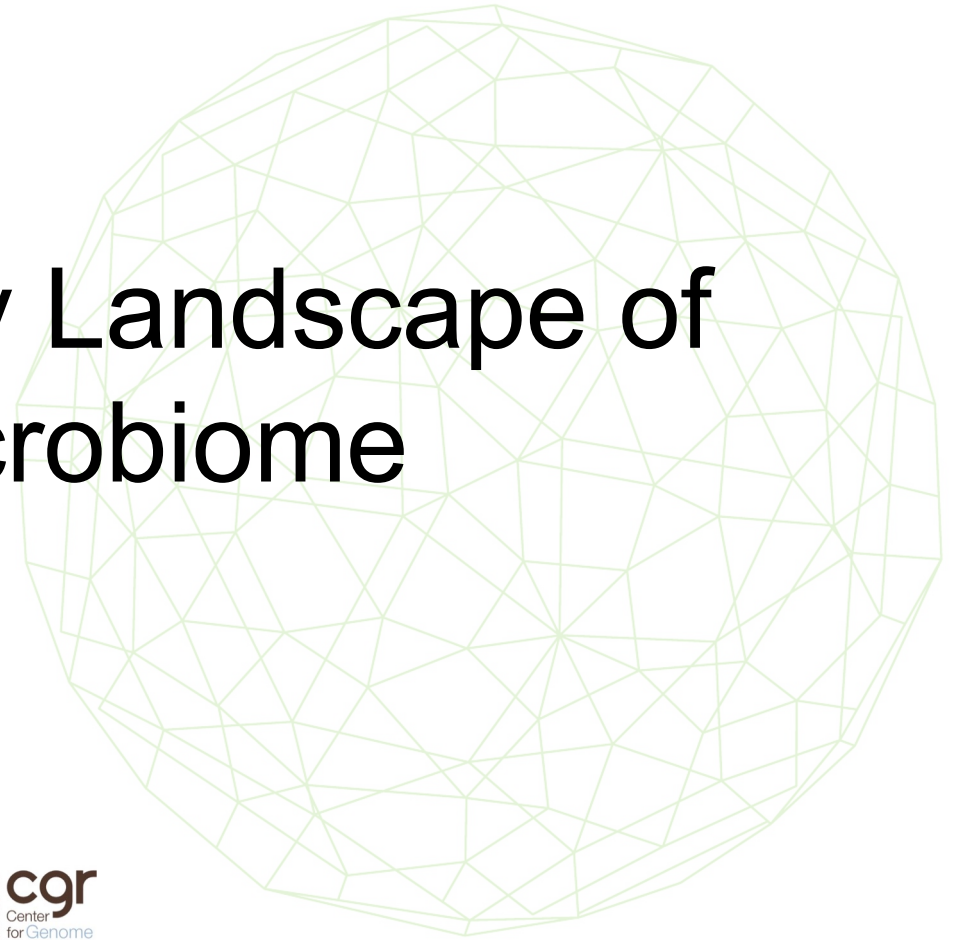
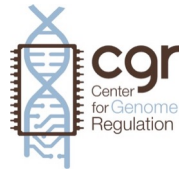


Gene Regulatory Landscape of Ocean Microbiome

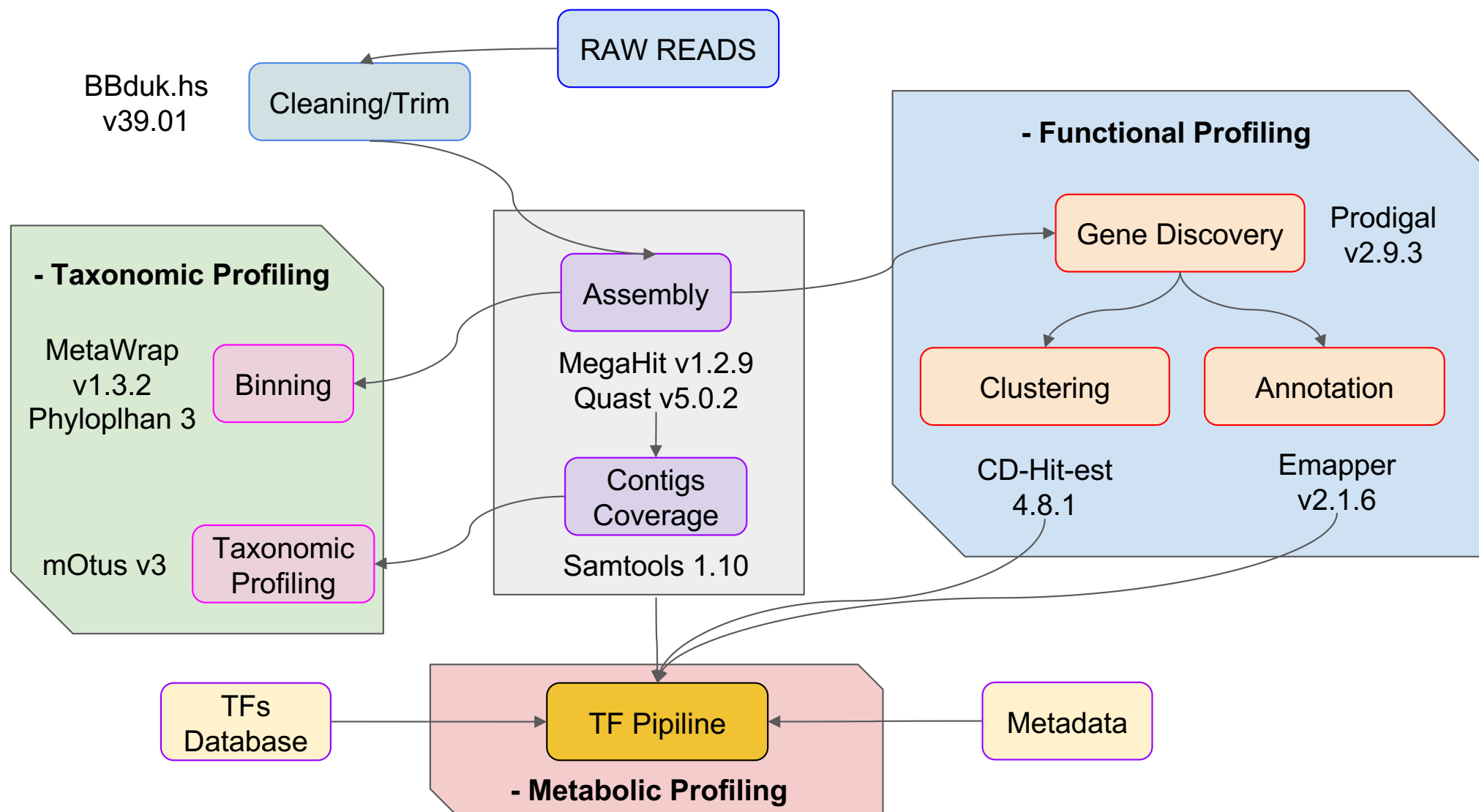


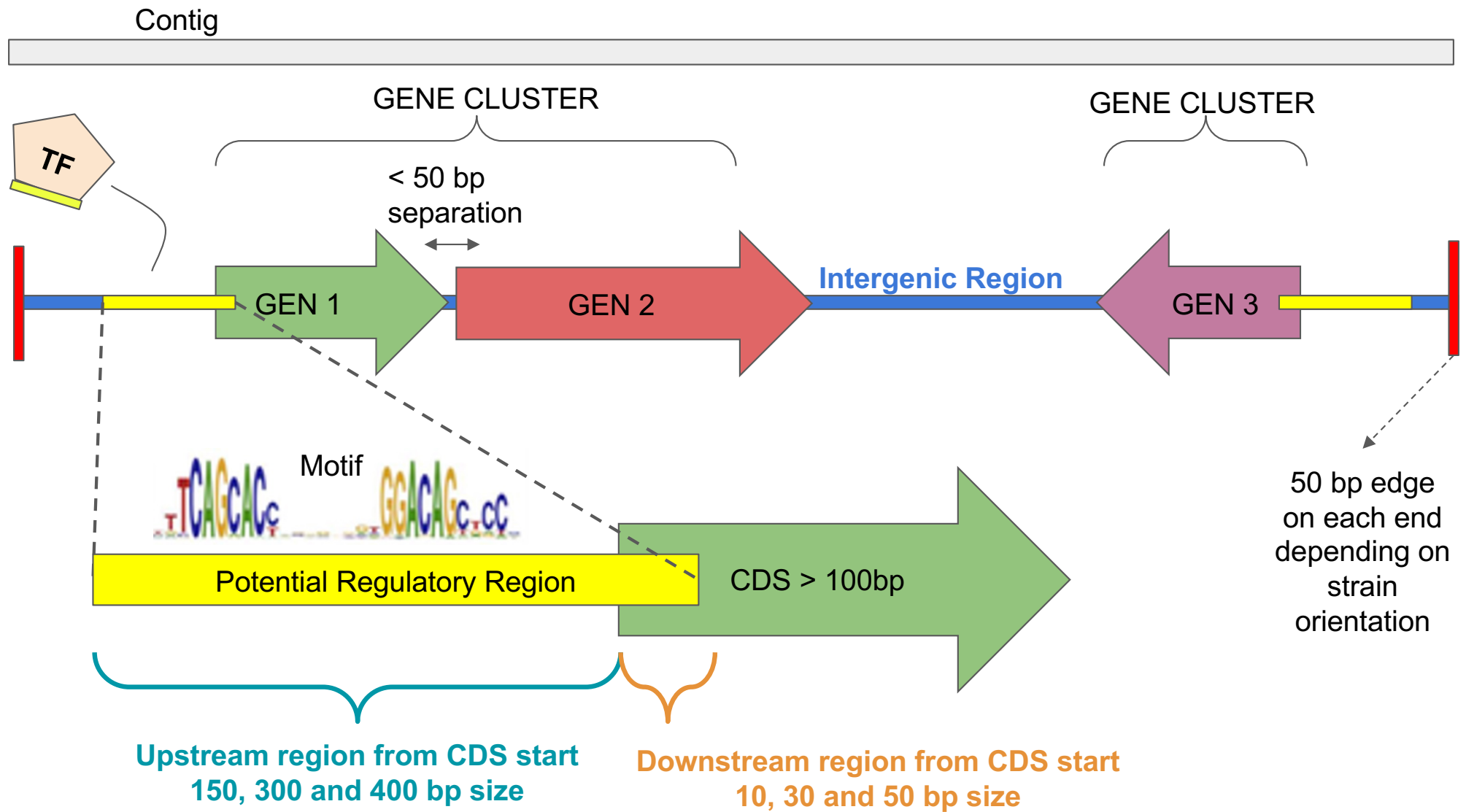
UNIVERSIDAD DE CHILE



Creating a resource ...







