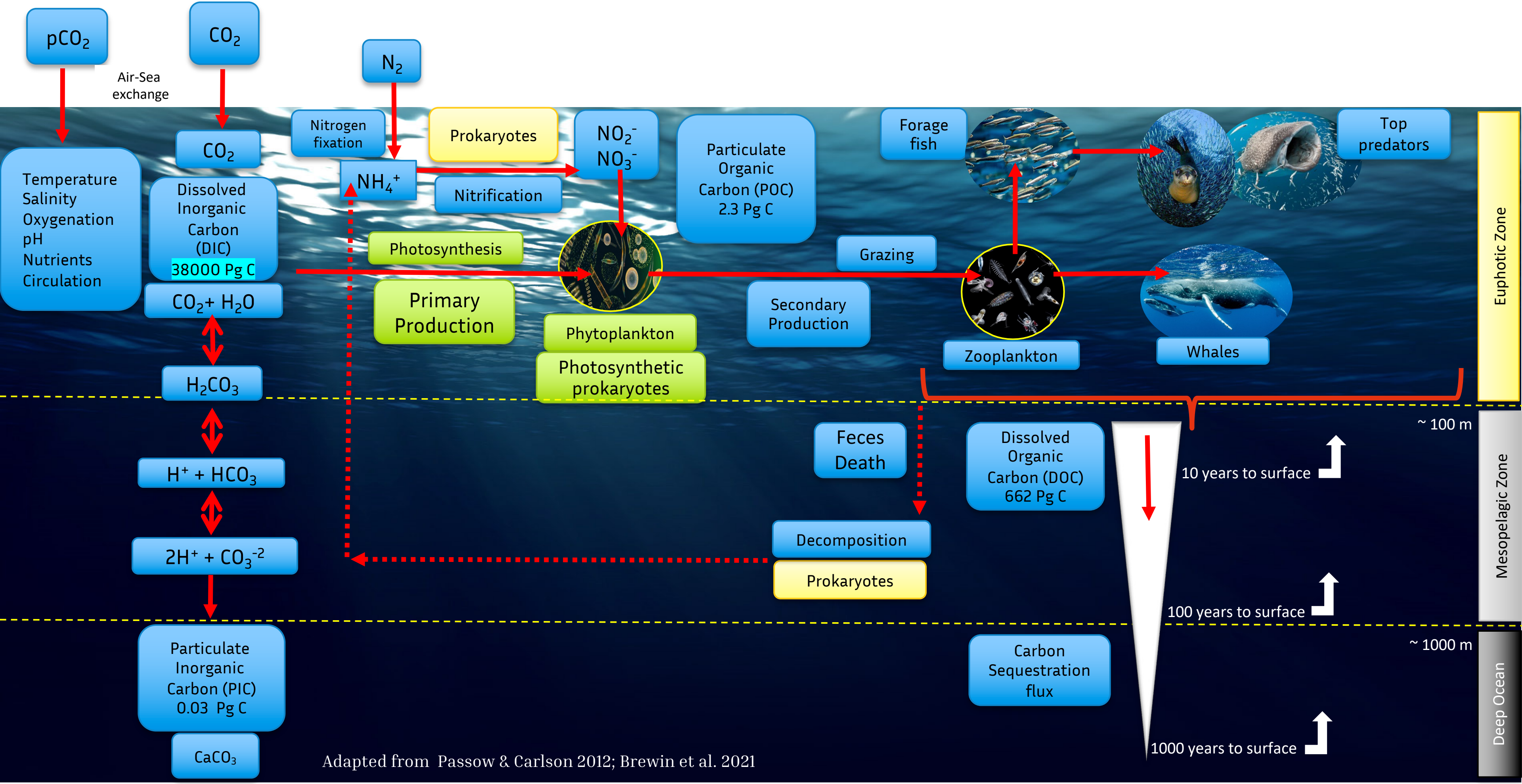
"Enhancing Interpretability of Global Plankton Communities Modeling through Multi-Omics and AI Techniques"

Luis Valenzuela

Workshop: Towards a modern analysis of omics data of the Ocean

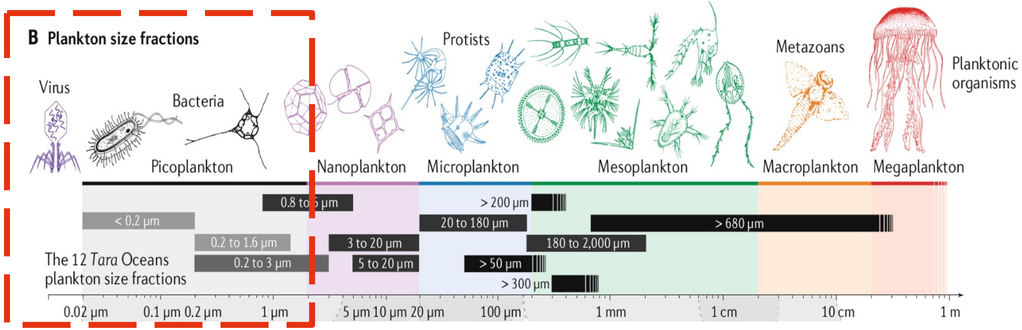
Instituto Sistemas Complejos, Valparaíso, Chile, 2023

Climate Change, Oceanic Carbon pumps and plankton roles

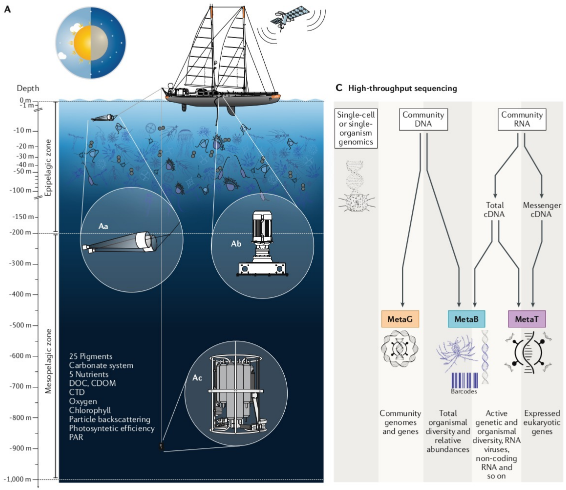


Adapted from Passow & Carlson 2012; Brewin et al. 2021

Ocean Microbial Reference Gene Catalog v2



Salazar et al. (2019)



Sunagawa et al. (2020)

Ocean Microbial Reference Gene Catalog v2:

Ocean Microbial Reference Gene Catalog v2

Metagenomic dataset:

~ 57.000 million reads (90 pb ±2.6pb)

ATGC…CTGGG$_{90}$	…	TTCA…TTTCC$_{90}$
GCCC…AAAAG$_{90}$		GGGGA…AGCTA$_{90}$

(meta)genome assembly

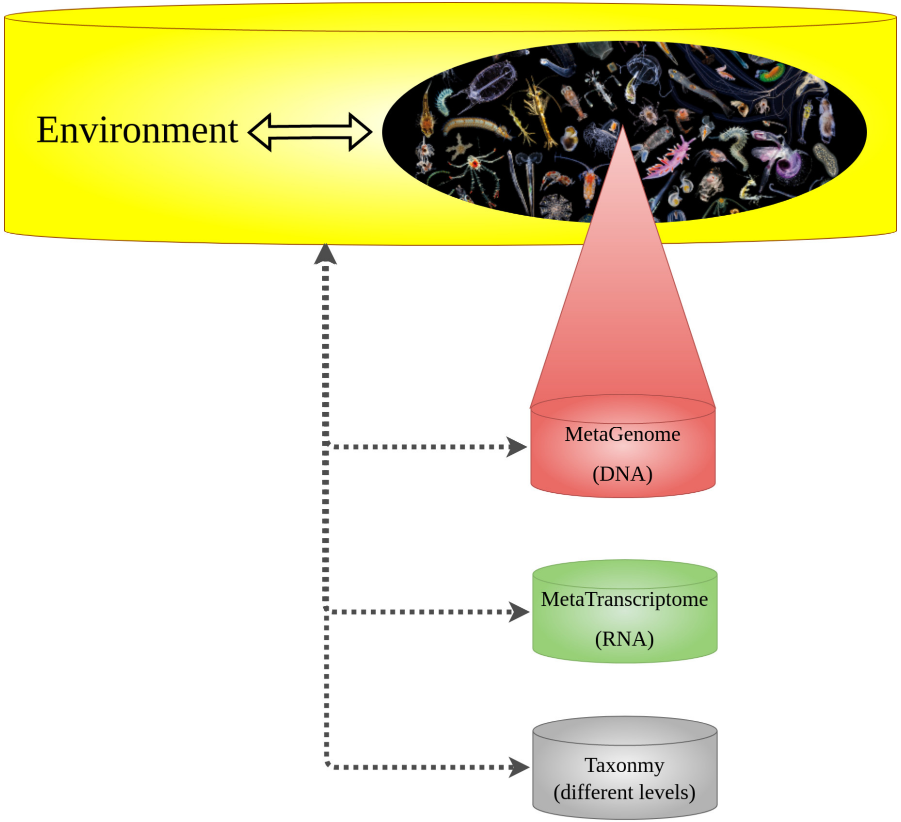
~ 200 metagenomes:

~ 230000 scaffolds, N50: 1300pb (average)

~ 42 million genes

Functional annotation
Quantification
Normalization

	Gene$_1$	Gene$_2$	…	Gene$_{42M}$
metaG$_1$				
metaG$_2$				
…				
metaG$_{200}$				

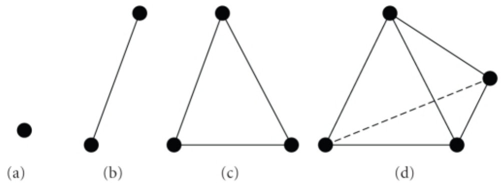


Salazar et al., Gene expression changes and community turnover differentially shape the global ocean metatranscriptome, Cell, 2019.

https://www.ebi.ac.uk/biostudies/studies/S-BSST297

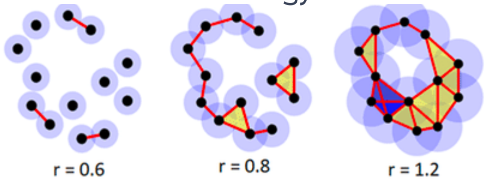
Persistent Homology and Uniform Manifold Approximation and Projection (UMAP)

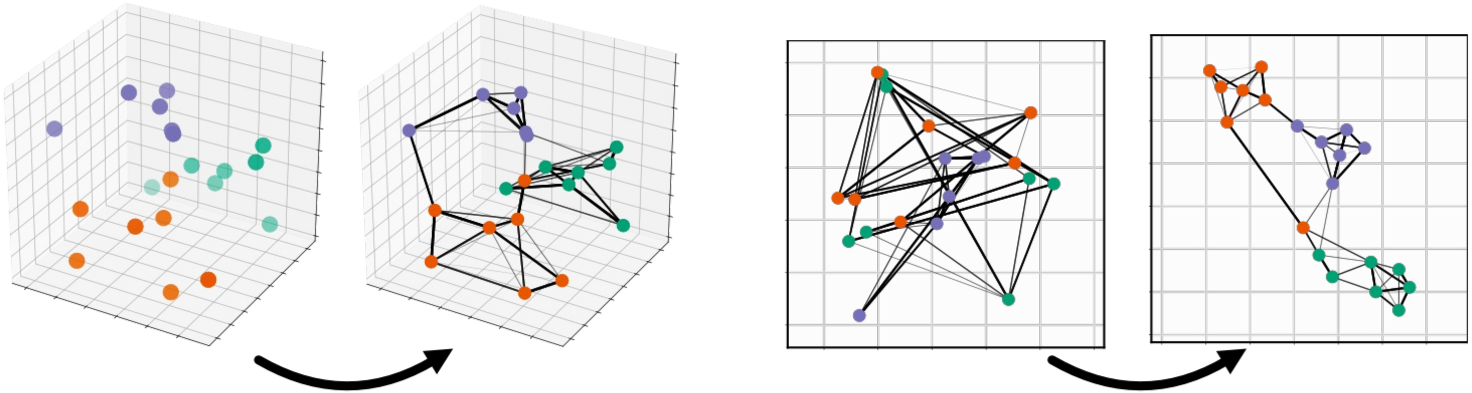
K-Simplicial Complexes



(a) 0-simplex is a vertex, (b) 1-simplex is a line,
(c) 2-simplex is a triangle, (d) 3-simplex is a tetrahedron.

Persistent Homology filtrations





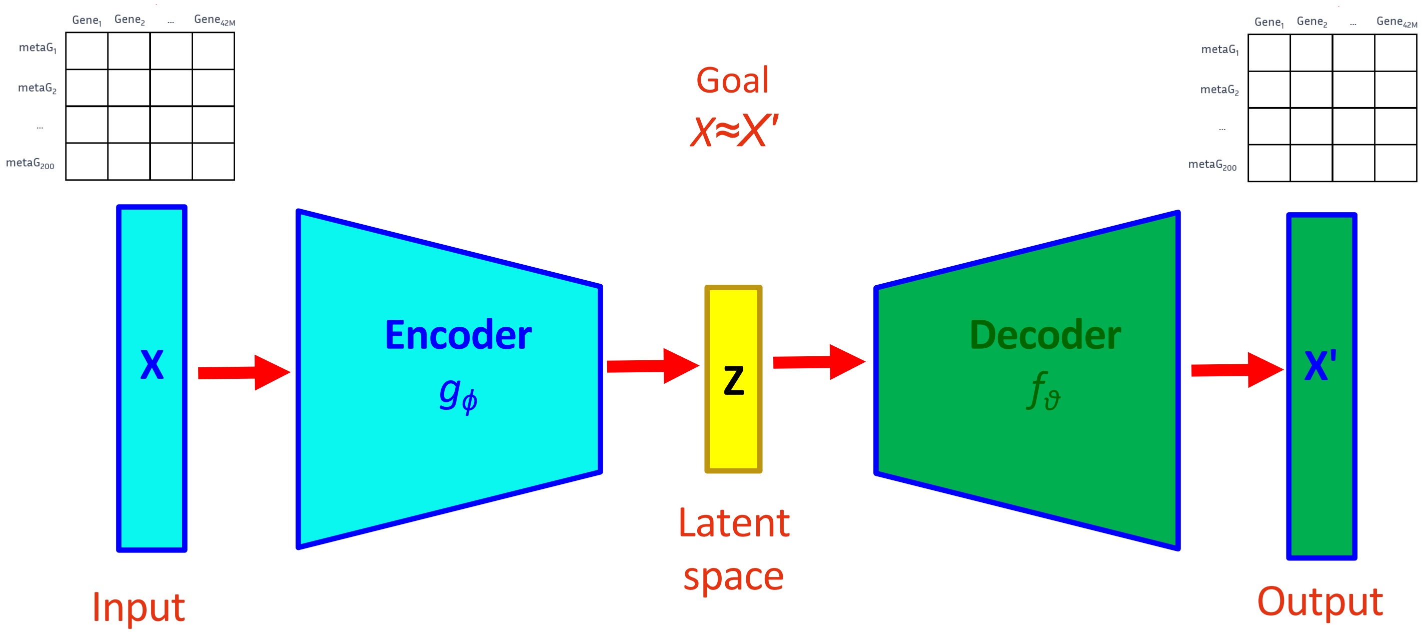
Compute a graphical representation of the dataset

Learn an embedding that preserves the structure of the graph

https://umap-learn.readthedocs.io/en/latest/

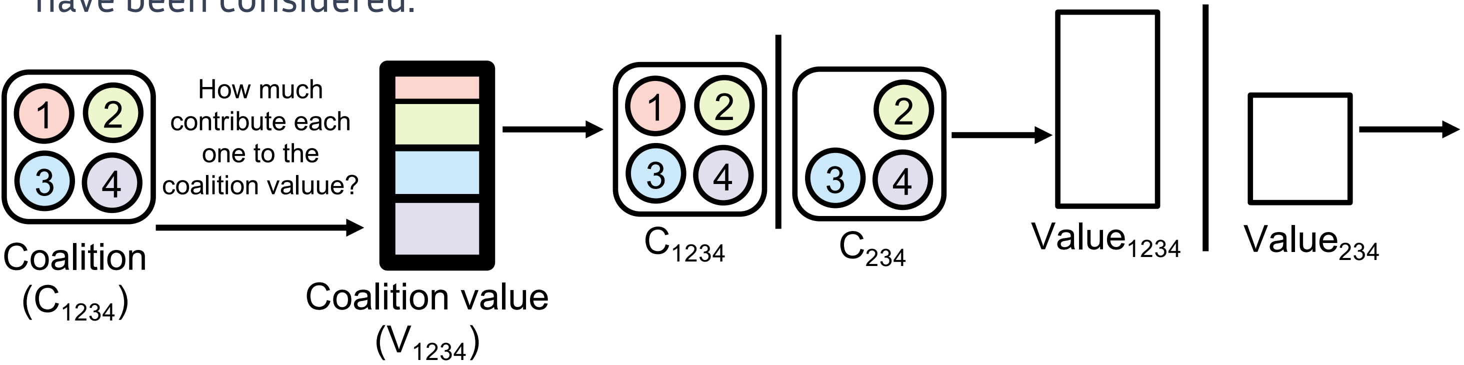
Autoencoder

An Autoencoder is a neural architecture designed to learn an identity function in an unsupervised way to reconstruct the original input while (usually) compressing the data in the process to discover a more efficient representation.



