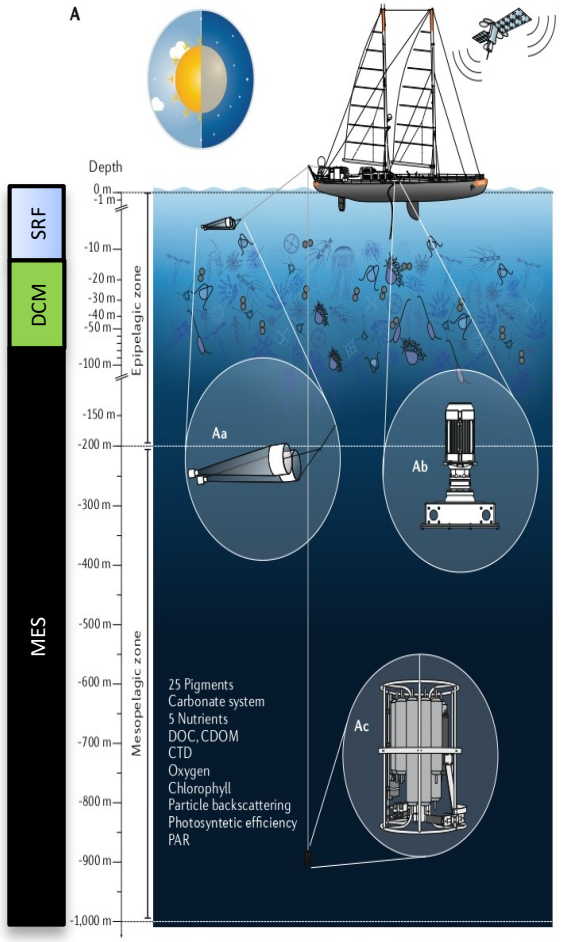
Glutathione Metabolism *(KO6048)*

Predicting environmental conditions from plankton features

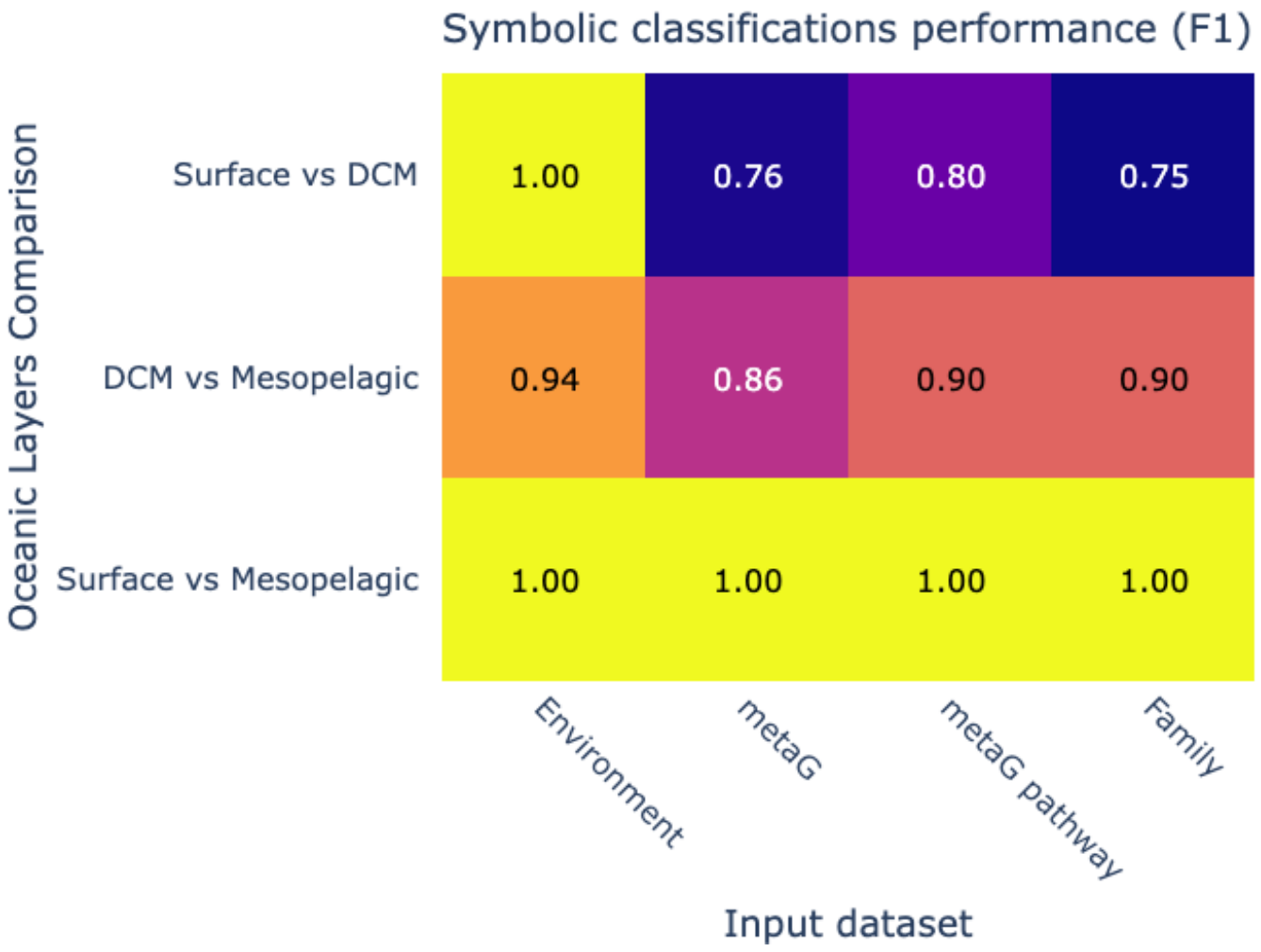
Target	Equation*	R^2
Depth	$(-\mathbf{Chl} \cdot (\mathbf{NO3} + 1.05) + \sin(\mathbf{N})) \cdot \cos(\mathbf{NPP} \cdot \mathbf{O2} - 0.34)$	0.89
Oxygen	$-1.23 \cdot \mathbf{NO3} \cdot \cos(0.80 \cdot \mathbf{L}) - \mathbf{T}$	0.66