- Matrix Creation Pipeline thru

Potential Regulatory Region

1. Obtain all possible potential regulatory regions



In house Python Script

2. Selection of Motifs

Regulon Database



motifs_complete	tf family		desc	general
AfrR_Rhizobiales	AfrR	LacI	1,5-Anhydro-d-Fructose utilization	Alcohol metabolism
AfrR_Rhodobacterales	AfrR	LacI	1,5-Anhydro-d-Fructose utilization	Alcohol metabolism
MSMEG_3606_Mycobacteriaceae	MSMEG_3606	DeoR	D-threitol utilization	Alcohol metabolism
MSMEG_3264_Mycobacteriaceae	MSMEG_3264	DeoR	Erythritol utilization; D-threitol utilization	Alcohol metabolism
EutR_Mycobacteriaceae	EutR	RpiR	Ethanolamine utilization	Alcohol metabolism
...	...	...	...	...
RbkR_Thermoproteales	RbkR	RbkR	Riboflavin biosynthesis	Vitamin metabolism
ThiR_Halobacteriales	ThiR	NaN	Thiamine transport; Thiamine biosynthesis	Vitamin metabolism
ThiR_Sulfolobales	ThiR	NaN	Thiamine transport; Thiamine biosynthesis	Vitamin metabolism

Motifs can be group by:

Transcription factor (89).

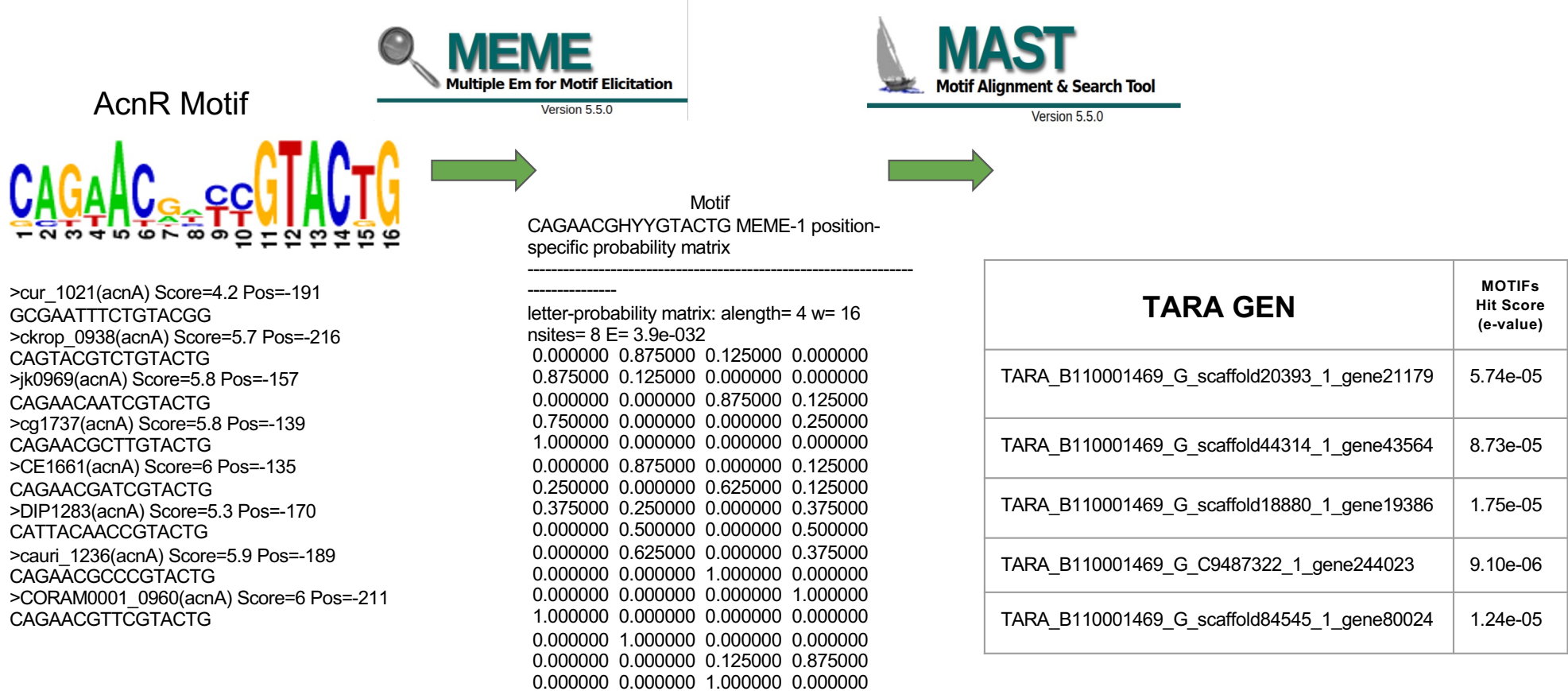
Family of Transcription factor (59)

General Regulon (695)

- Matrix Creation Pipeline thru

Potential Regulatory Region

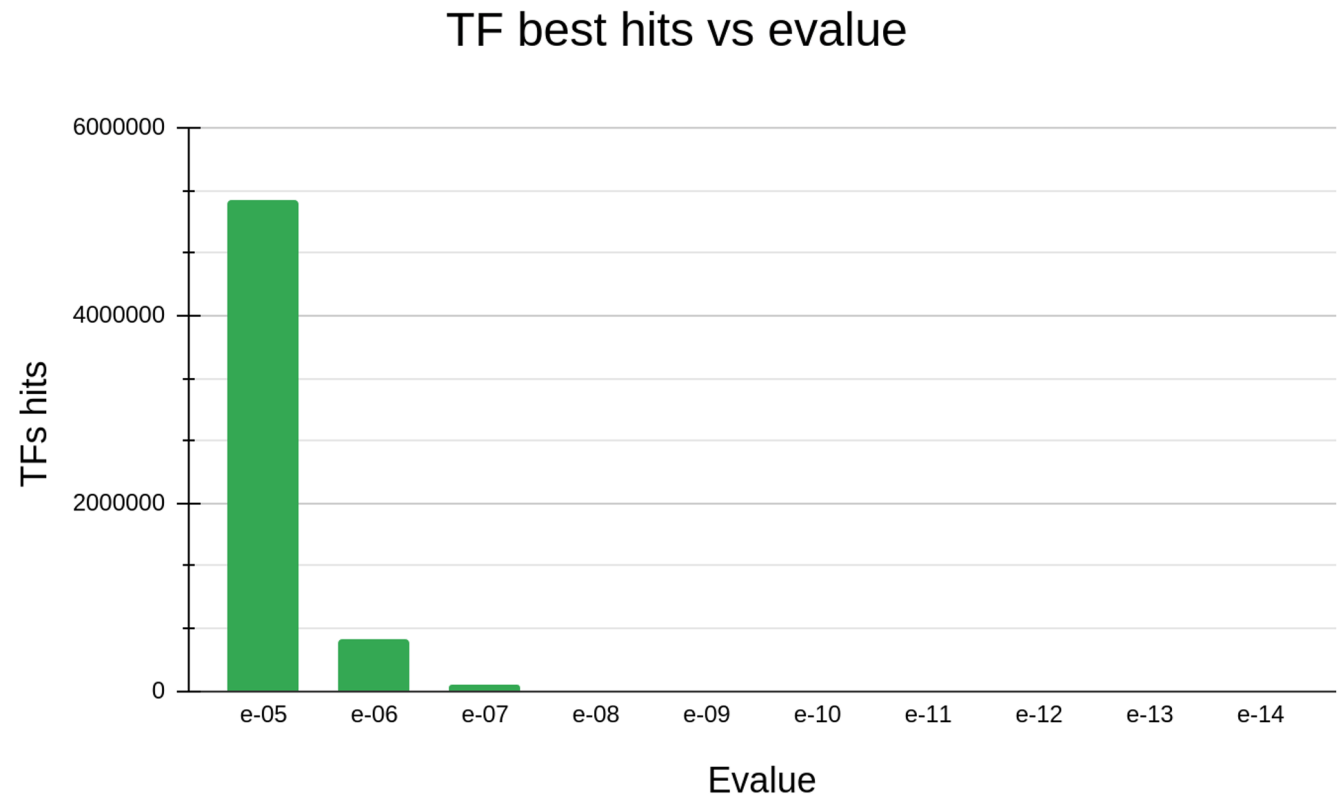
3. Calculate the probability of each motifs in the Regulatory Region



The statistics of motif search

Evalue	TF hits
e-05	5223742
e-06	548449
e-07	64231
e-08	12468
e-09	4108
e-10	2221
e-11	1128
e-12	378
e-13	91
e-14	22