SHapley Additive exPlanations (SHAP) values

Shapley value is the average expected marginal contribution of one feature after all possible combinations have been considered.



How much contribute each one to the coalition valuue?

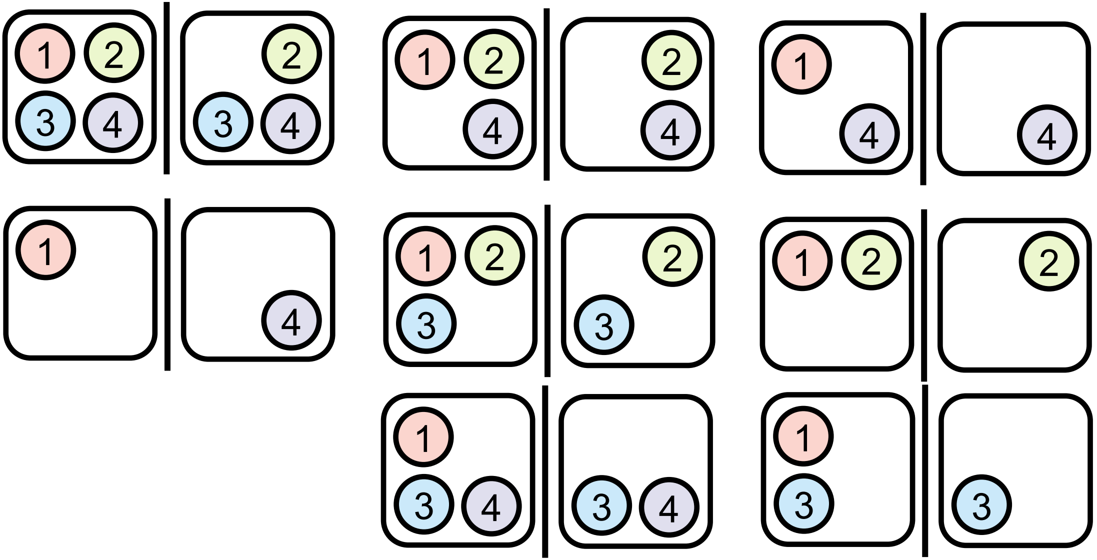
Coalition (C_{1234})

Coalition value (V_{1234})

C_{1234} C_{234}

Value_{1234} Value_{234}

V_{1234} - V_{234} = Marginal contribution of member 1 to C234



Shapley values are the mean marginal contribution.

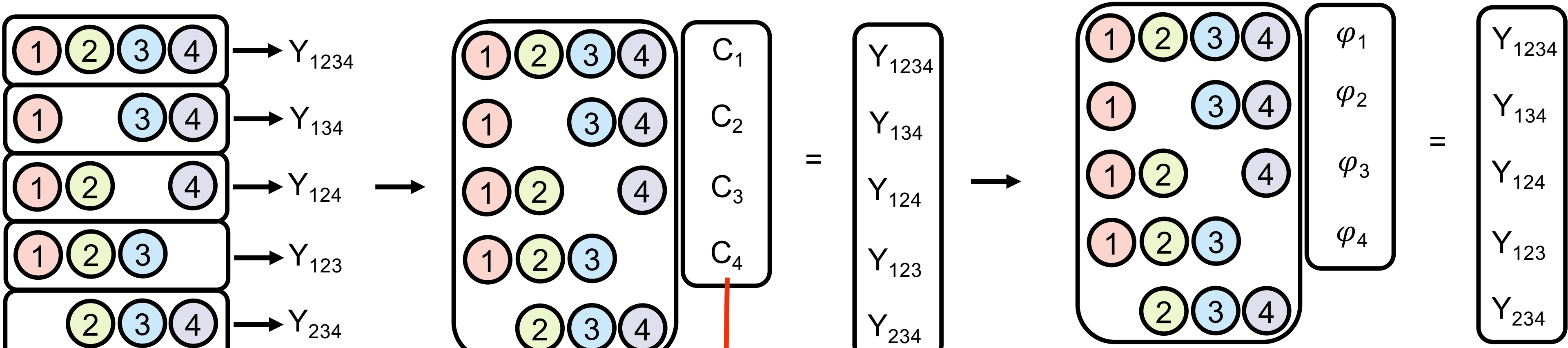
$$\varphi_i = \frac{1}{\#\,Members} \sum_{\forall C\ s.t.\ i \neg\in C} \frac{Marginal\ Contribution\ of\ i\ to\ C}{\#\,Coalitions\ of\ size\ |C|}$$

Lundberg & Lee, 2018. A Unified Approach to Interpreting Model Predictions. 31st Conference on Neural Information Processing Systems.

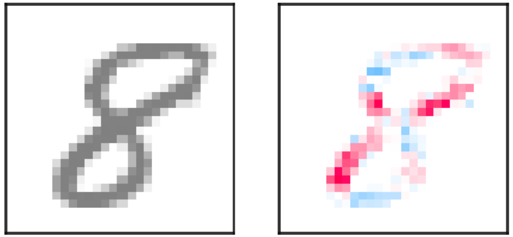
SHapley Additive exPlanations (SHAP) values

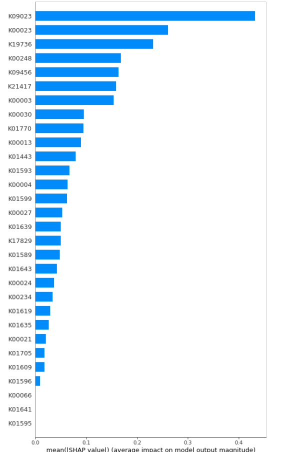
SHAP is a game theoretic approach to explain the output of any machine learning model.

SHAP Kernel approximates Shapley values through much fewer samples.



$$W_C = \frac{\#\,total\ features\ -\ 1}{\#\,coalitions\ of\ size\ |C| * \#included\ features\ in\ C\ * \#excluded\ features\ in\ C}$$

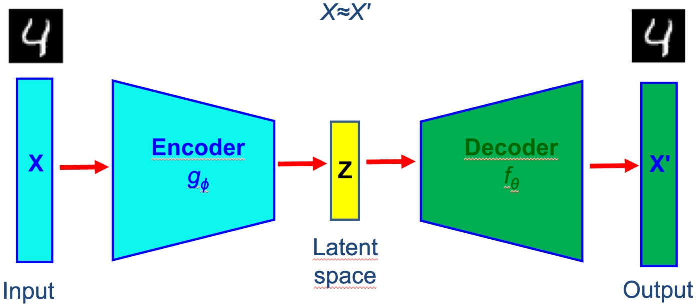


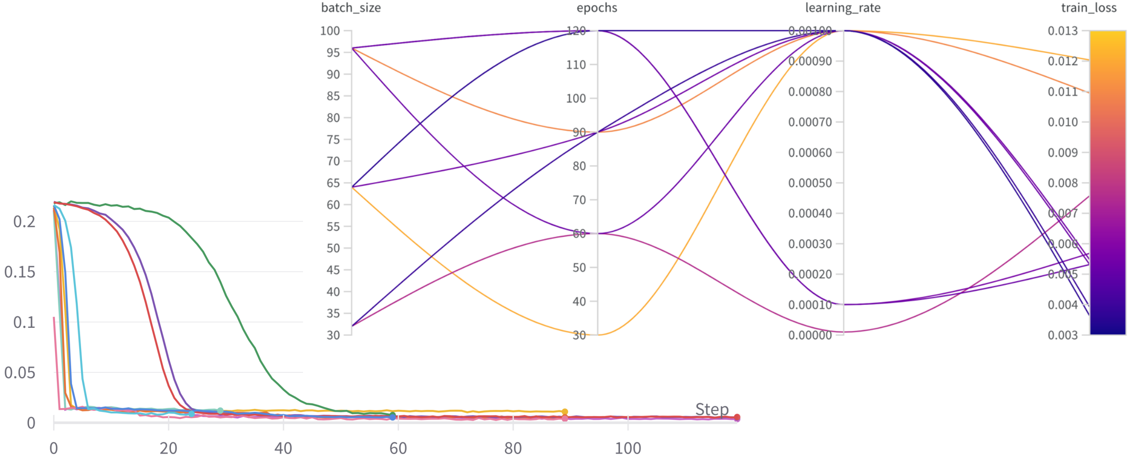


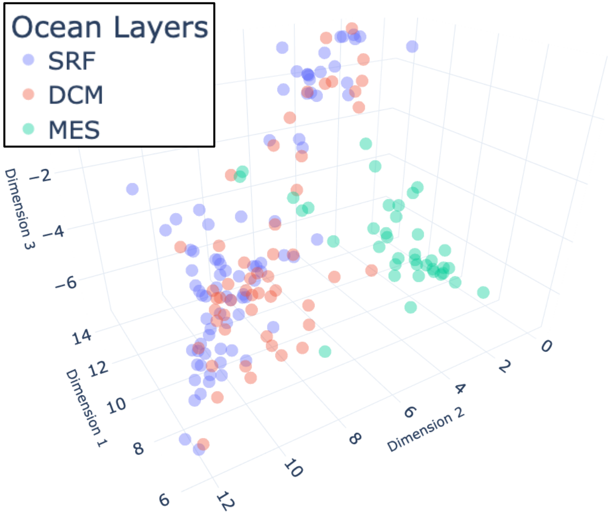
Lundberg & Lee, 2018. A Unified Approach to Interpreting Model Predictions. 31st Conference on Neural Information Processing Systems.

Exploration of the Omics datasets

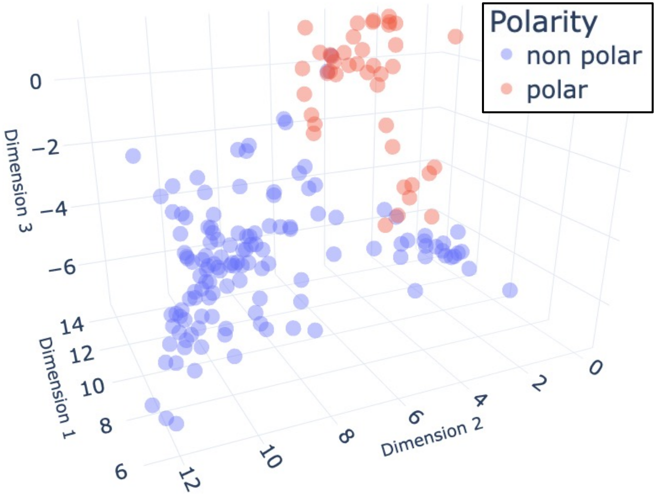
Metagenomic assays: Functional Composition (KEGG)