Temperature	$-0.92 \cdot \mathbf{D} + 0.81 \cdot \mathbf{S} + 0.19 \cdot \cos(1.22 \cdot \mathbf{D} - 1.88 \cdot \mathbf{S})$	0.99
Nitrate	$0.08 \cdot \mathbf{L} + \mathbf{N} - 0.01 \cdot (0.74 - \mathbf{I})/\mathbf{Si}$	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



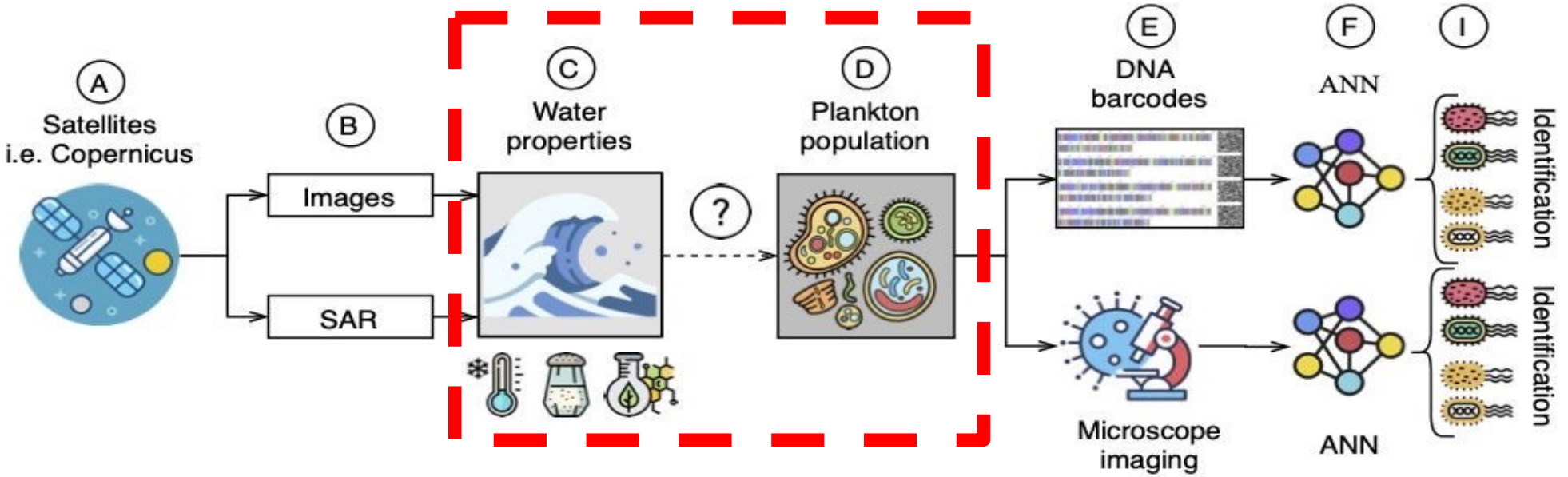
Sunagawa et al. (2020)