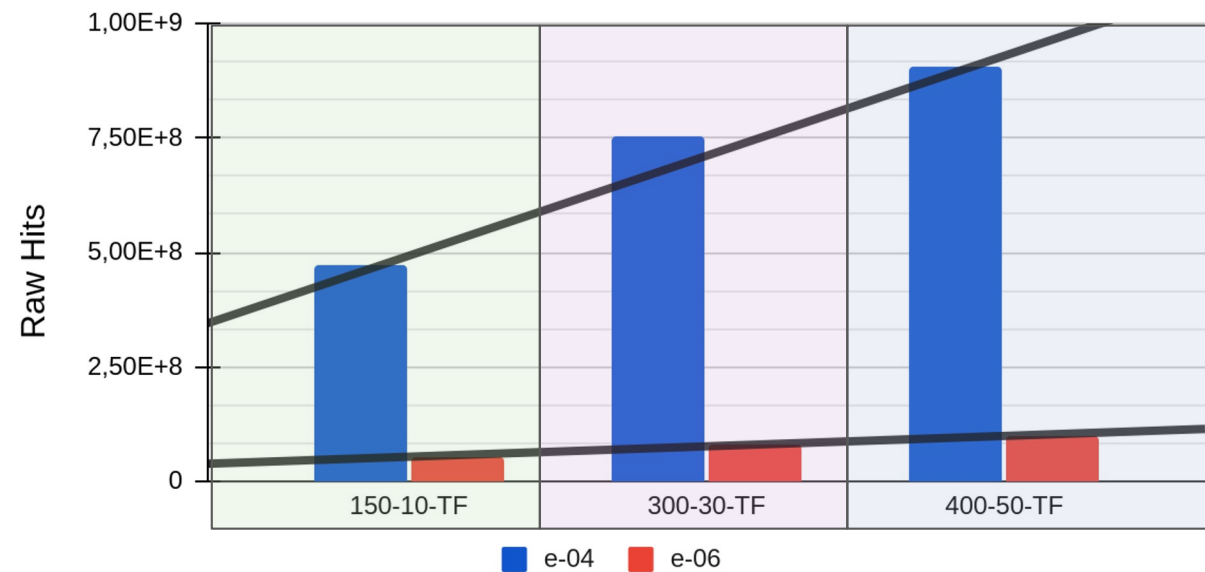
TARA_004_DCM_0.22-1.6
Statistics



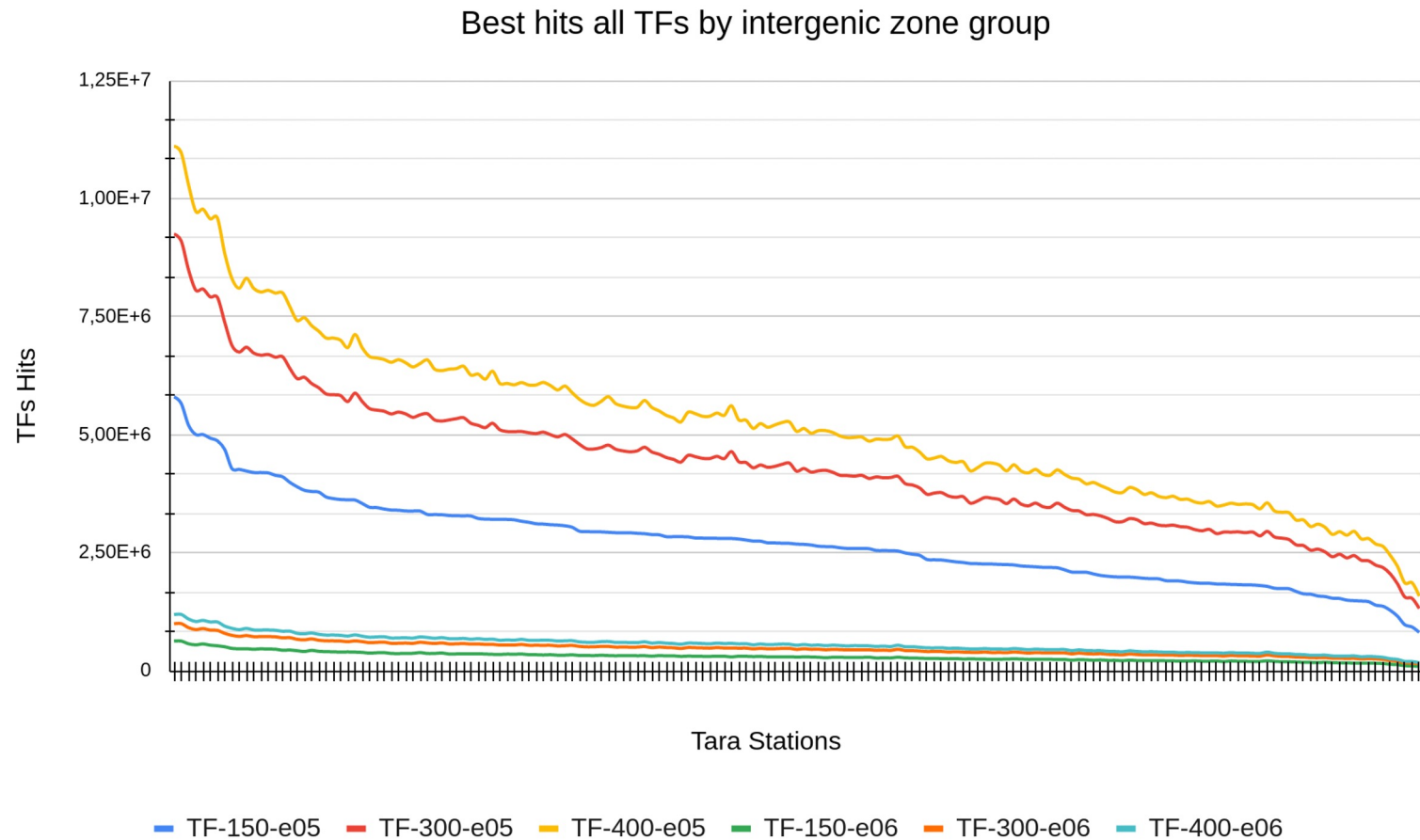
The statistics of motif search

TF hits against Intergenic group evalule

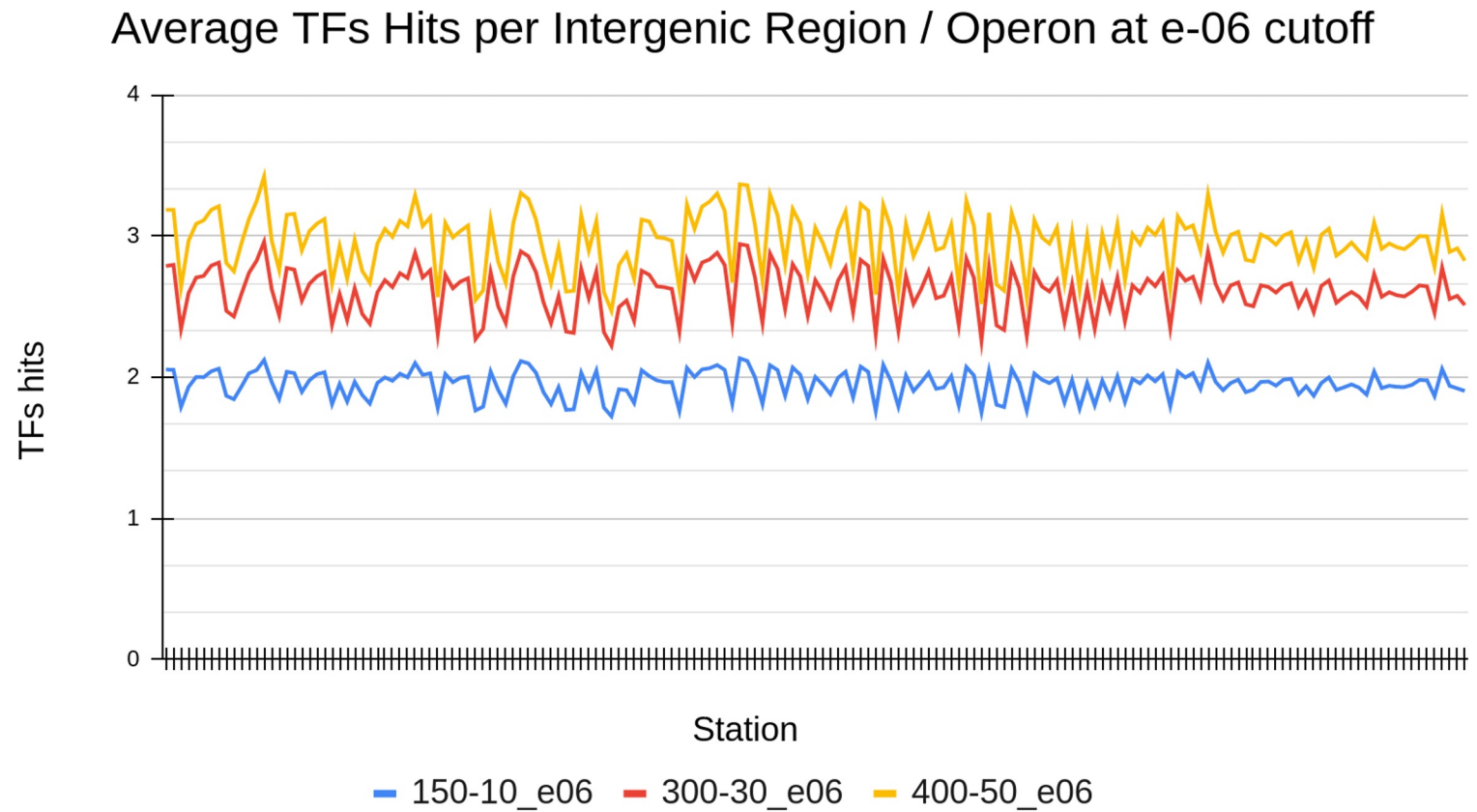
Intergenic	TF Hits
150-05-TF	470.727.510
150-06-TF	51.537.708
300-05-TF	755.193.693
300-06-TF	81.493.136
400-05-TF	907.313.613
400-06-TF	97.107.316