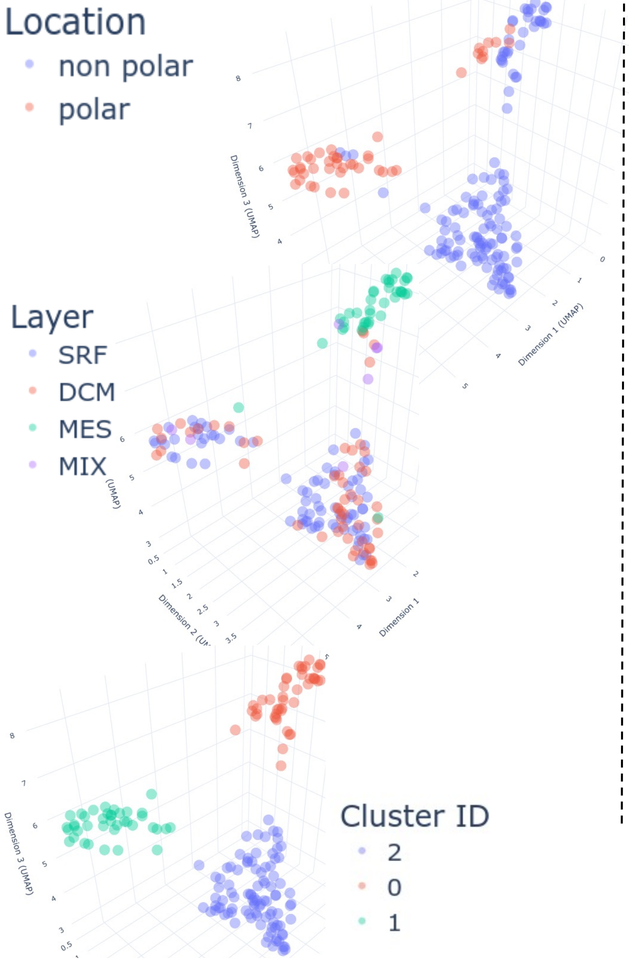


Latent Space

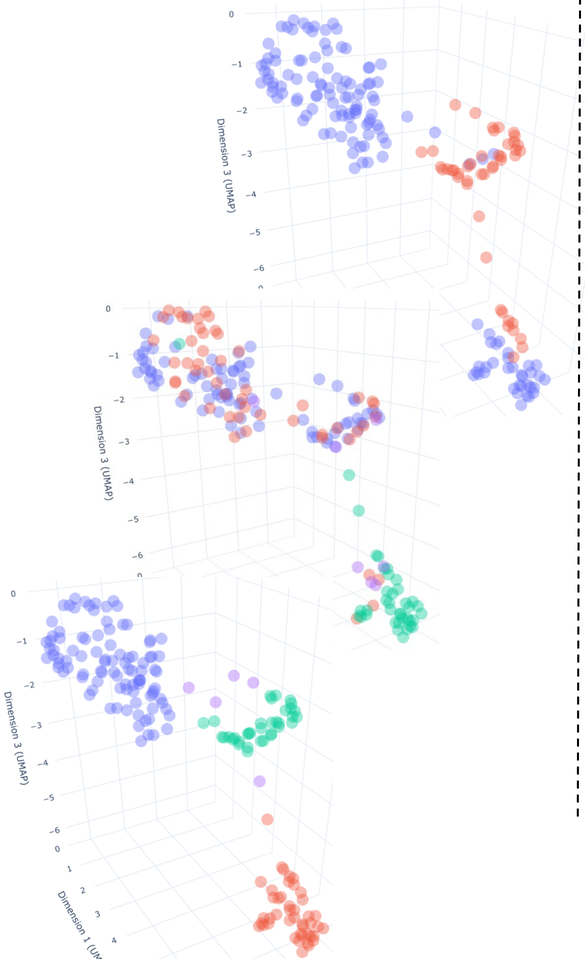


Metagenomic assays: Functional Composition

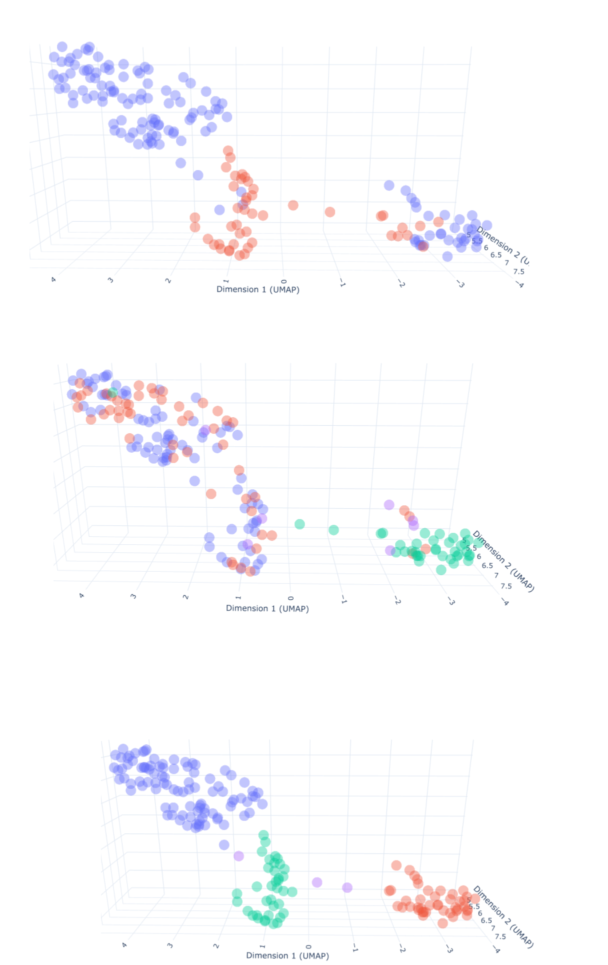
i) KEGG db: 9026 features



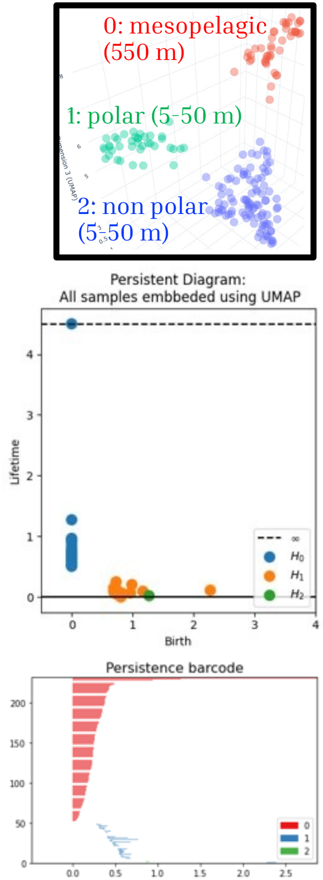
ii) eggNOG db 76022 features



iii) eggNOG+GC db 325936 features

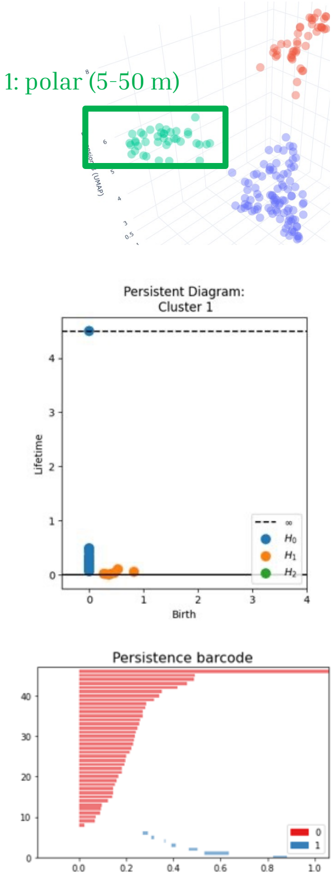


Persistent Homology from Metagenomic composition



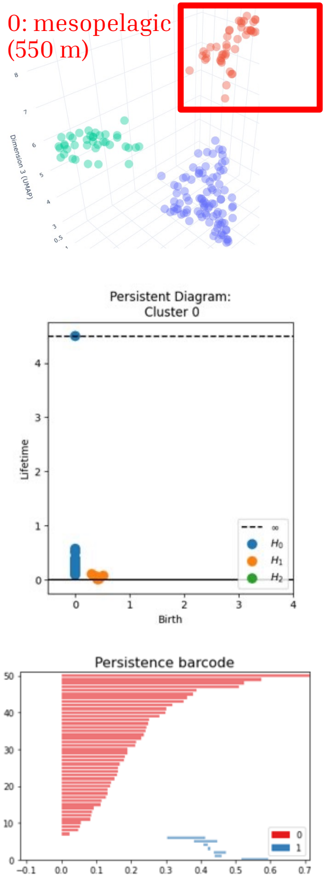
Persistent Diagram: All samples embbeded using UMAP

Persistence barcode



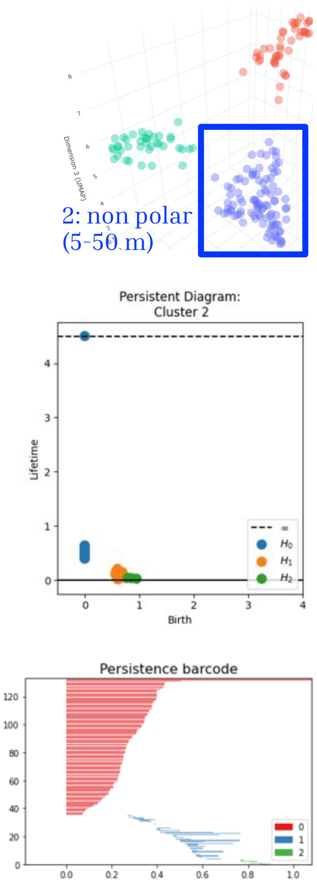
Persistent Diagram: Cluster 1

Persistence barcode



Persistent Diagram: Cluster 0

Persistence barcode

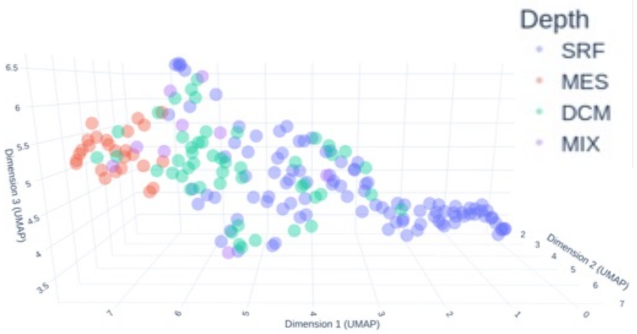


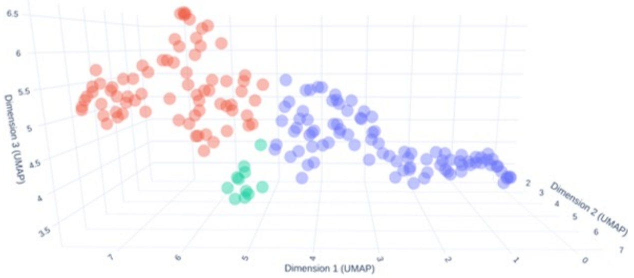
Persistent Diagram: Cluster 2

Persistence barcode

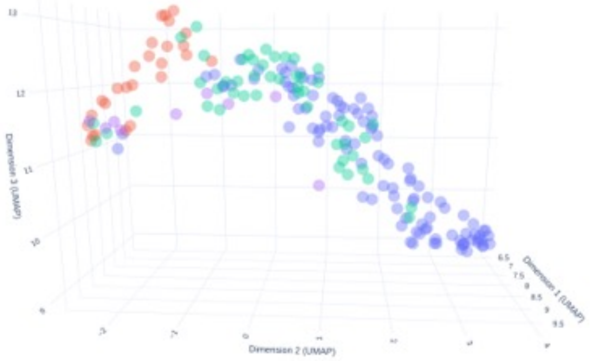
Metatranscriptomic assays: Gene expression

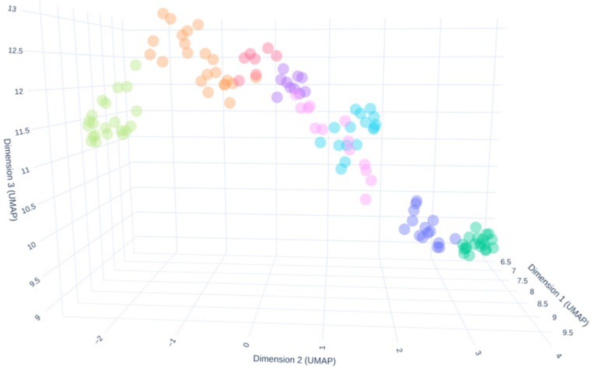
i) KEGG db: 8937 features



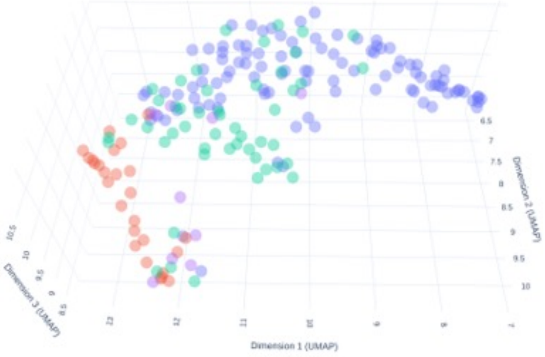


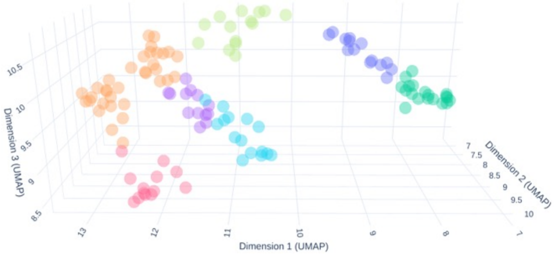
ii) eggNOG db 71662 features





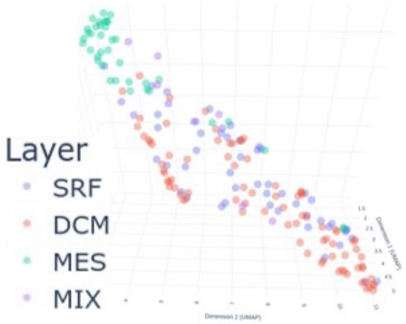
iii) eggNOG+GC db 314715 features



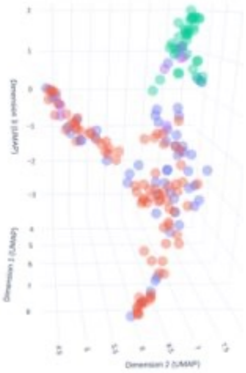


Taxonomy datasets

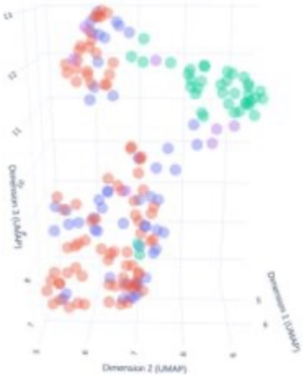
Domain level



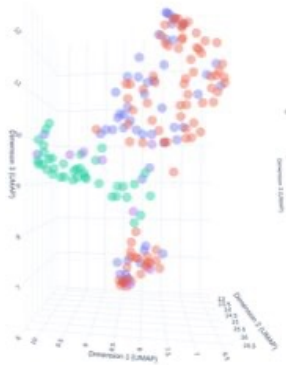
Phylum level



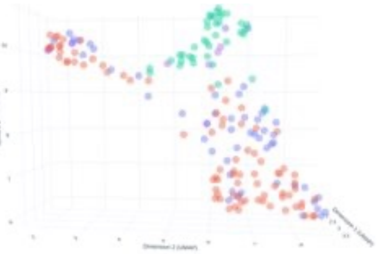
Class level



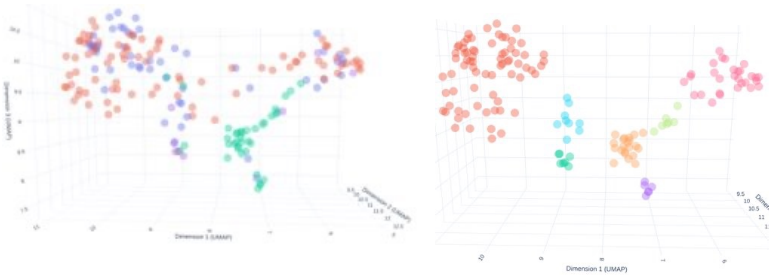
Order level

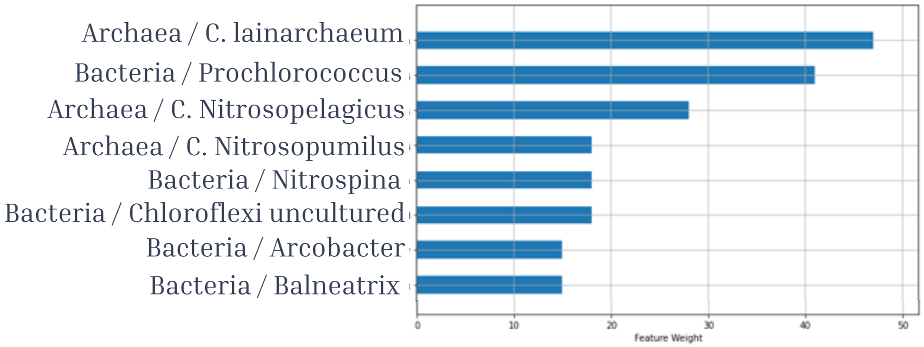


Family level

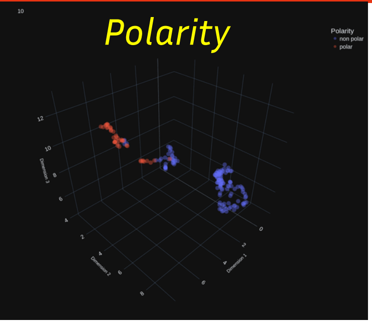


Genus level

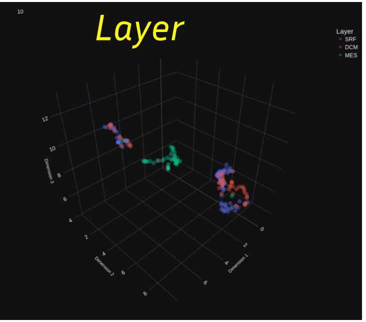


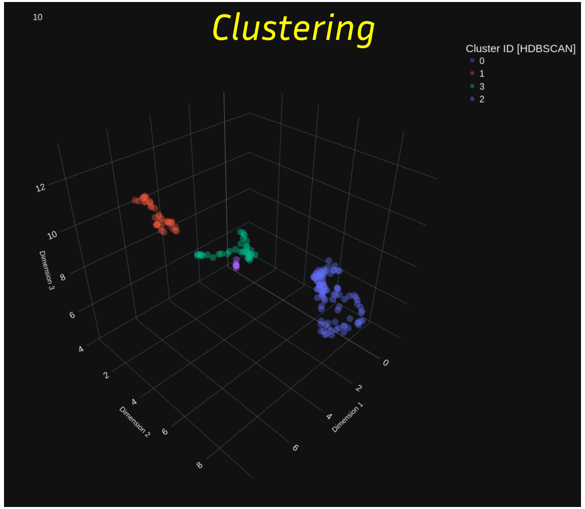


General exploration: Stability of clusters (metaG example)



$\mathbb{R}^{9024} \longrightarrow \mathbb{R}^3$



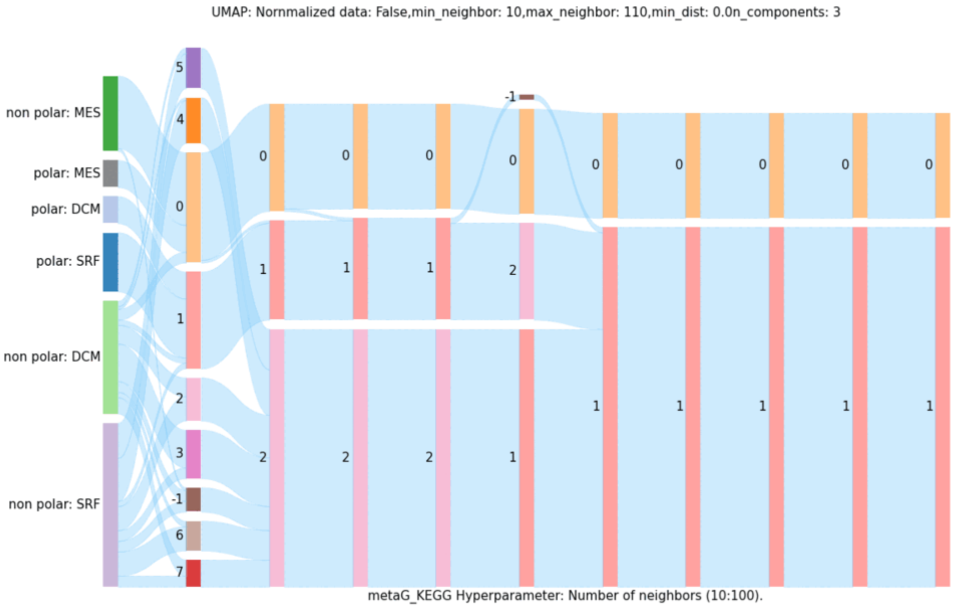


min dist: 0.1

[10 models]