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

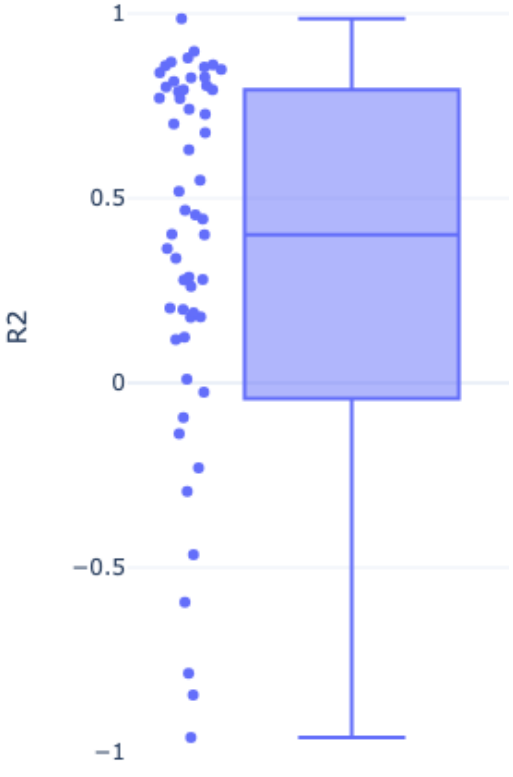
Mapping C -> D

Image by Luis Martí

Predicting plankton features from environmental conditions

Using environmental features to predict directly key metagenomic features

61 Molecular Functions
(from metagenomic dataset)

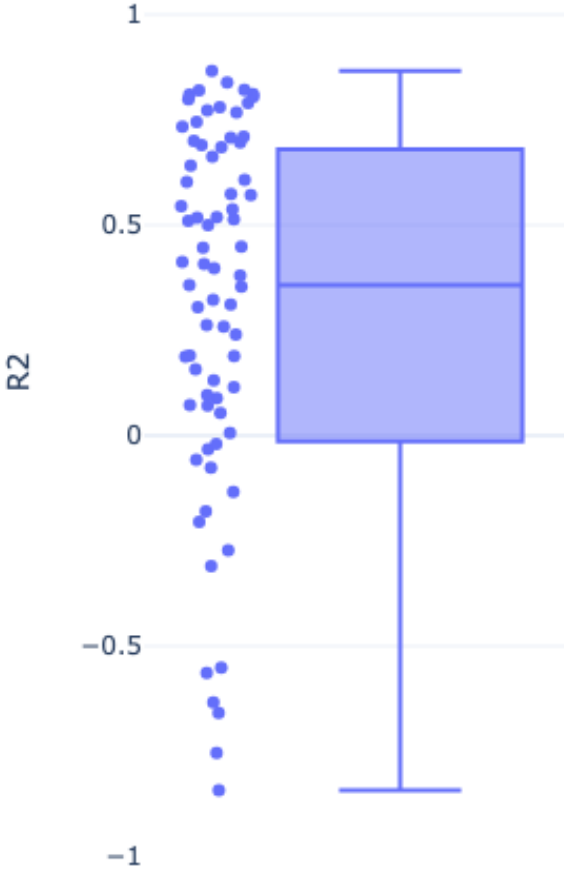


K05724	0.9868	5.4993143e-6/(Iron5m - 0.94)
K00275	0.8970	-0.072*DepthN - 0.072*PO4 + 0.072*Temp + 0.851
K11927	0.8796	0.173*ChlorophyllA - 0.173*Temp + 0.173*sin(O2) + 0.307
K15551	0.8693	-0.021*DepthN - 0.071*Temp + 0.1127
K03433	0.8610	0.333*log(DepthN + 1.612)
K02037	0.8584	(0.0457*Iron5m + 0.200)/(PO4 + 1.156)
K15635	0.8553	0.202*log(DepthN + 1.570)
K03163	0.8489	DepthN*(0.075 - 0.0186*Longit) + 0.0435
K07574	0.8394	-0.062*Temp + 0.062*sin(Latit*(-CarbonT + ChlorophyllA - Latit)) + 0.09
K02533	0.8270	0.093*Temp + 0.297
K13525	0.8264	0.470*log(DepthN + 1.654)