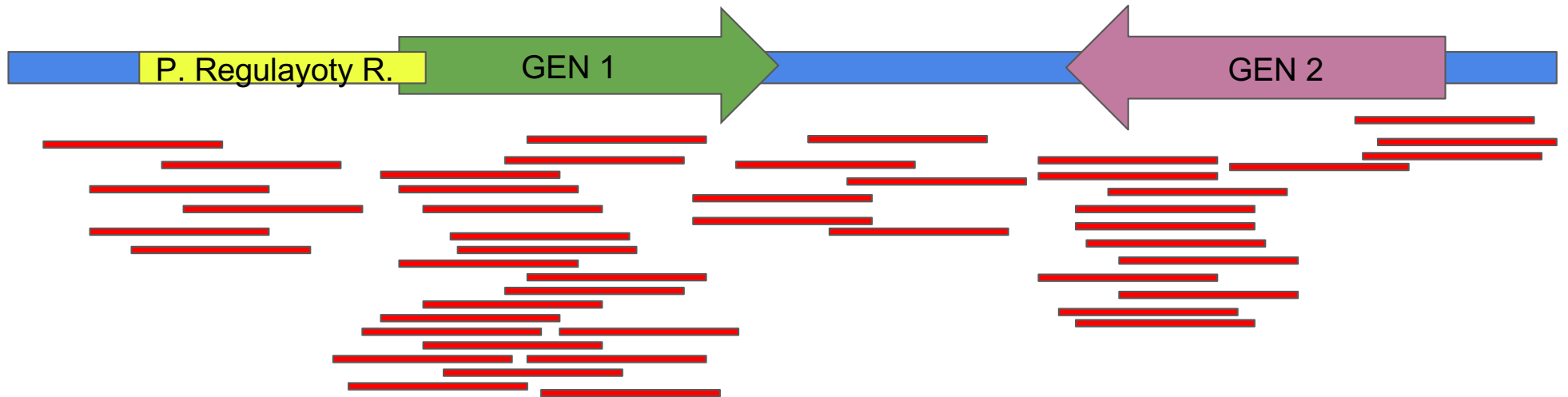
The statistics of motif search



The statistics of motif search



The coverage of motif search



Abundance of Genes and Regulatory regions are assigned like
Binned Coverage

The Matrix of motif search

Matrix 300-30-TF-TF_e06

TF	TSC000	TSC001	TSC002	TSC003	TSC004	TSC005	TSC006	TSC008	TSC011
ActR	5449.474	6516.430	3673.460	4770.489	7693.896	10096.909	6343.407	5272.018	2902.201
AggR	71267.592	66111.926	18814.866	38325.834	66869.347	79636.519	79356.733	80836.727	20867.764
AmrR	33683.897	34414.898	10693.026	20478.140	34459.949	37672.469	41373.965	43376.787	11049.661
AraR	51339.673	44395.368	9700.479	25886.892	48144.283	51081.496	57642.673	50257.319	14935.192
ArgR	439168.662	419368.591	90922.335	202748.686	411819.958	471520.890	511589.841	488367.284	119952.841
BioQ	17132.630	21615.052	8037.449	9106.279	15975.552	22550.898	20745.437	22258.134	6105.070
BirA	254092.418	241533.256	53740.059	116763.439	242294.992	290509.892	279646.663	285174.879	63785.169
CadR-PbrR	64297.315	65362.214	20736.893	33810.697	62315.288	76712.958	78132.458	67043.595	22006.705
CblR	38338.103	33307.331	10587.046	19691.547	34961.175	38124.249	32024.584	39846.846	12164.691
CcpA	59315.783	64836.621	21478.900	32765.052	59999.497	64069.396	77400.354	66562.268	16885.769
CsoR	16899.952	13260.475	6672.148	9650.225	15486.372	19544.081	14450.944	19199.807	5752.605
CueR	52931.597	51506.620	24655.926	40041.763	55369.865	78105.204	55171.311	63301.621	21131.211
CzrA	209056.593	189877.871	33113.477	80408.205	185985.168	210504.681	223635.309	212859.188	51479.308
DasR	11759.475	10894.698	5412.348	9598.126	15242.918	20202.684	15035.145	13643.223	6795.322
ExuR-UxuR	49728.469	51096.370	18877.168	29401.212	50223.210	63828.317	54584.109	54673.549	15868.136
FUR	400416.585	384236.836	82165.359	182964.683	375621.517	426976.573	463238.603	444737.676	105665.477
FabR	46445.996	37385.551	17870.458	23450.947	39081.300	50700.897	40908.049	41928.802	14520.379
FadP	4064.958	4335.889	2703.398	3435.432	4191.033	8129.286	3969.047	3732.393	2703.406
FadR	18096.960	18945.919	8860.431	13153.509	17488.187	27068.506	16681.050	19041.991	7593.444
FixJ	9749.443	7254.612	6365.168	5888.486	13140.367	10619.704	10402.920	10014.939	4261.228
FixN-FixK	25099.876	29699.067	11296.382	17509.027	32699.125	44064.869	33654.926	28076.084	12264.173
FruR	19529.905	22656.086	11957.721	15511.116	21762.611	28778.448	21398.306	19897.102	8226.579
GguR	20830.392	18336.143	8596.298	9593.684	16341.648	13290.372	12257.124	12388.142	3944.956
GlcC	26539.991	31853.173	10595.720	17421.212	28456.418	41182.683	32259.753	30956.606	10654.320
GntR	43067.551	41477.922	21526.824	25945.515	48459.302	62020.151	42611.274	44358.427	17815.194
GulR	38736.238	30141.981	7525.154	16571.920	33309.233	40148.723	42724.003	41601.718	9387.999
HcpR	39516.904	46602.127	14098.221	28978.455	42399.836	55444.695	48882.443	40799.106	15172.986
HexR	83441.803	77046.147	26528.960	52117.113	87368.814	91245.525	98890.148	93128.186	25502.978
HisR	60525.376	53050.600	19785.643	34063.675	55026.189	69291.575	70907.573	66036.963	16234.991
HrcA	89877.127	90120.681	33630.559	49524.183	88307.499	103822.846	97013.366	94241.556	30861.269

Specific Matrix of motif search

	TFs	Genes	
Salazar Matrix	89	57	Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome Guillem Salazar ³² • Lucas Paoli ³² • Adriana Alberti • ... Matthew B. Sullivan • Patrick Wincker • Shinichi Sunagawa ³³  • Show all authors • Show footnotes
Stress Matrix	89	63	Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation LUCAS J. USTICK , ALYSE A. LARKIN , CATHERINE A. GARCIA , NATHAN S. GARCIA , MELISSA L. BROCK , JENNA A. LEE , NICOLA A. WISEMAN , J. KEITH MOORE , AND ADAM C. MARTINY  Authors Info & Affiliations
Guidi Matrix	89	856	Published: 10 February 2016 Plankton networks driving carbon export in the oligotrophic ocean Lionel Guidi , Samuel Chaffron , Lucie Bittner , Damien Eveillard , Abdelhalim Larhlimi,

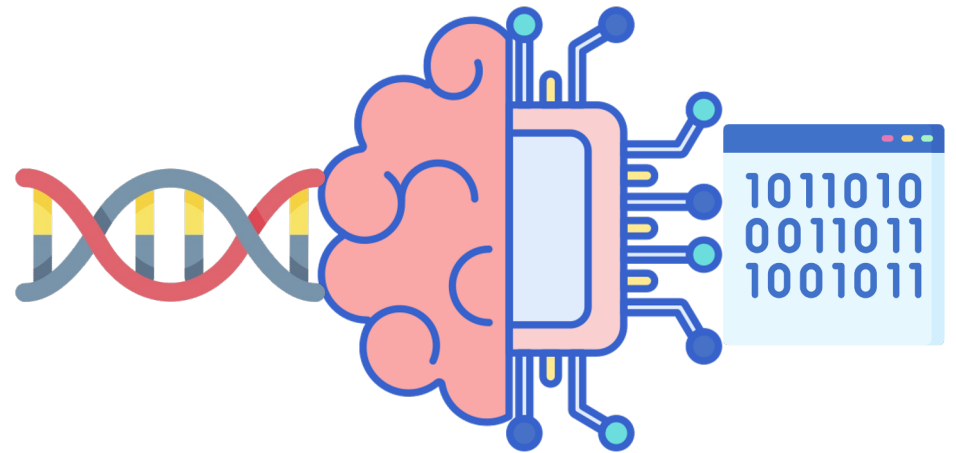
The Matrix Catalog

Matrix Type	Description
MX	TFs are counted if they appear in an intergenic region
M0	TFs are counted if they regulate any of 32k GENEs
M1	TFs are counted if they regulate any of 9k metabolic defined GENEs
M2	TFs are counted if they regulate any of 3k metabolic genes with ECs
M3	TFs are counted if they regulate any of 7k in metabolic genes with KOs
M4	GENEs are counted by raw coverage