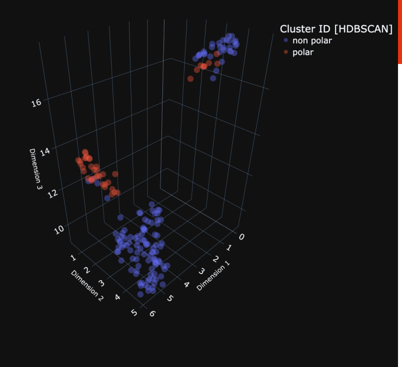
Sankey plot: exploring min dist hyper-parameter:

gif: 0.0, 0.05, 0.10, 0.15, 0.20, 0.25,0.30,0.35,0.40,0.45,0.50



[110 models]

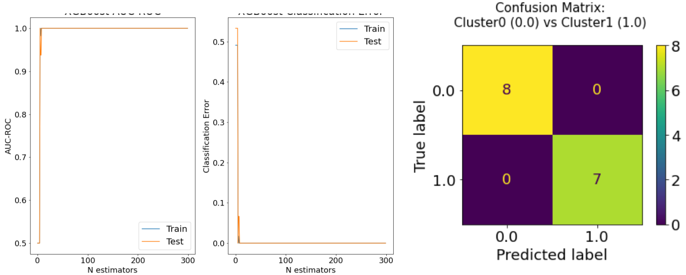
General exploration: Binary Classification [KEGG data]

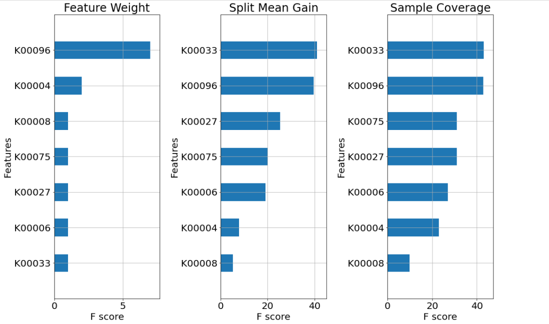


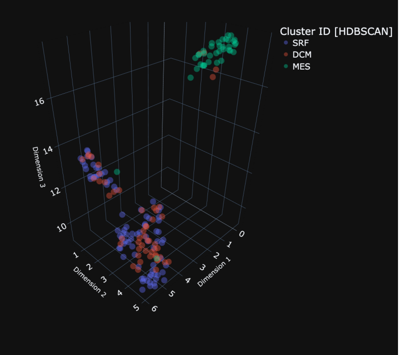
Polar Data:
(42, 9027)
SRF 22
DCM 10
MES 10

Non Polar Data:
(131, 9027)
SRF 61
DCM 42
MES 28

Cluster 0 vs Cluster 1 (mes vs polar)

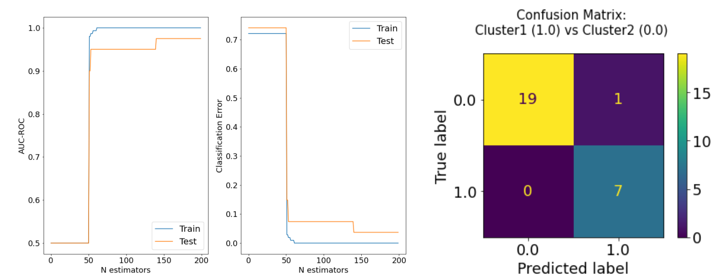


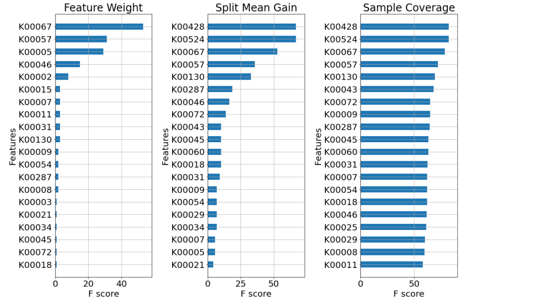


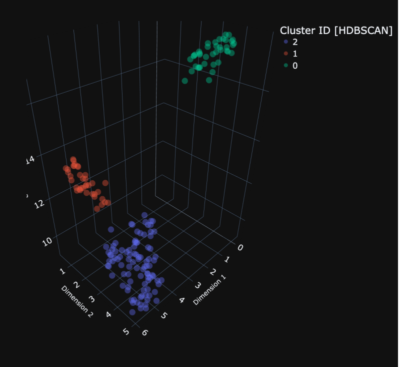


All Data:
(173, 9073)
SRF 83
DCM 52
MES 38

Cluster 1 vs Cluster 2 (polar vs non polar)

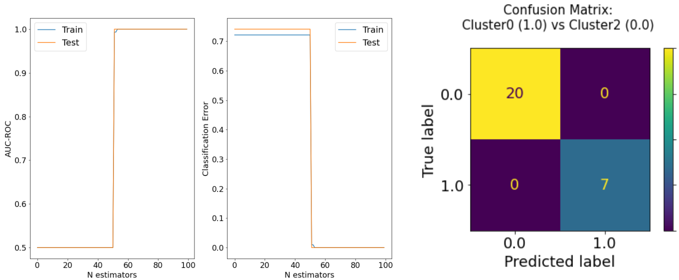


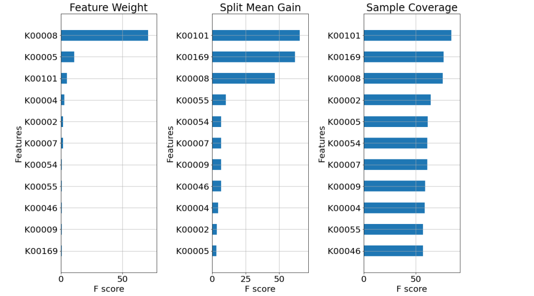




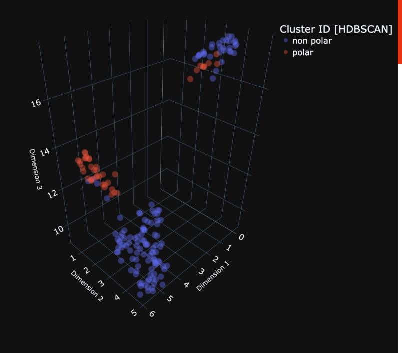
Cluster0	36
Cluster1	32
Cluster2	95

Cluster 0 vs Cluster 2 (mes vs non polar)





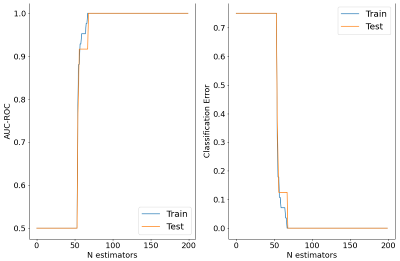
General exploration: Binary Classification [KEGG data]

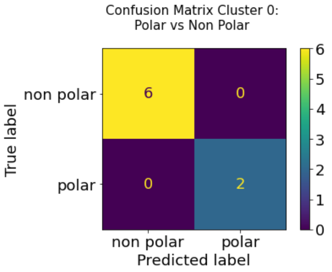


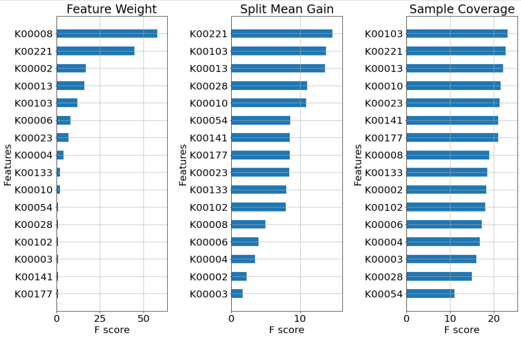
Polar Data:
(42, 9027)
SRF 22
DCM 10
MES 10

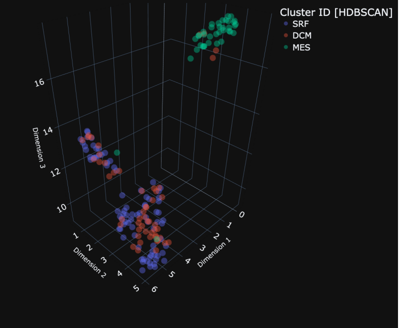
Non Polar Data:
(131, 9027)
SRF 61
DCM 42
MES 28

Cluster 0 (mes)



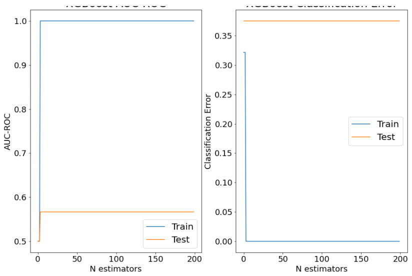


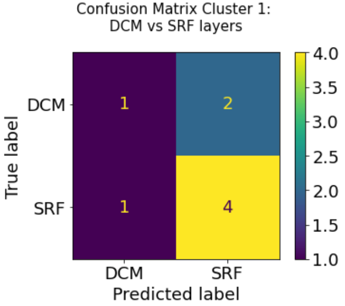


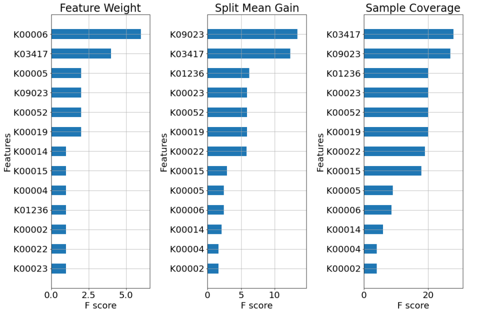


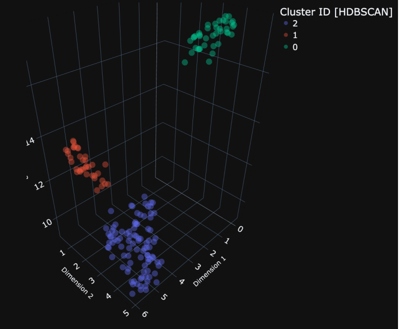
All Data:
(173, 9073)
SRF 83
DCM 52
MES 38

Cluster 1 (polar)



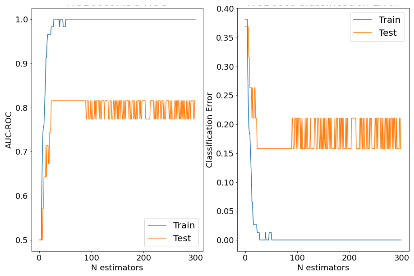


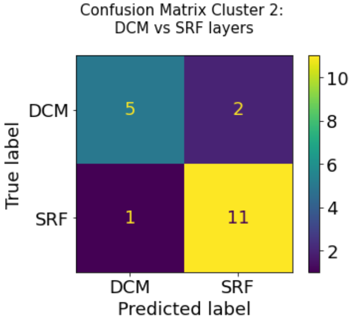


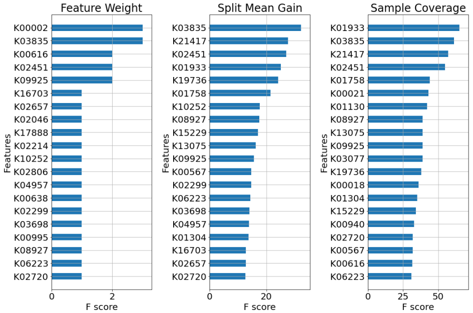


Cluster0	36
Cluster1	32
Cluster2	95

Cluster 2 (non polar)





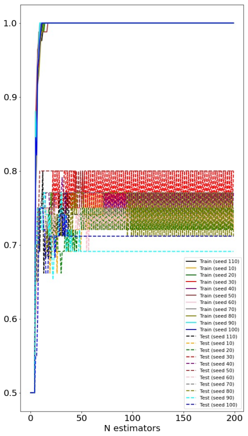


Layers characterization - Comparison 1: SRF vs DCM (5 vs 55 m depth)

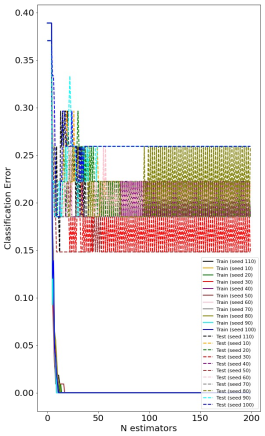
No parametric test: Wilcoxon, adjusted p value via fdr.

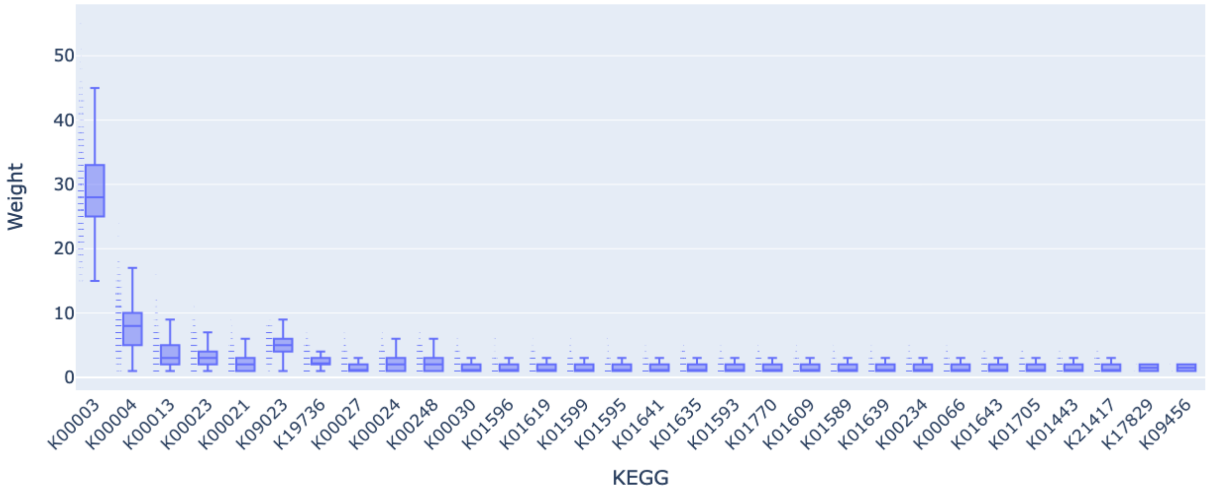
1808/9027 features with adj. P val < 0.05 as input for XGBoost hyperparameter tuning.