K05501	0.8166	0.030*ChlorophyllA + 0.030*O2 + 0.108
K05934	0.8053	0.061*DepthN + 0.061*sin(DepthN) + 0.0619
K00324	0.8007	0.216*Temp + 1.201

Predicting plankton features from environmental conditions

Using environmental features to predict directly key metagenomic features

79 Pathways
(from metagenomic dataset)

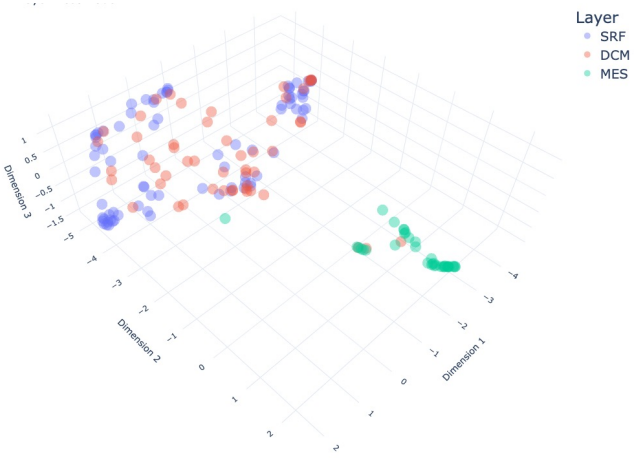


Glycine, serine and threonine metabolism	0.865	ChlorophyllA - Temp + 25.171
Retinol metabolism	0.838	0.126*DepthN - 0.126*Temp + 0.470
Leishmaniasis	0.821	0.089*DepthN + 0.104
Phenylalanine metabolism	0.820	0.536*ChlorophyllA - 0.536*Temp + 6.649
Fanconi anemia pathway	0.811	0.181*DepthN + 0.195
Basal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

Predicting plankton features from environmental conditions

metagenomic dataset
(**9024 Molecular Functions**)

UMAP
+
Autoencoder

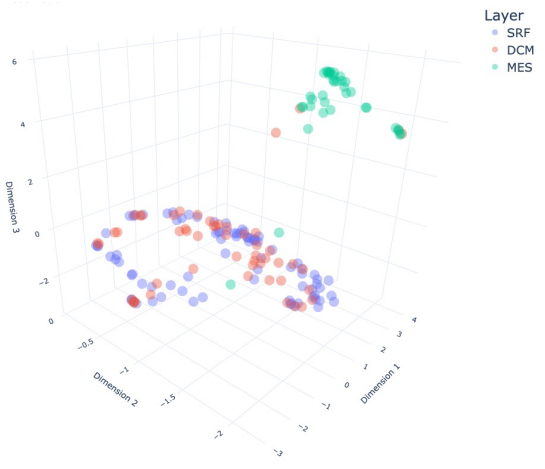


Symbolic
Regressions

Environmental features

metagenomic dataset
(**453 pathways**)