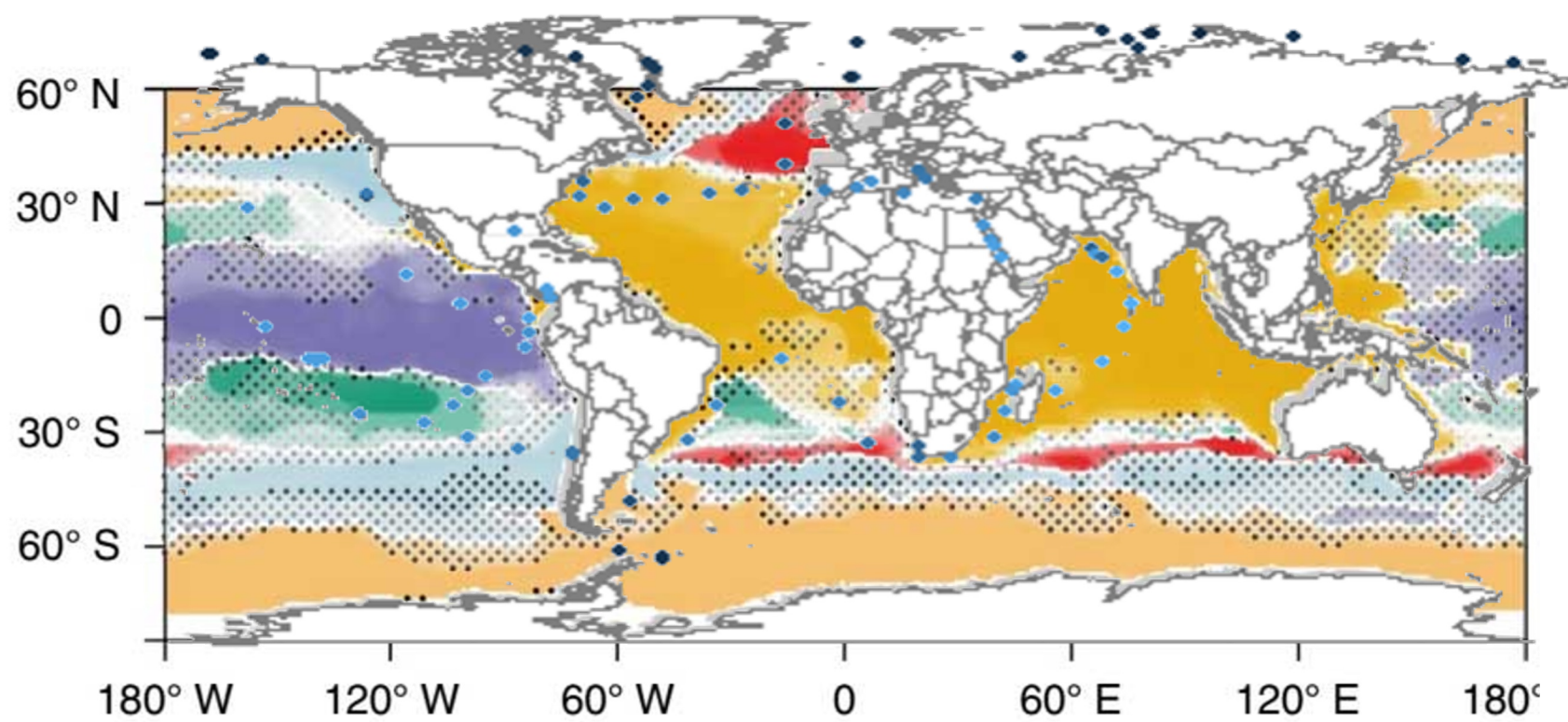
Currently 62 Matrices
with custom

- Evaluate cutoff
- Gene or feature list
- TF list
- Cluster group

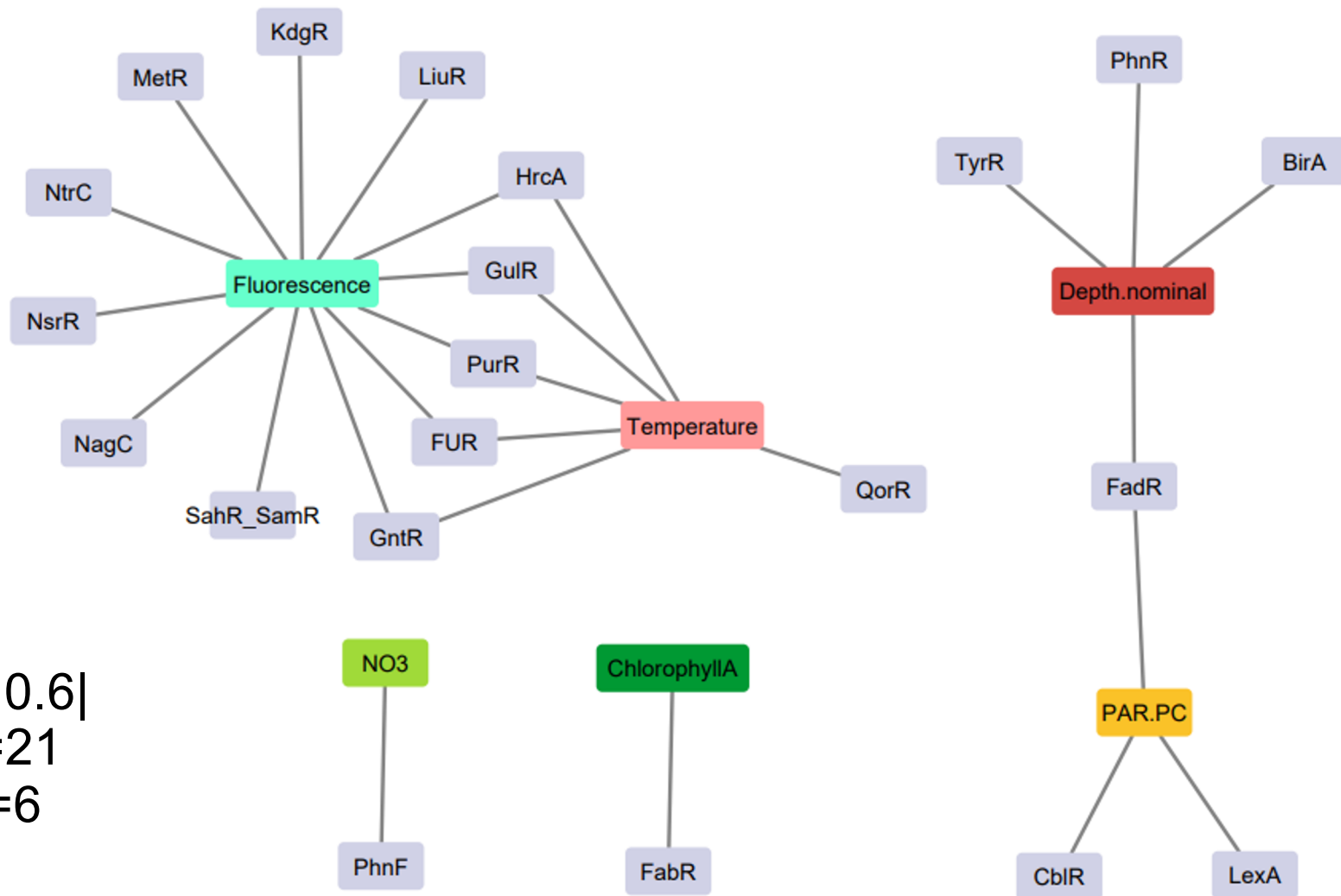
1st exercise: bioprovinces vs regulated genes