SRF vs DCM [all data] (significant KOs)

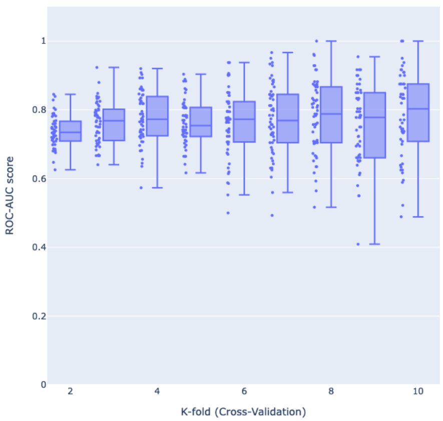


SRF vs DCM [all data] (significant KOs)

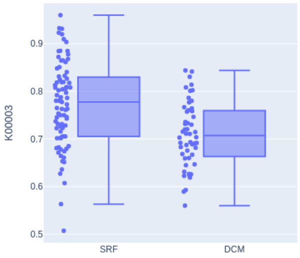




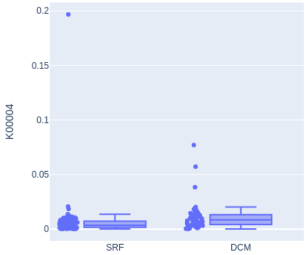
Classification Performance (SRF vs DCM)



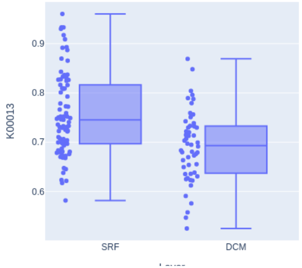
Homoserine dehydrogenase



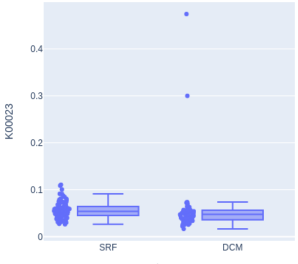
Butanediol dehydrogenase



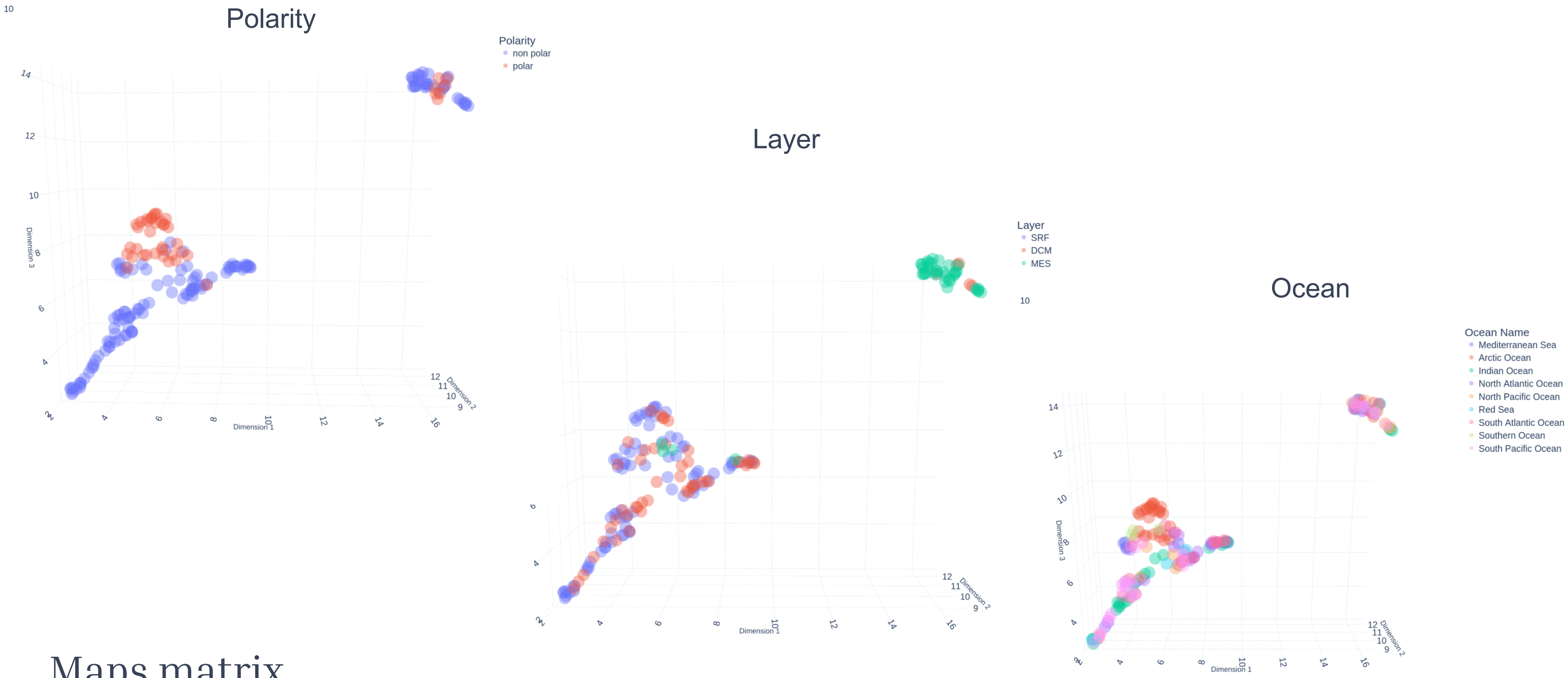
histidinol dehydrogenase



Acetoacetyl-CoA reductase



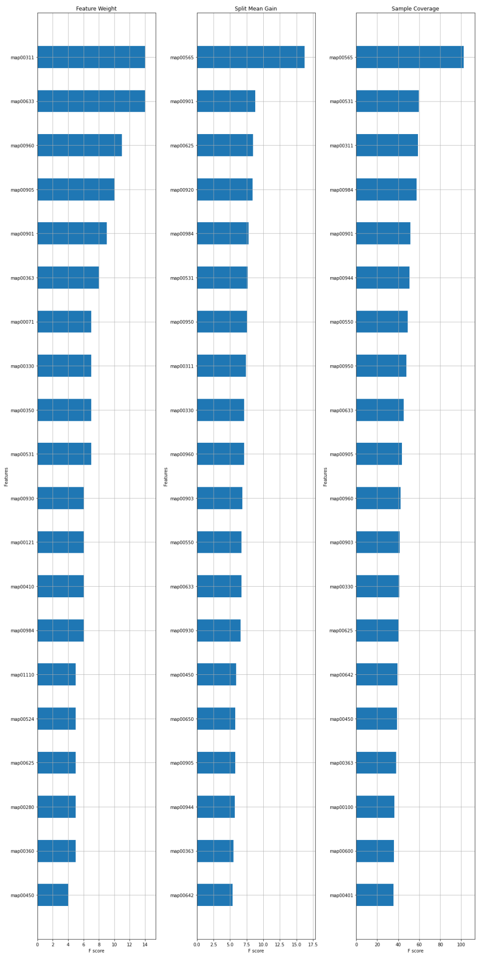
Transcription Factors regulating KEGG pathways



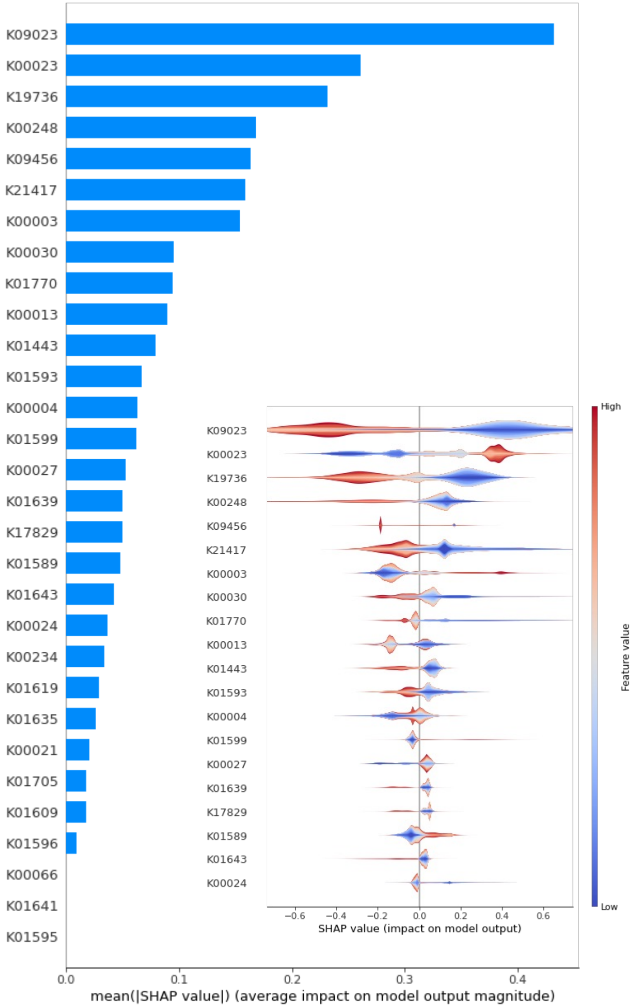
Maps matrix
173 x 163

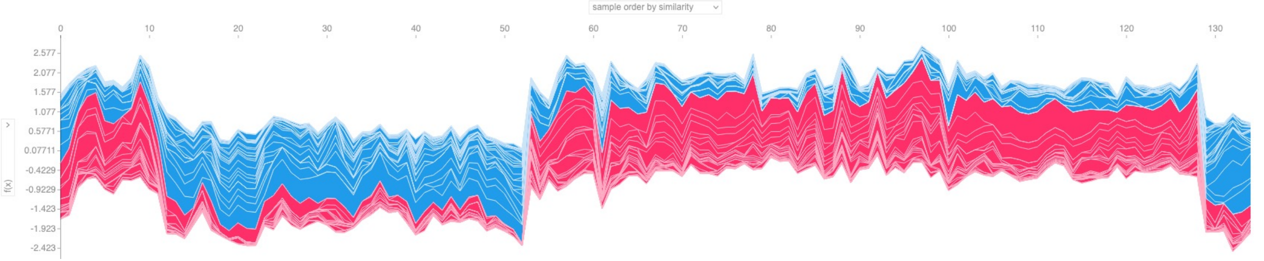
Layers characterization (TF) - Comparison 1: SRF vs DCM

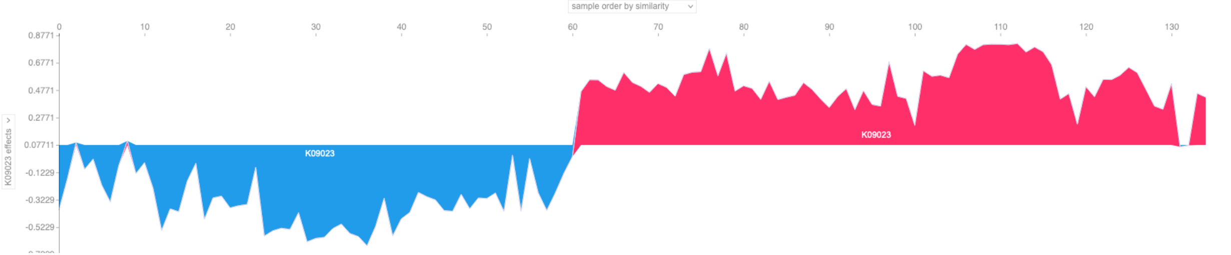
Map00311: Penicillin and cephalosporin biosynthesis
Map00633: Nitrotoluene degradation
Map00960: Tropane, piperidine and pyridine alkaloid biosynthesis
Map00905: Brassinosteroid biosynthesis
Map00901: Indole alkaloid biosynthesis
Map00363: Bisphenol degradation
Map00071: Fatty acid degradation
Map00330: Arginine and proline metabolism
Map00350: Tyrosine metabolism
Map00930: Caprolactam degradation
Map00121: Secondary bile acid biosynthesis
Map00410: beta-Alanine metabolism
Map00984: Steroid degradation
Map01110: Biosynthesis of secondary metabolites
Map00524: Neomycin, kanamycin and gentamicin biosynthesis
Map00625: Chloroalkane and chloroalkene degradation
Map00280: Valine, leucine and isoleucine degradation
Map00360: Phenylalanine metabolism
Map00450: Selenocompound metabolism

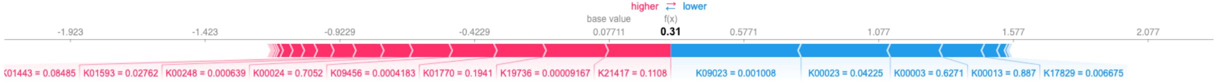


Layers characterization - Comparison 1: SRF vs DCM (5 vs 55 m depth)

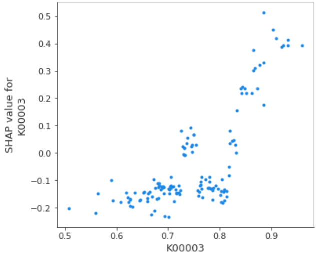


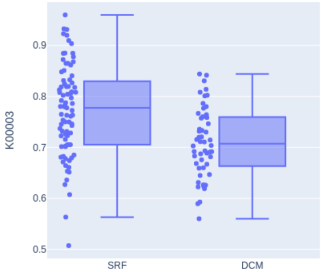




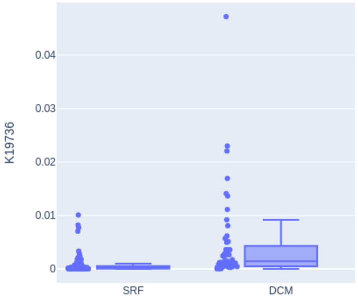


Aminoacrylate hydrolase (pyrimidine metabolism)





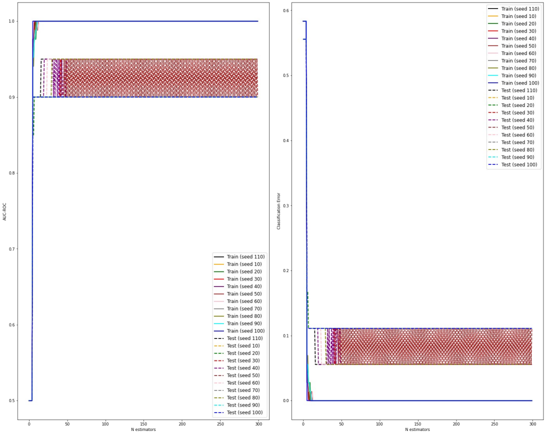
TetR/AcrR family transcriptional regulator

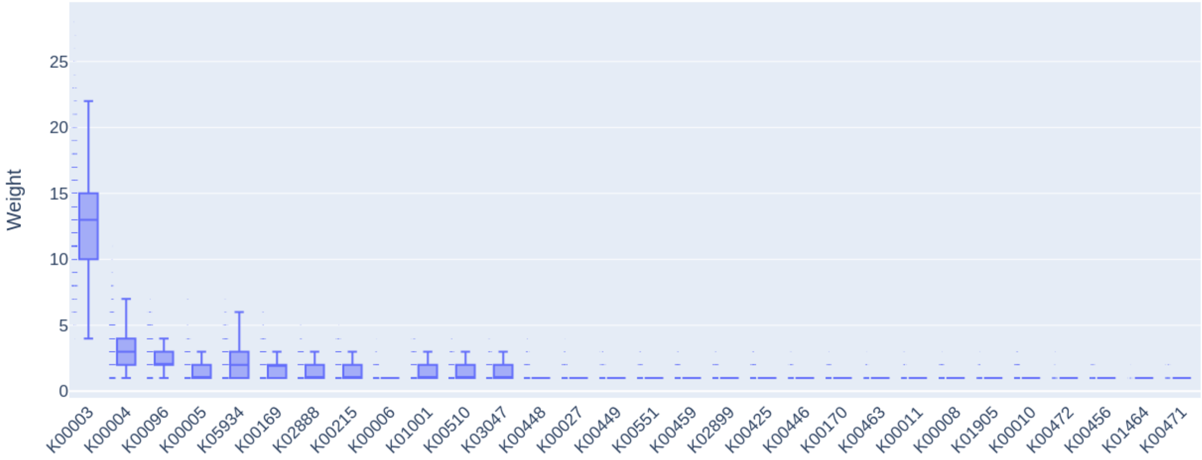


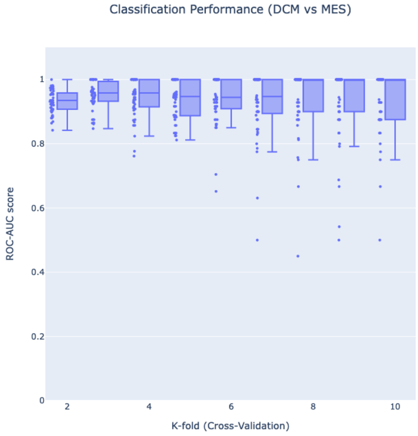
Layers characterization - Comparison 2: DCM vs MES (5 vs 550 m depth)

No parametric test: Wilcoxon, adjusted p value via fdr.

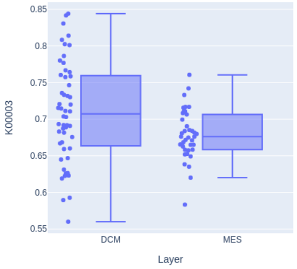
6937/9027 features with adj. P val < 0.05 as input for XGBoost hyperparameter tuning.