UMAP
+
Autoencoder

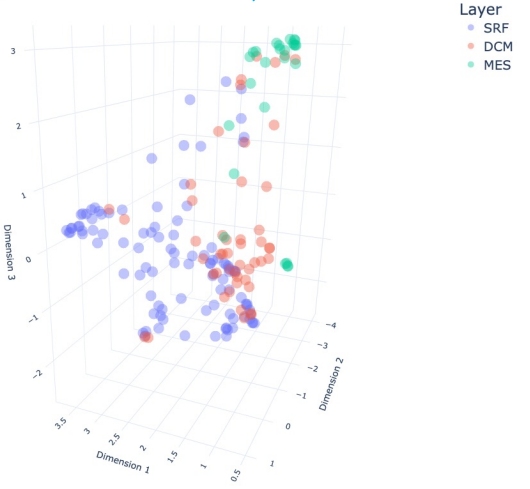


Symbolic
Regressions

Environmental features

metatranscriptomic dataset
(**8935 Molecular Functions**)

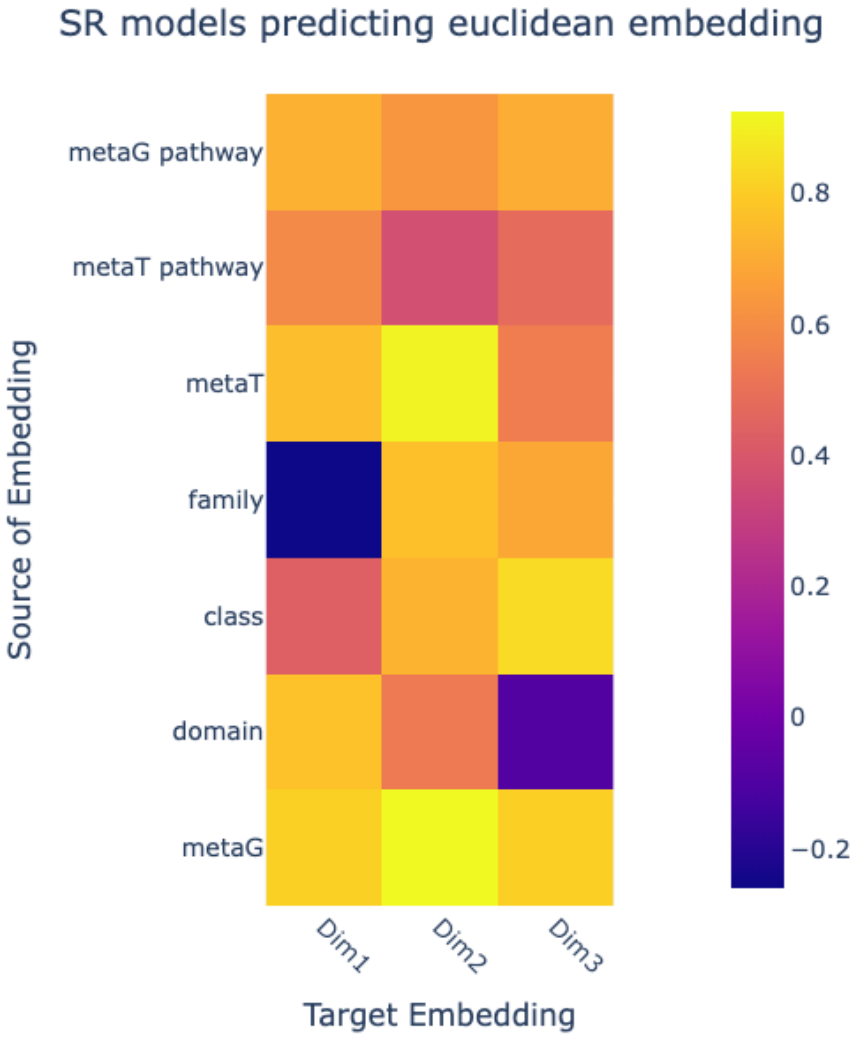
UMAP
+
Autoencoder

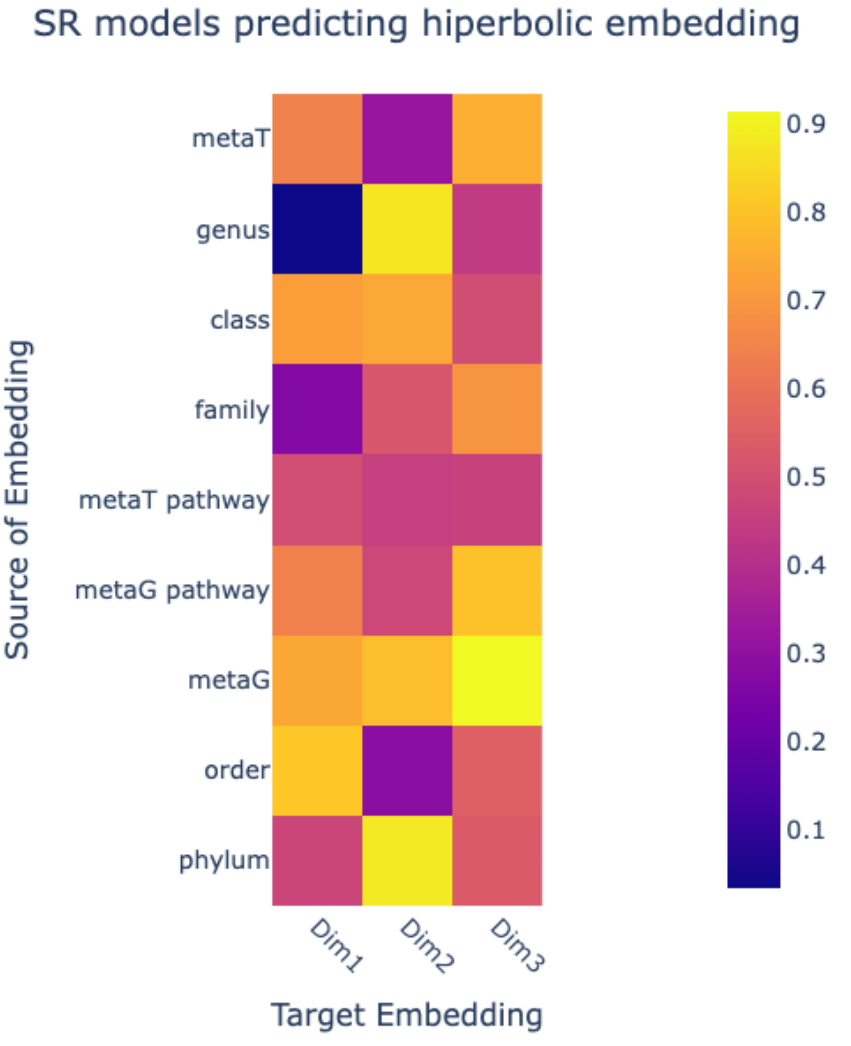


Symbolic
Regressions

Environmental features

Predicting plankton features from environmental conditions





Conclusion

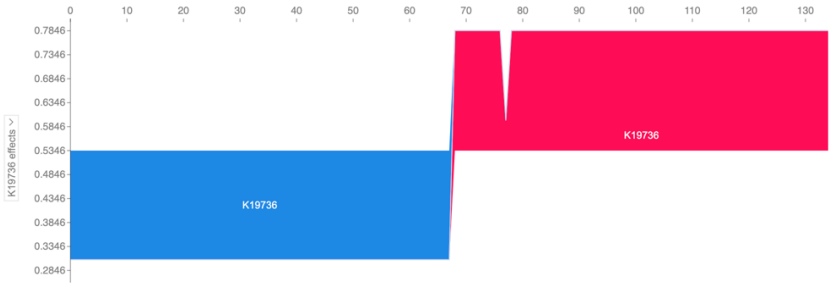
- After reducing the dimensionality of each of the datasets and visualizing general patterns, the influence of polarity and the depth at which the sample was taken stand out. Categories such as provinces or oceans only group samples when they come from the surface.
- There is a subset of environmental 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

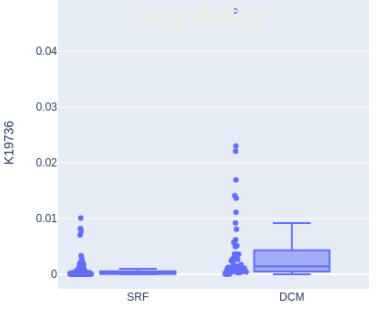
- Joint work with Nahuel Pilquinao, José Vásques, Luis Martí, Nayat Sánchez-Pi.