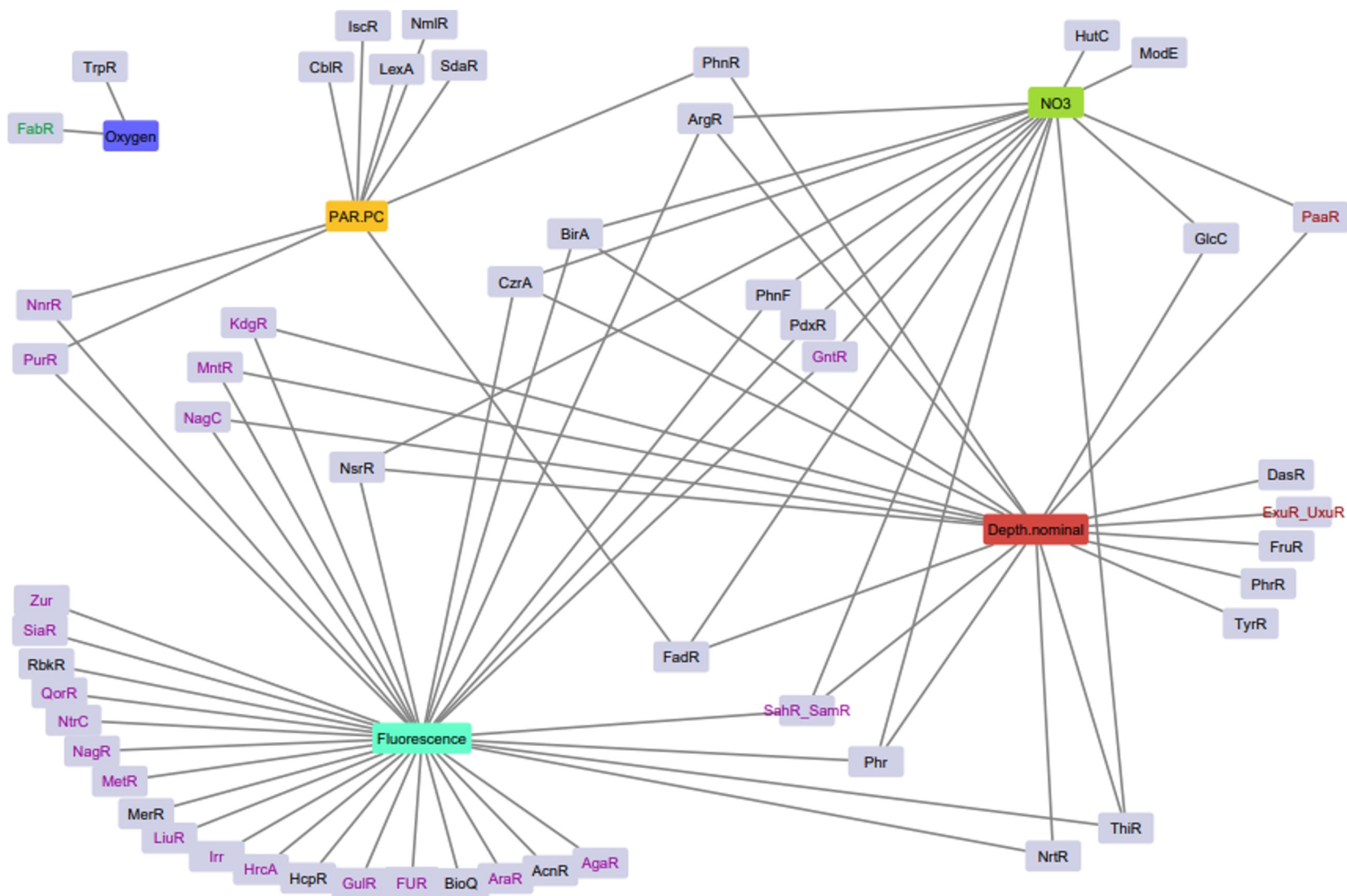
Province

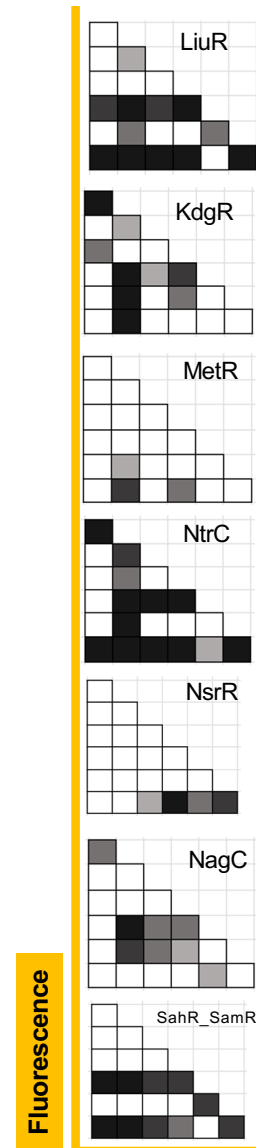
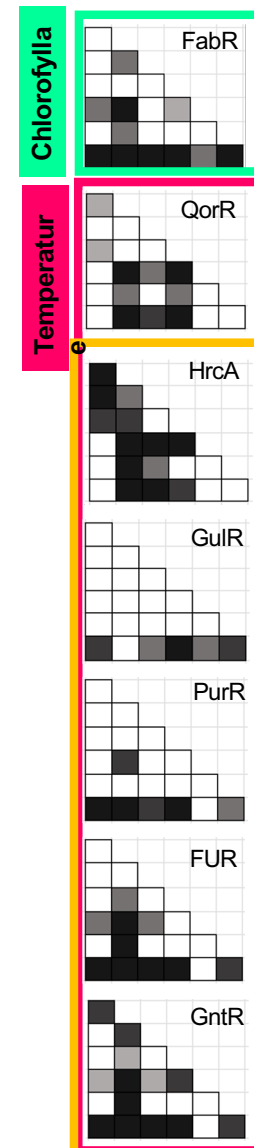
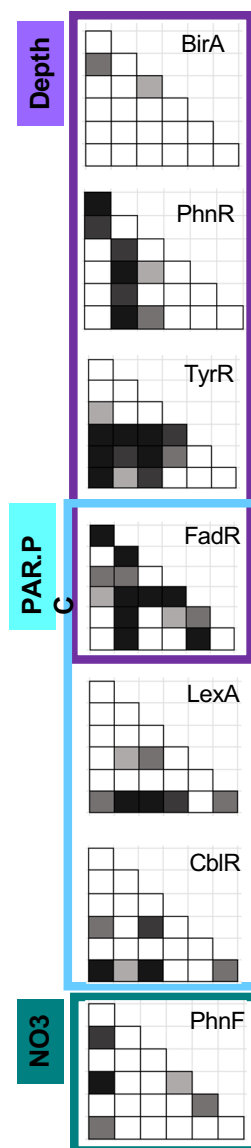
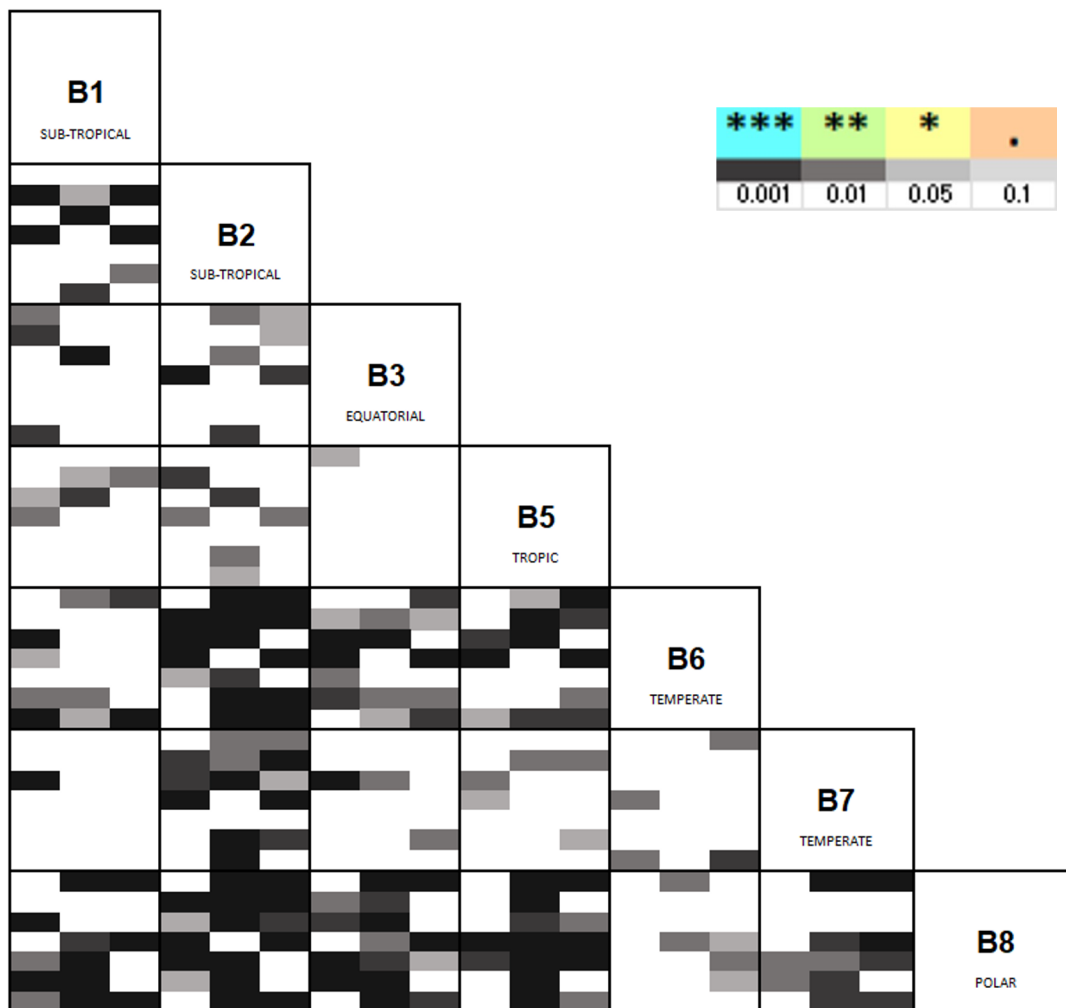