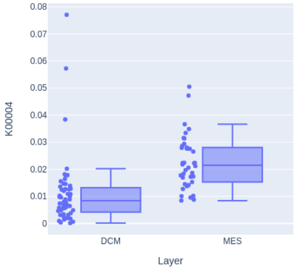




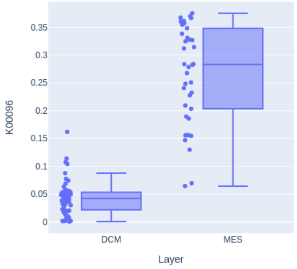
Homoserine dehydrogenase



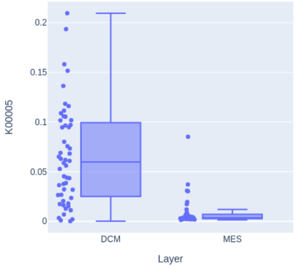
Butanediol dehydrogenase



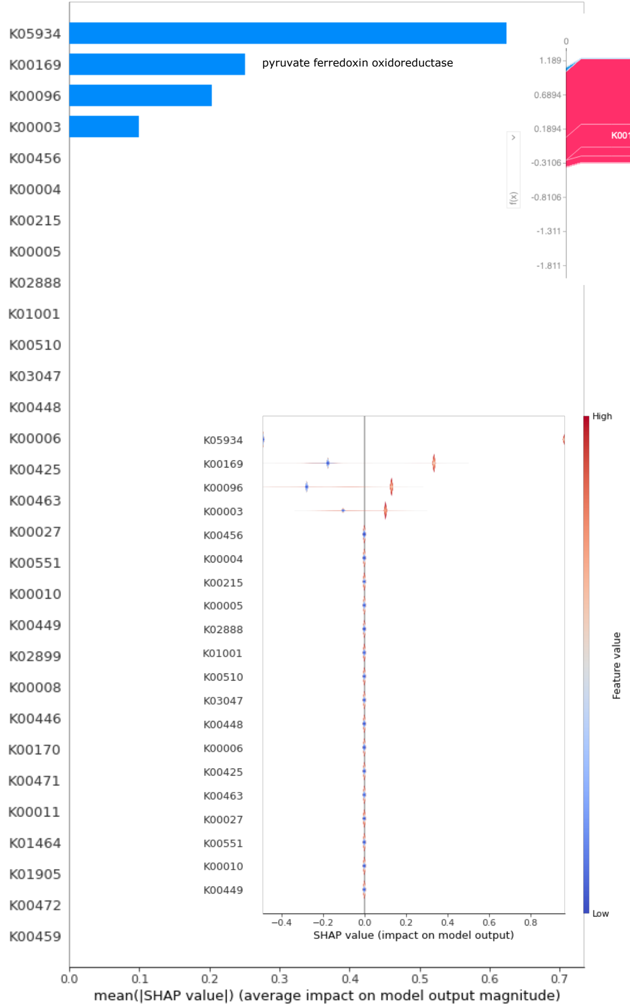
glycerol-1-phosphate dehydrogenase

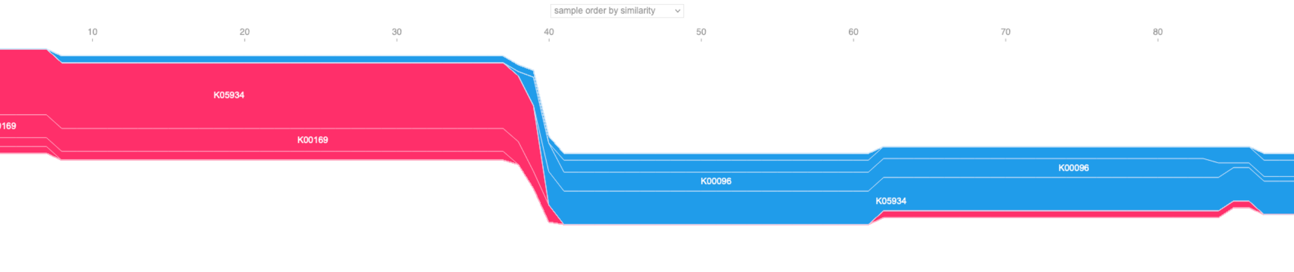


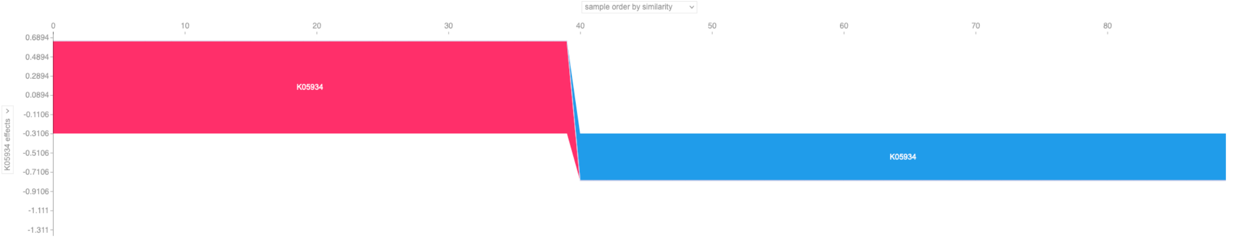
glycerol dehydrogenase

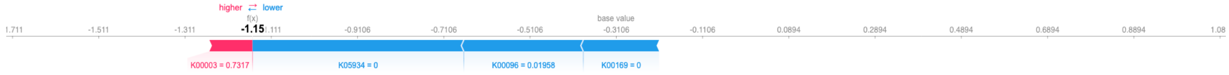


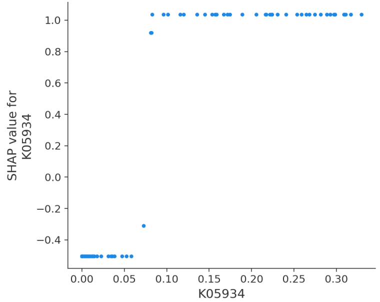
Layers characterization - Comparison 2: DCM vs MES (5 vs 550 m depth)



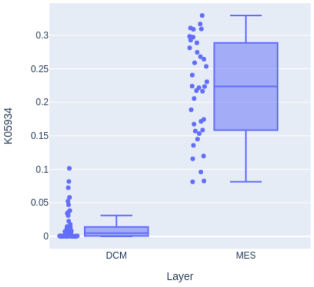








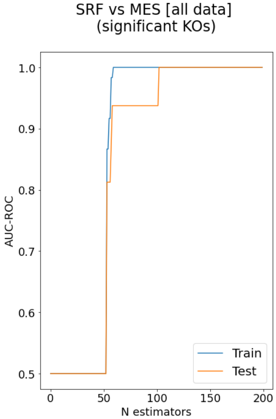
Cobalamin biosynthesis

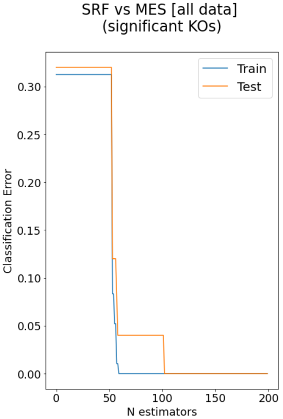


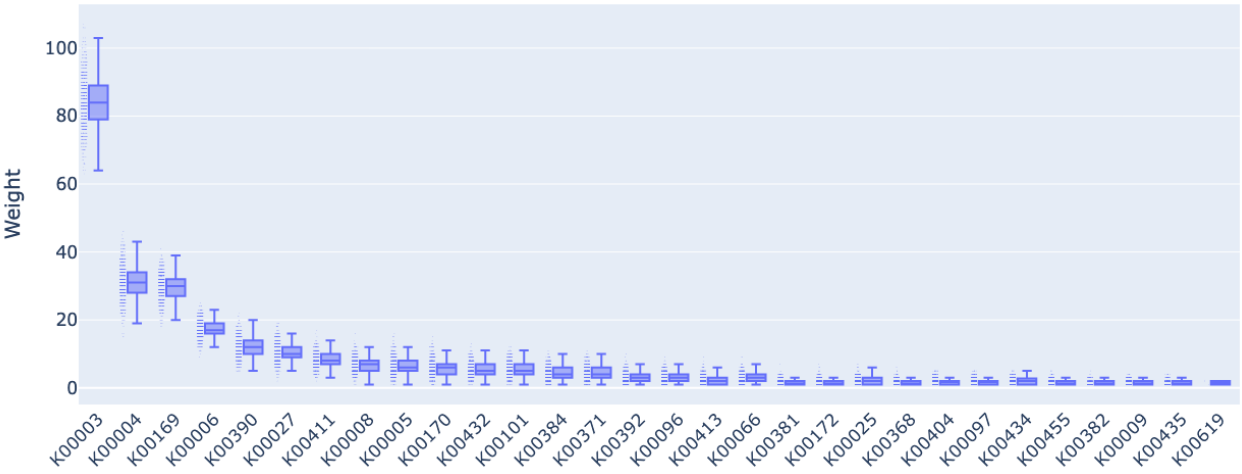
Layers characterization - Comparison 3: SRF vs MES (5 vs 550 m depth)

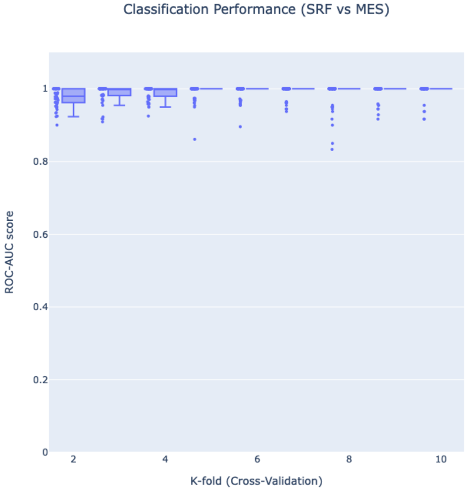
No parametric test: Wilcoxon, adjusted p value via fdr.

7517/9027 features with adj. P val < 0.05 as input for XGBoost hyperparameter tuning.

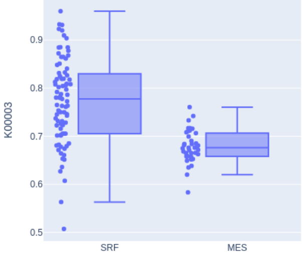




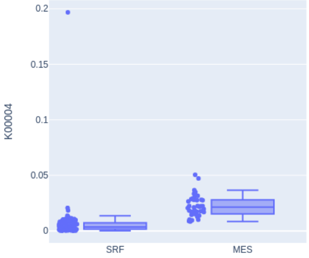




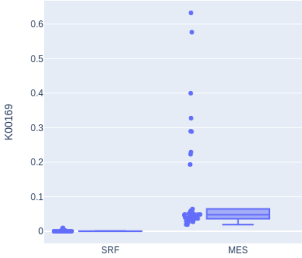
Homoserine dehydrogenase



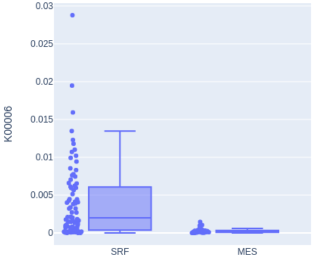
Butanediol dehydrogenase



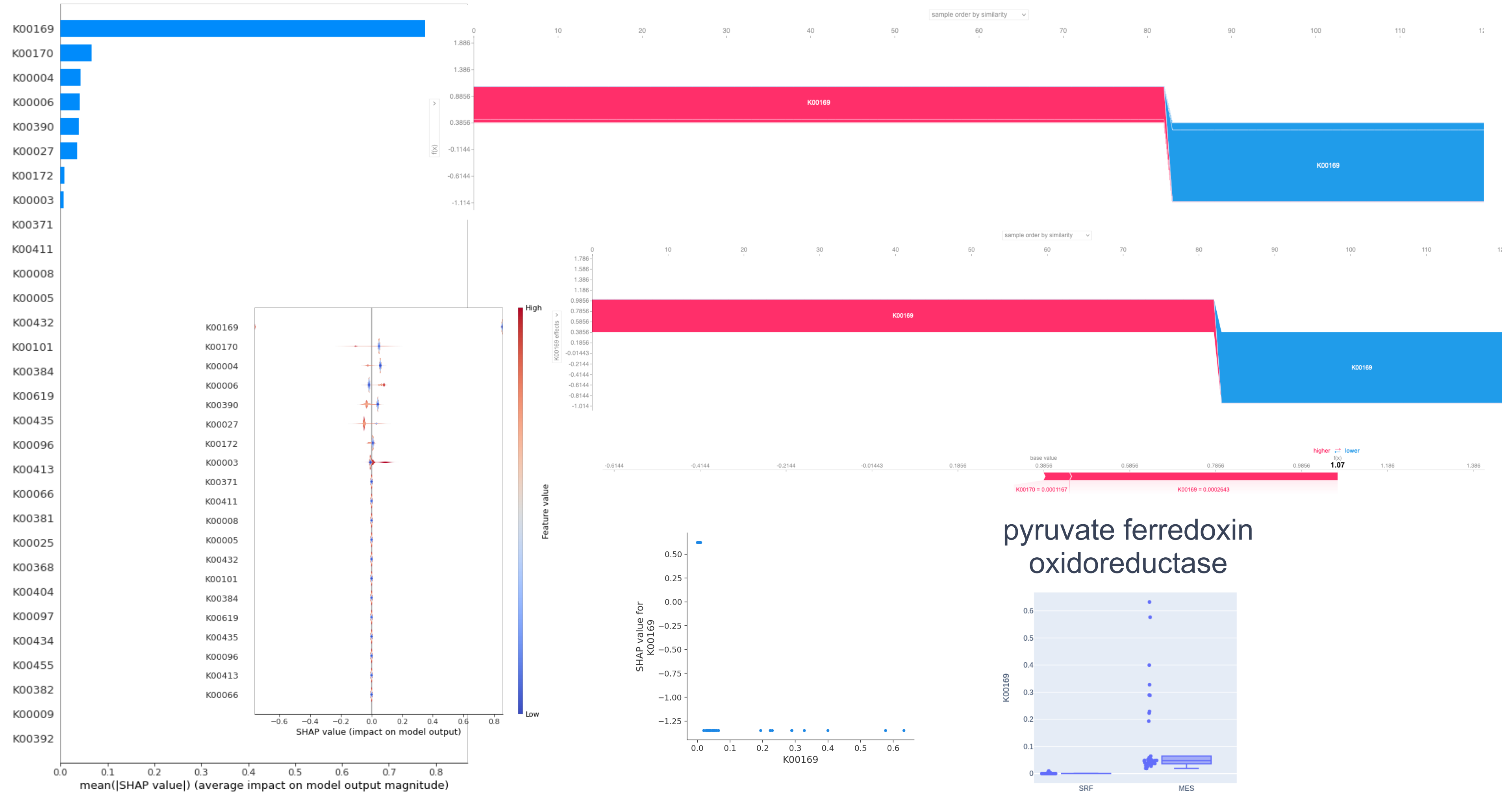
pyruvate ferredoxin oxidoreductase



glycerol-3-phosphate dehydrogenase

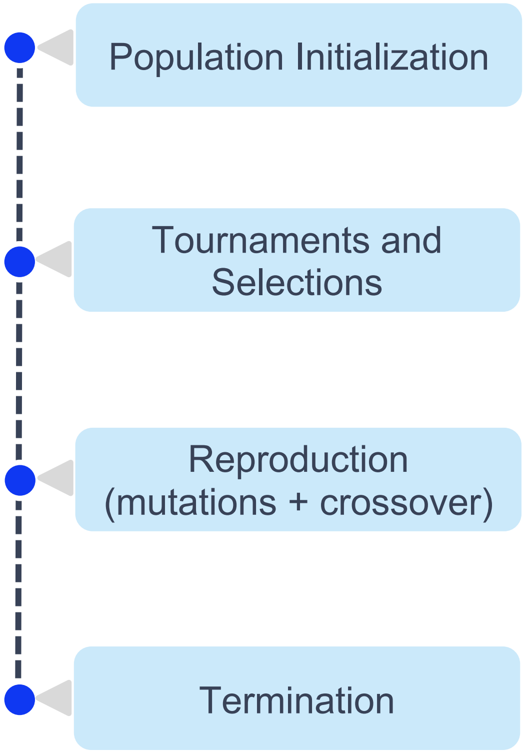


Layers characterization - Comparison 3: SRF vs MES (5 vs 550 m depth)