Thank you! Obrigado! Merci ! ¡Gracias!
https://inria.cl

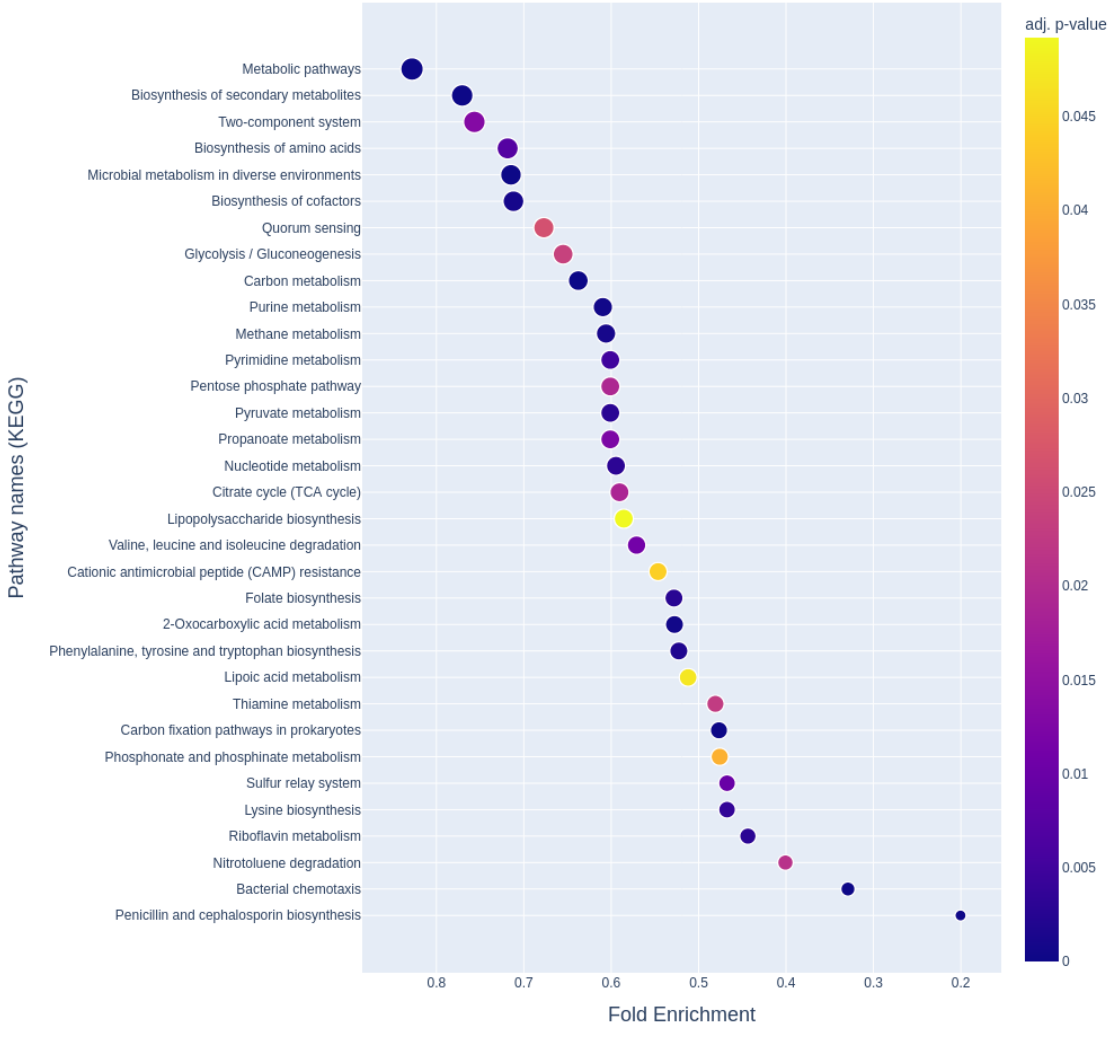
(Metagenomic) Layers Characterizacion: SRF vs DCM (5 vs 55 m depth)



TetR/AcrR family transcriptional regulator



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



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