corr. > |0.6|
TFs: n=21
vars: n=6

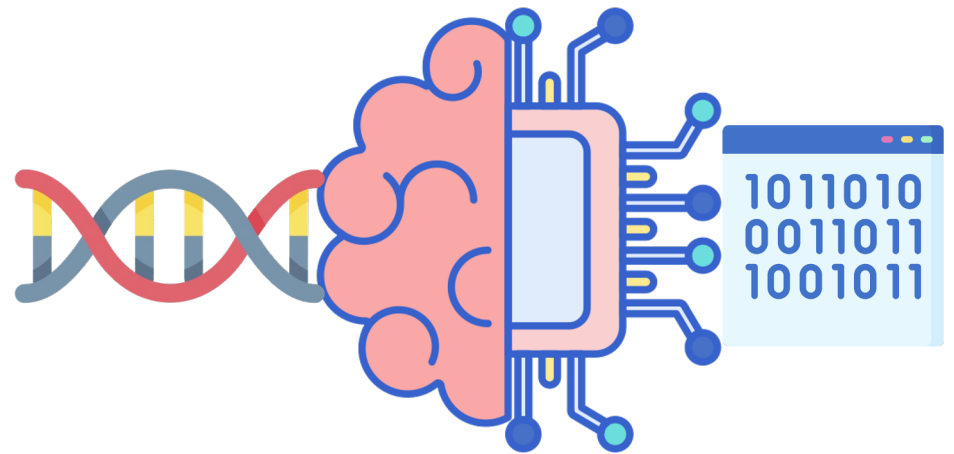


corr. > |0.5|
TFs: n=52
vars: n=8





2nd exercise: predicting the environment

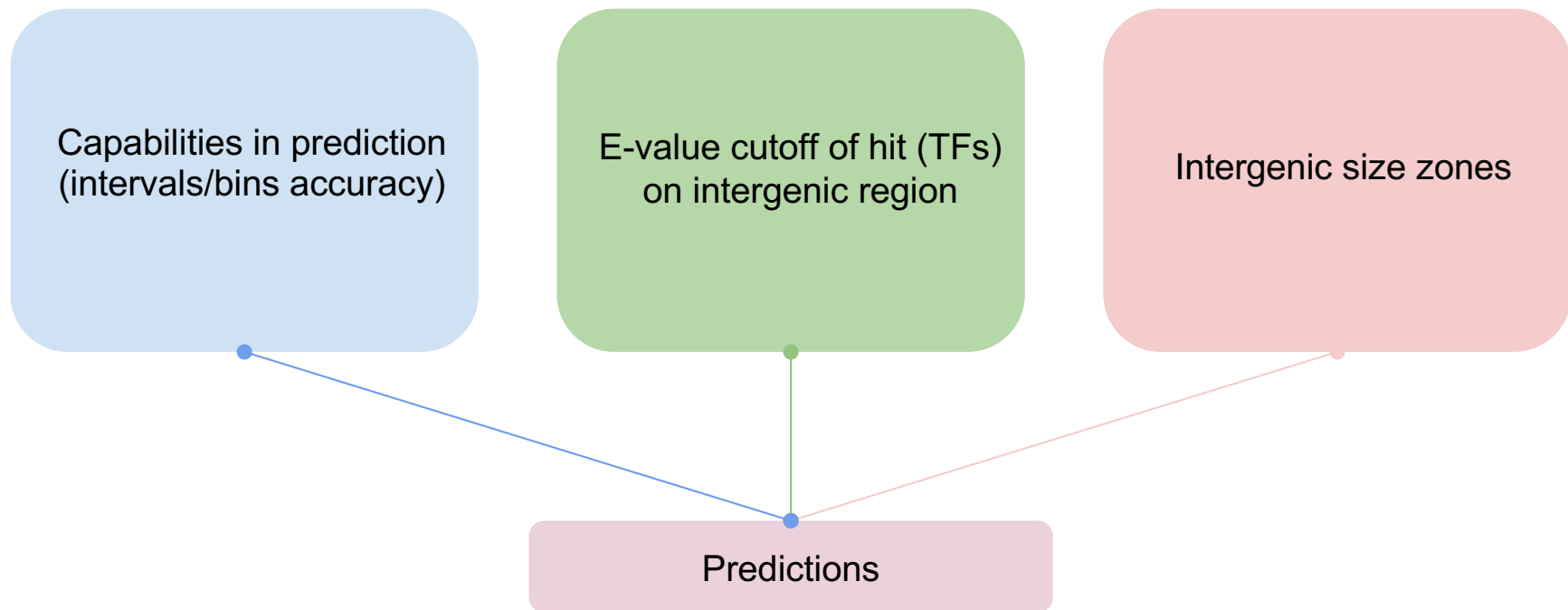


Random Forest v0.22

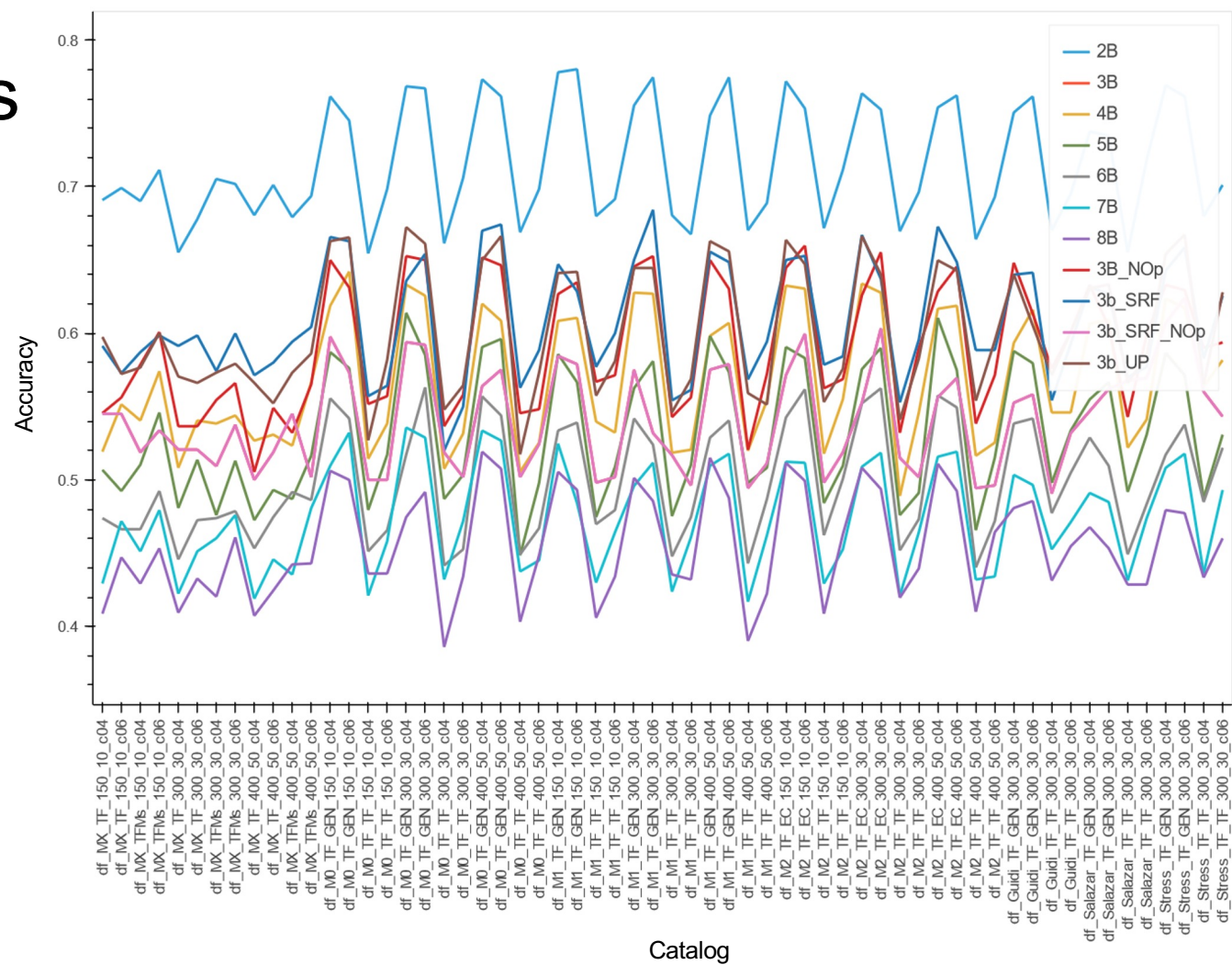
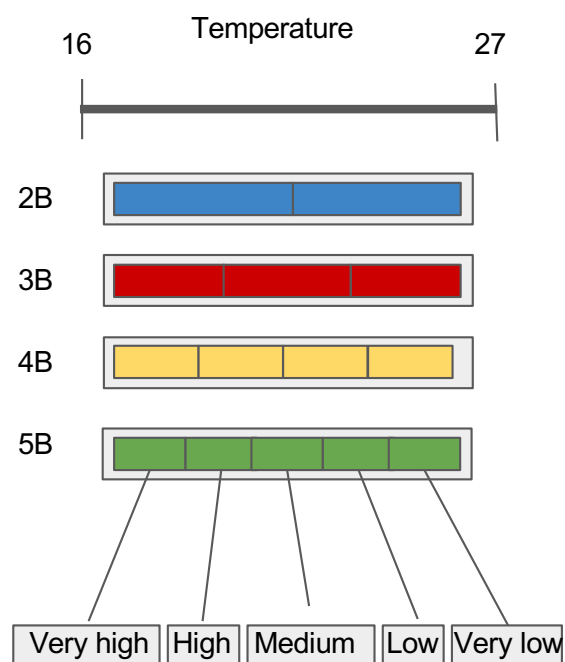
- Metrics Calibration :
- Test Size = 0.3
- Estimators = 100
- Criterion = Entropy
- Warm start = True
- Average = weighted
- Max features = sqrt
- Balanced = $n_samples / (n_classes * np.bincount(y))$



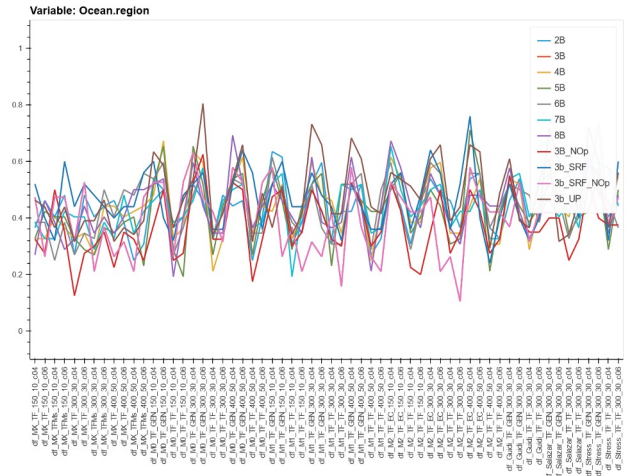
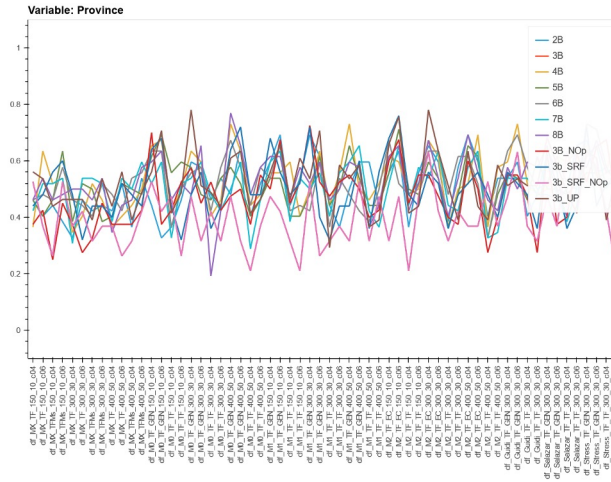
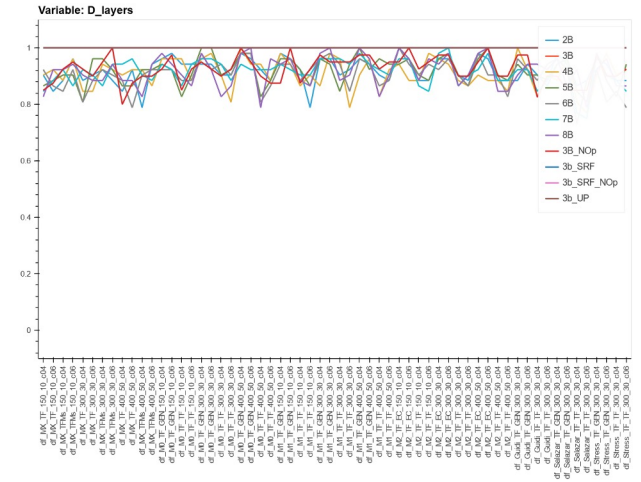
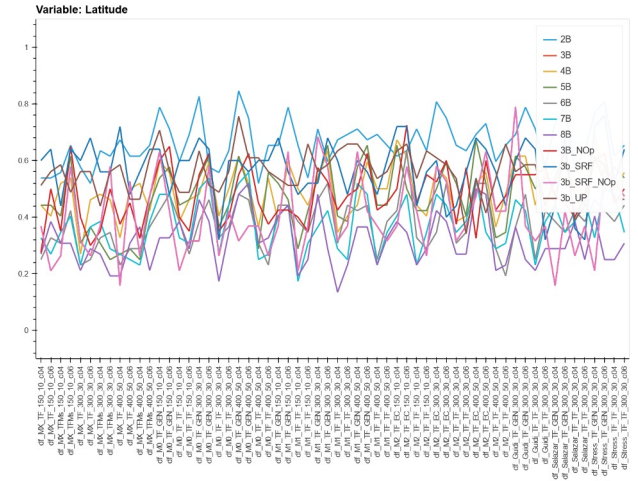
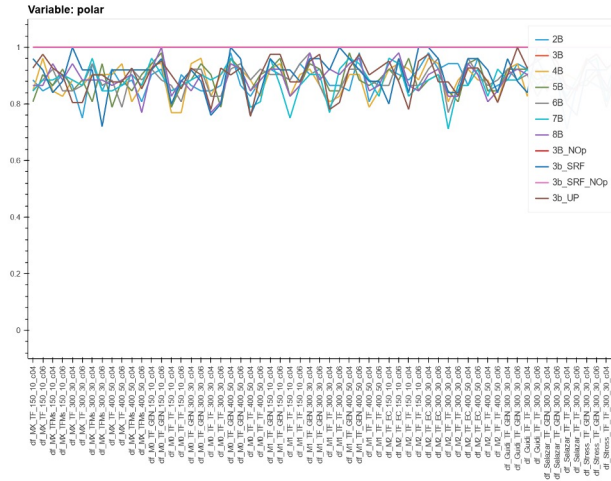
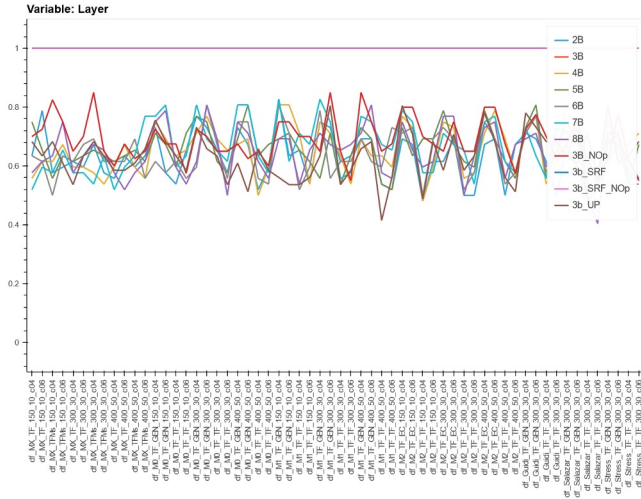
Catalog(Data) Calibration