pyruvate ferredoxin oxidoreductase

Genetic Programming-based Symbolic Regressions



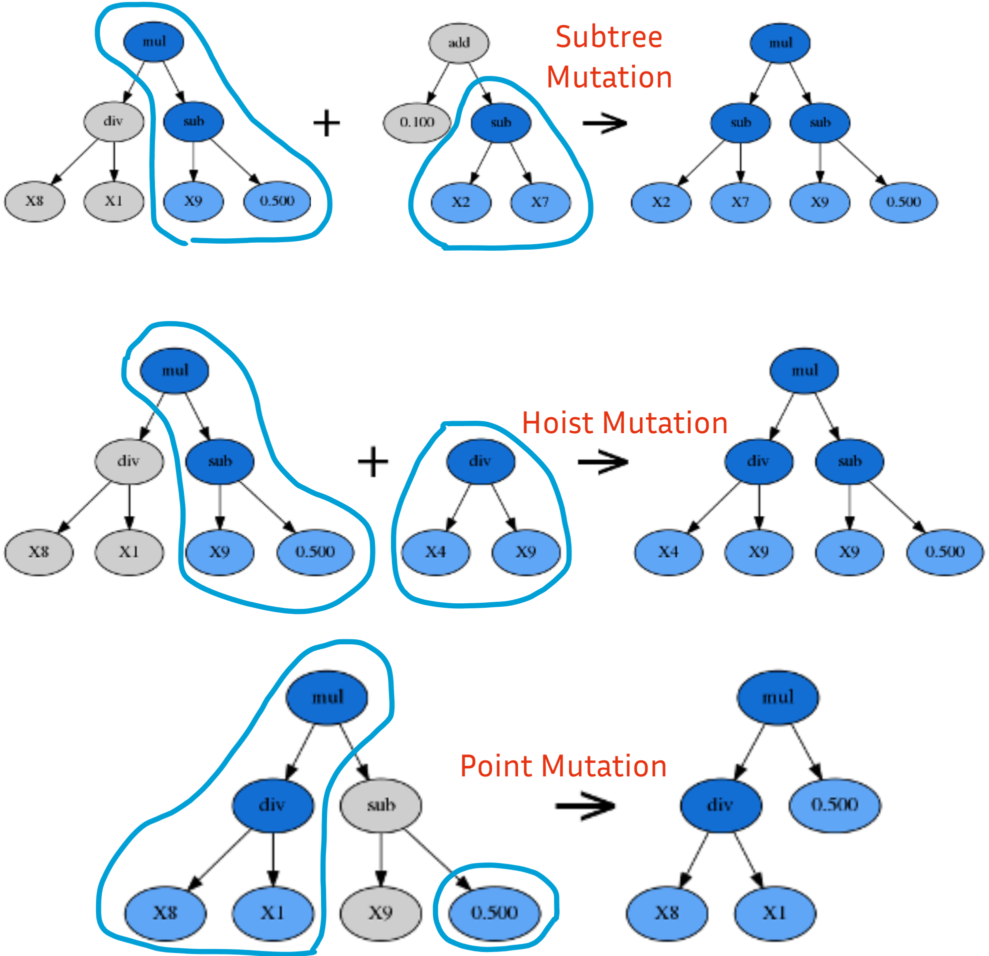
Population Initialization

Tournaments and Selections

Reproduction (mutations + crossover)

Termination

$Y = (X8/X1) * (X9 - 0.500)$

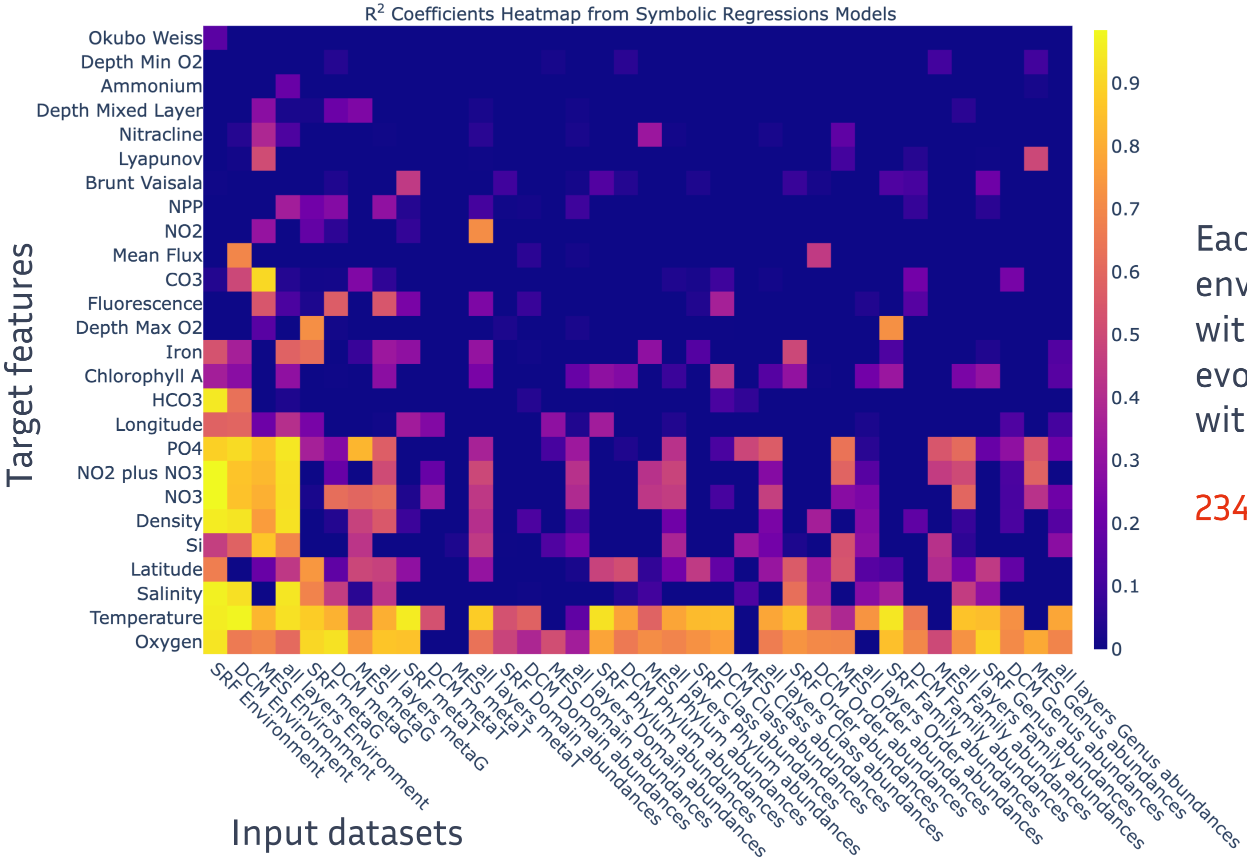


Subtree Mutation

Hoist Mutation

Point Mutation

Genetic Programming-based Symbolic Regressions

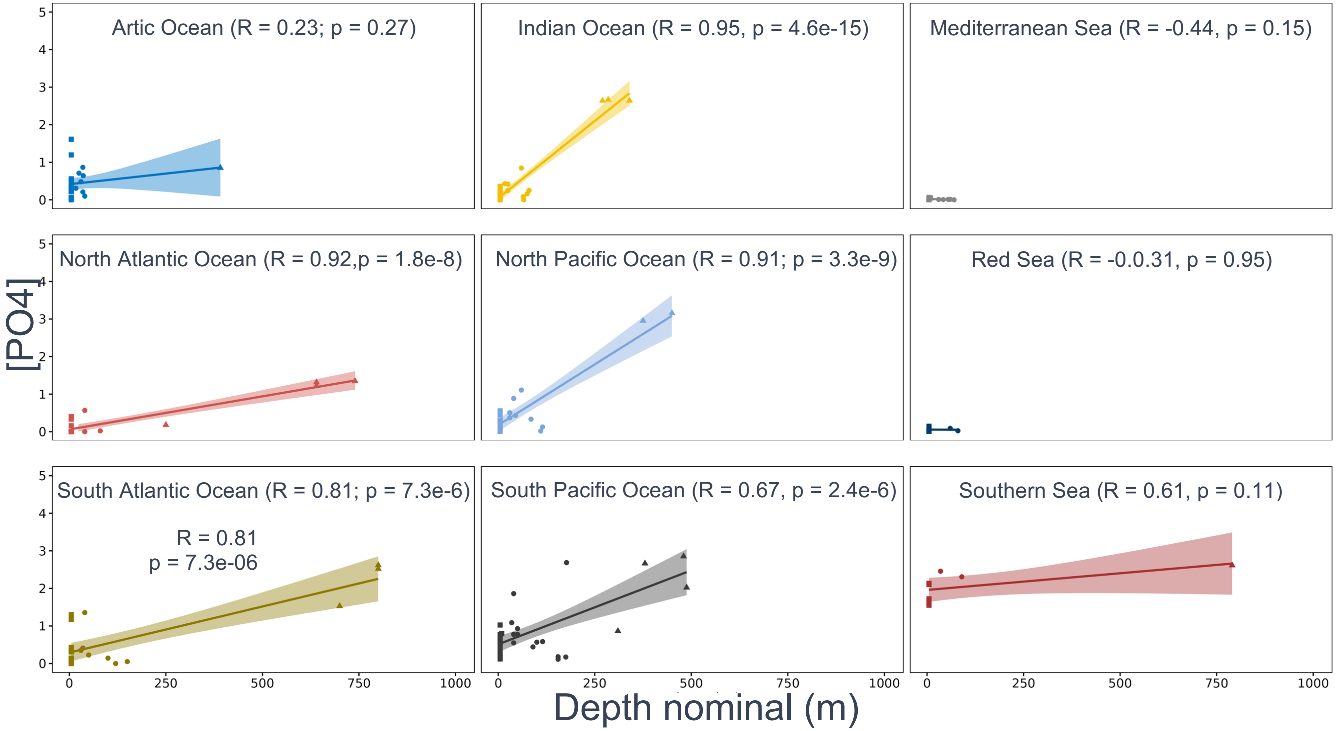


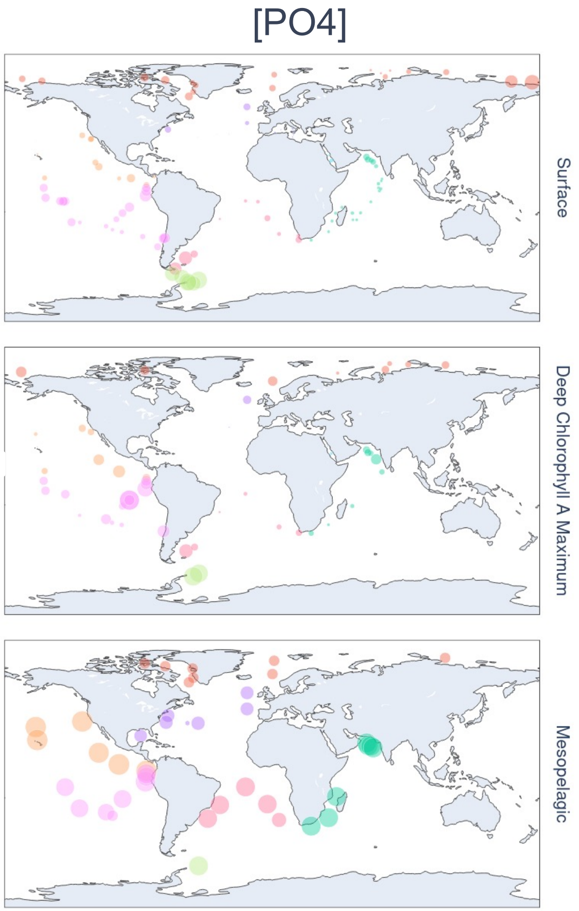
Each dataset was used to predict one environmental variable at a time starting with a **population size of 20000**, and evolving the models during **20 generations** with a crossover probability of 0.65.

234 models

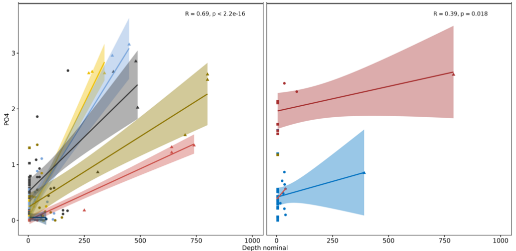
$$Target_i = SymReg_{GeneticProgramming}(dataset_j)$$

Exploration: Phosphate concentration vs Depth nominal (m)





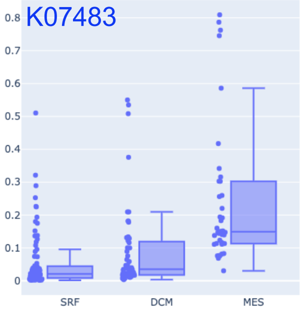
Non Polar:
Indian Ocean
Mediterranean Sea
North Atlantic Ocean
North Pacific Ocean
Red Sea
South Atlantic Ocean
South Pacific Ocean



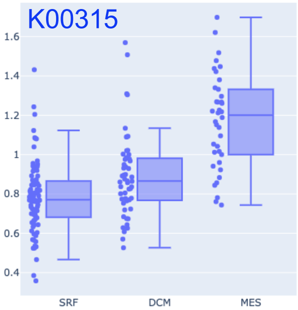
Polar:
Southern Ocean

Artic Ocean

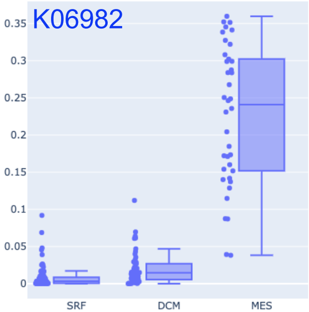
PO4 prediction from metaG dataset



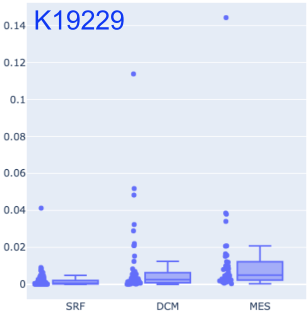
Replication and repair (Enzyme Transposase)



DMGDH; dimethylglycine dehydrogenase (Enzyme Oxidoreductases)



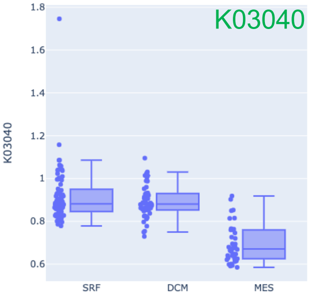
Pantothenate and CoA biosynthesis (Fatty acid, secondary metabolites and cofactors metabolism)



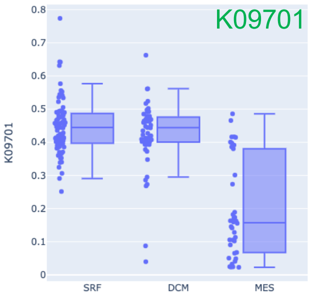
Cationic antimicrobial peptide (CAMP) resistance

PO4 = K00315 - K03040 + K07483 - K09701 + log(log(K06982*K19229))

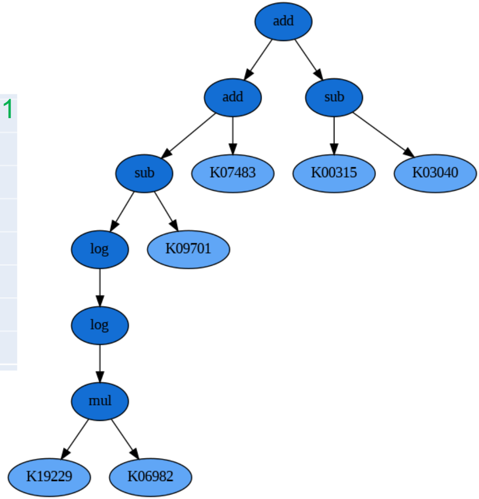
(R^2: 0.66)

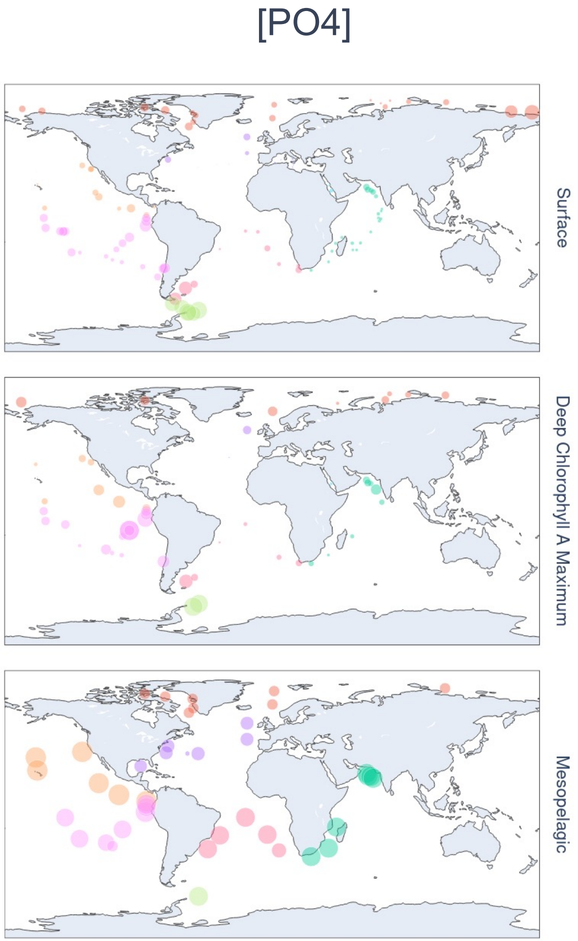


DNA-directed RNA polymerase subunit Alpha (Enzyme Transferases)



Uncharacterized protein





Ocean Name

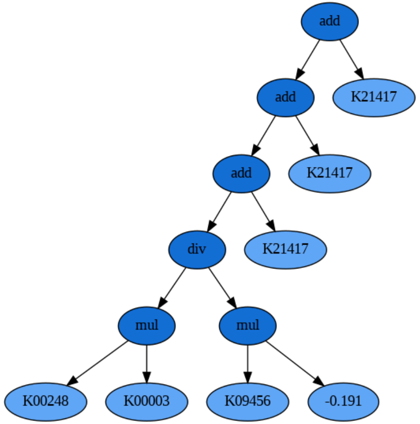
- ● Mediterranean Sea
- ● Arctic Ocean
- ● Indian Ocean
- ● North Atlantic Ocean
- ● North Pacific Ocean
- ● Red Sea
- ● South Atlantic Ocean
- ● Southern Ocean
- ● South Pacific Ocean

Layers characterization [KEGG] - SRF Layer (5 m depth)

Input: 'K09023', 'K00023', 'K19736', 'K00248','K09456','K21417','K00003'
Output: 1 environmental feature.

Symbolic Regressions: population_size=20000, generations=20, p_crossover=0.65

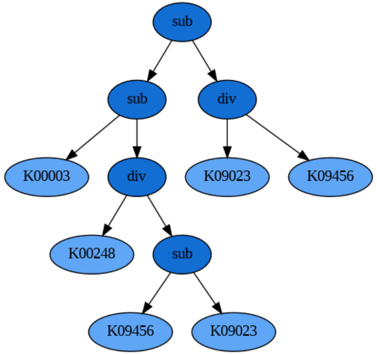
Latitude



log((K09456**0.5/(-K00003*K00248/K09456 + K21417))**0.5)

(R^2: 0.6600)

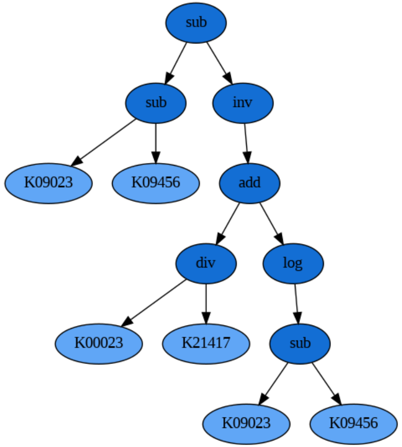
Temperature



log((K09456**0.5/(-K00003*K00248/K09456 + K21417))**0.5)

(R^2: 0.6688)

Salinity



log((K09456**0.5/(-K00003*K00248/K09456 + K21417))**0.5)

(R^2: 0.5883)

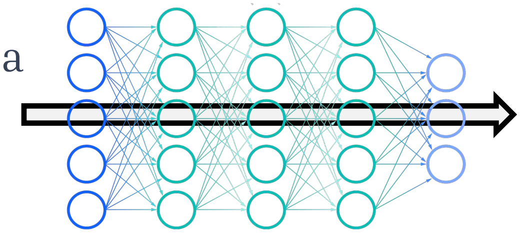
Exploring telemonitoring metaG composition from environment

Meta-Genomic data
R^{9024}

(direct transform)

R^{3} 3D embedding

Environmental data
R^{29}



R^{3} 3D embedding reconstruction

3D embedding reconstruction
R^{3}

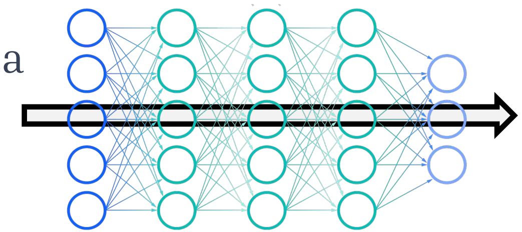
(reverse transform)

R^{9024} Meta-Genomic data reconstruction

Telemonitoring meta-genomic composition from environment

Environmental data

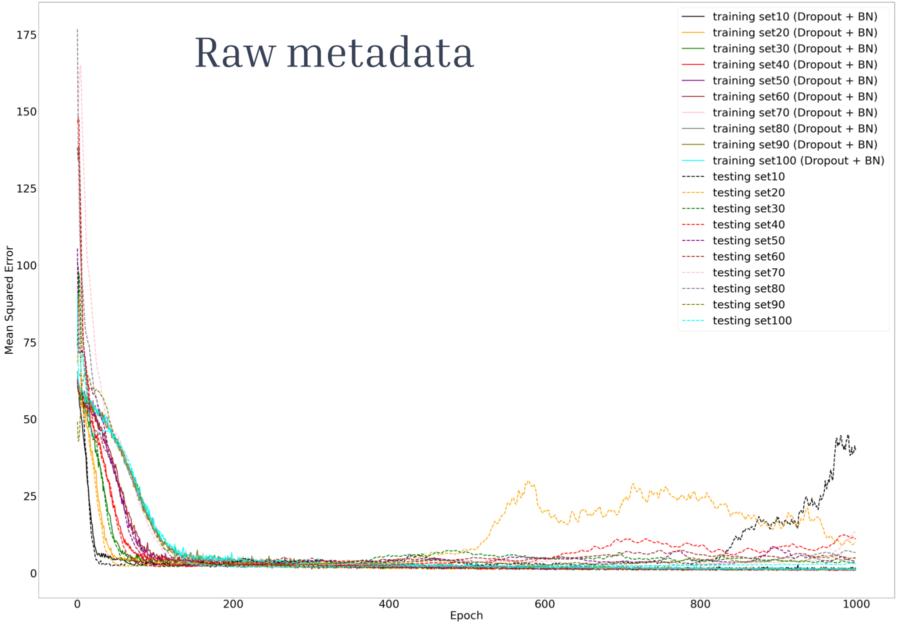
R^{29}



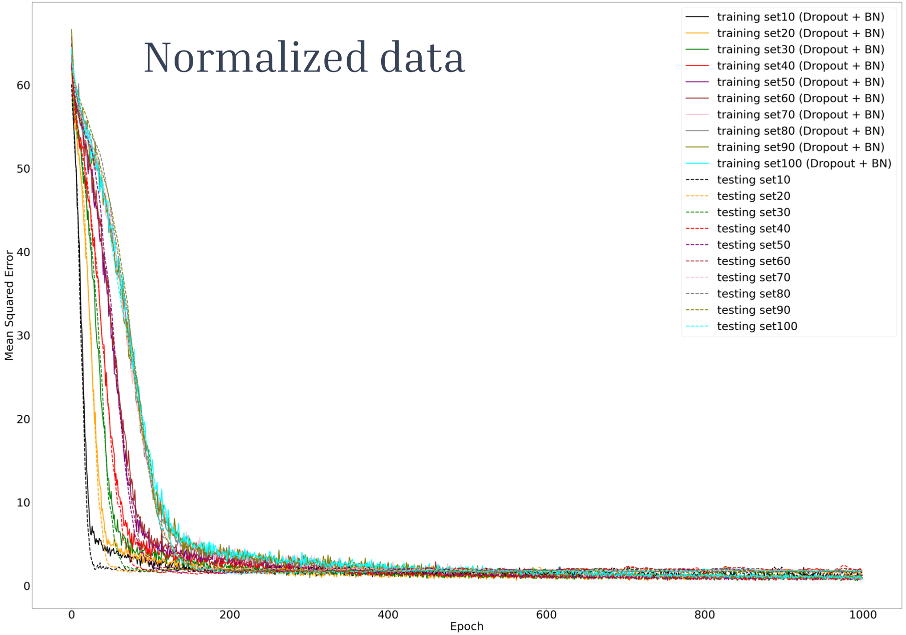
R^3

3D embedding

reconstruction

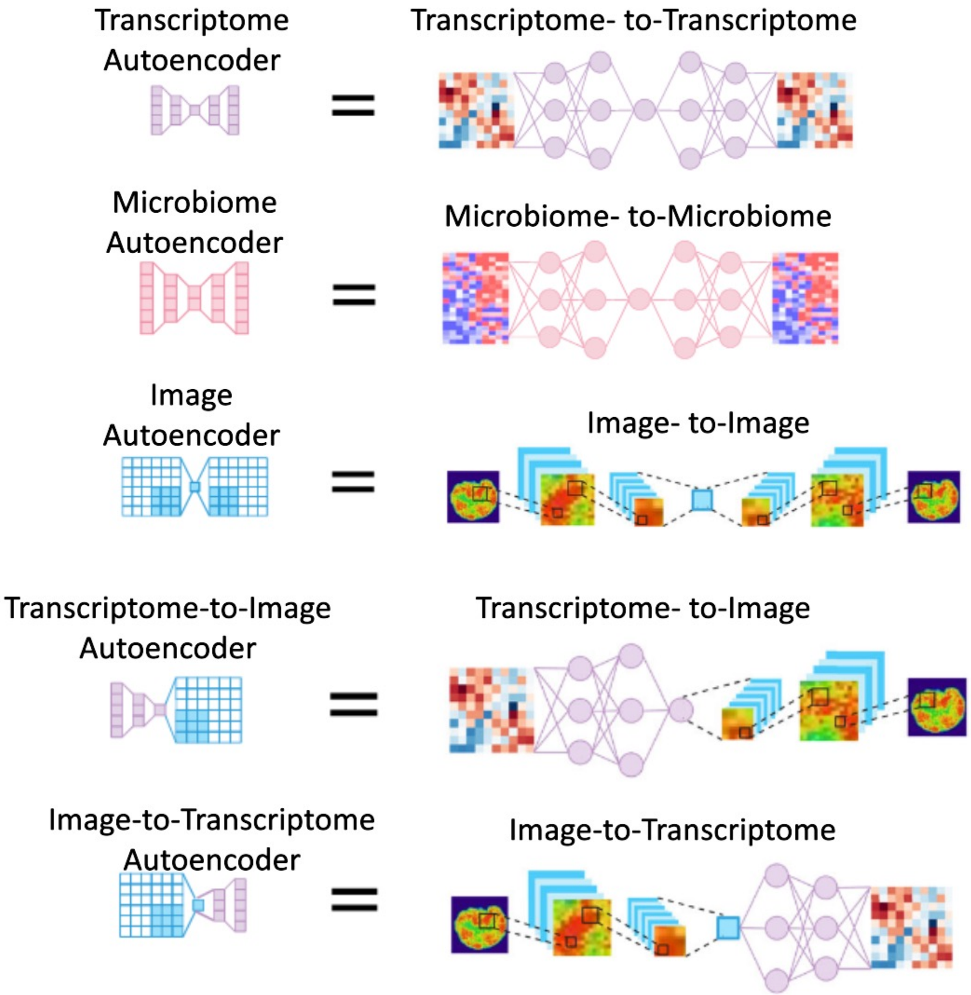


Raw metadata



Normalized data

Moving towards cross-modal autoencoders


Transcriptome Autoencoder
=
Transcriptome- to-Transcriptome
Microbiome Autoencoder
=
Microbiome- to-Microbiome
Image Autoencoder
=
Image- to-Image
Transcriptome-to-Image Autoencoder
=
Transcriptome- to-Image
Image-to-Transcriptome Autoencoder
=
Image-to-Transcriptome

Challenges for implementations of generative models

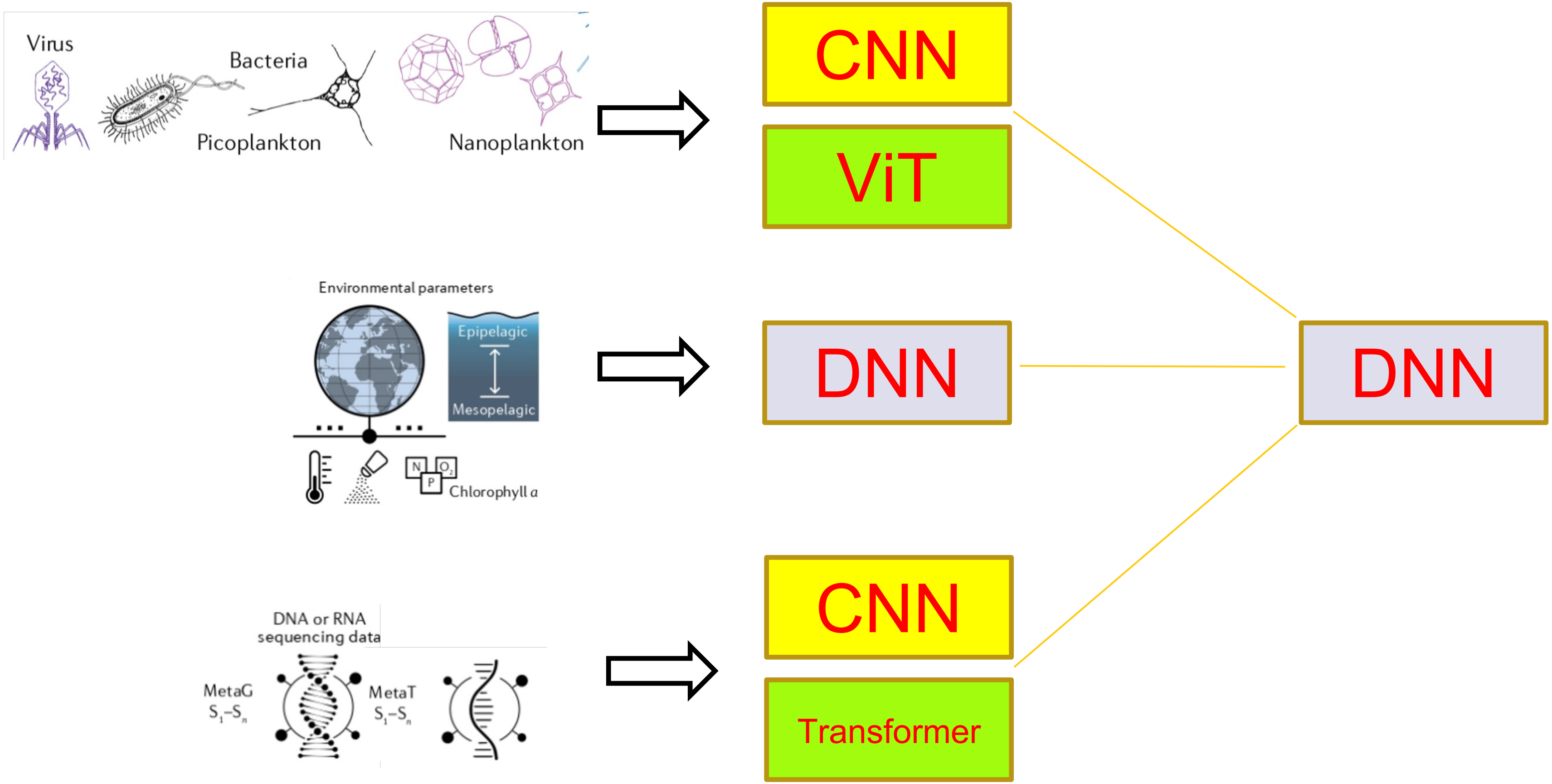
Data augmentation estrategy:

Variational autoencoders

Optimal Transport

Edited and adapted from Yang, K. D., Belyaeva, A., Venkatachalapathy, S. (2021). Multi-domain translation between single-cell imaging and sequencing data using autoencoders. Nature Communications, 12, 31. doi: 10.1038/s41467-020-20249-2.

Perspectives: integration with other members of the team



CNN

ViT

DNN

CNN

Transformer

DNN

Thank you!