Intervals / Bins

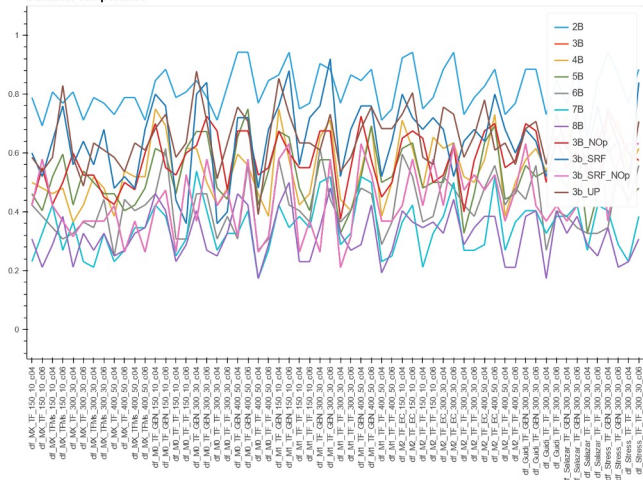


Geo Data

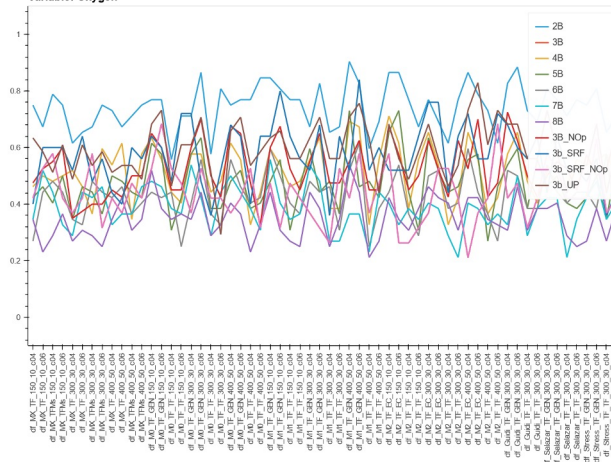


Physicochemical data

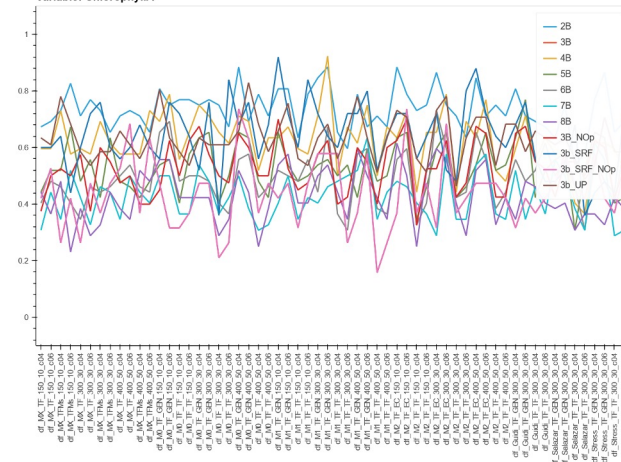
Variable: Temperature



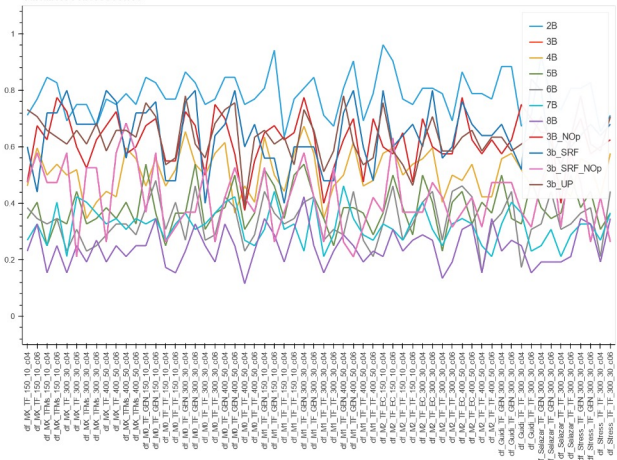
Variable: Oxygen



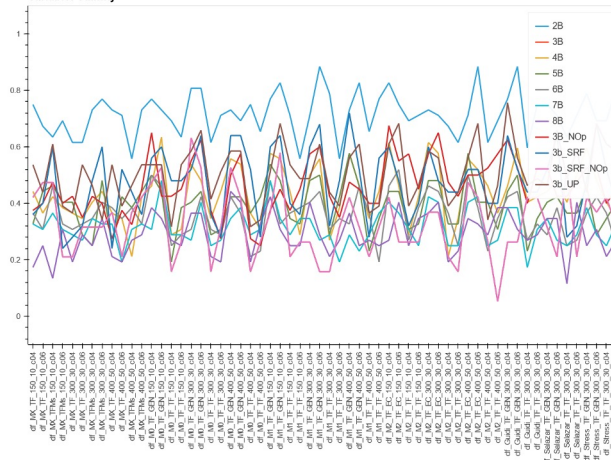
Variable: Chlorophylla



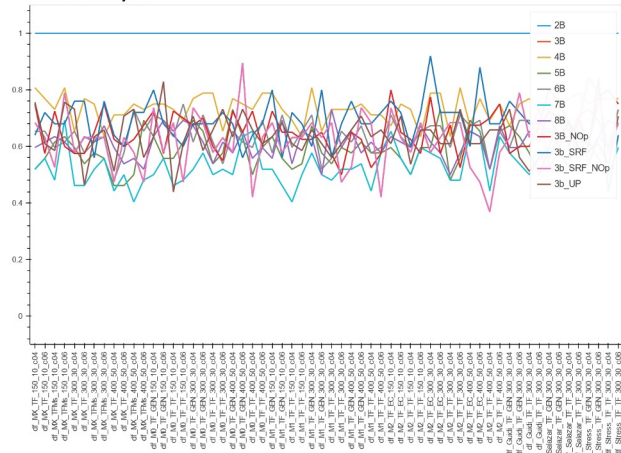
Variable: Fluorescence



Variable: Salinity

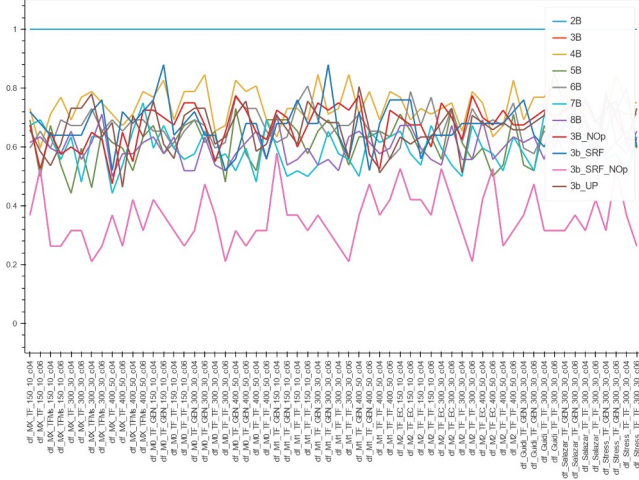


Variable: Alkalinity.total

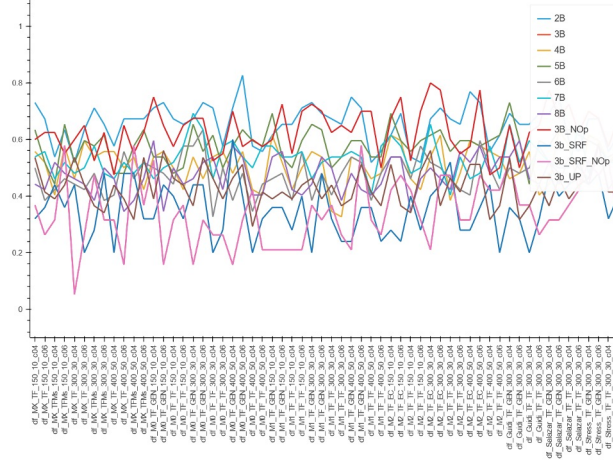


Physicochemical data

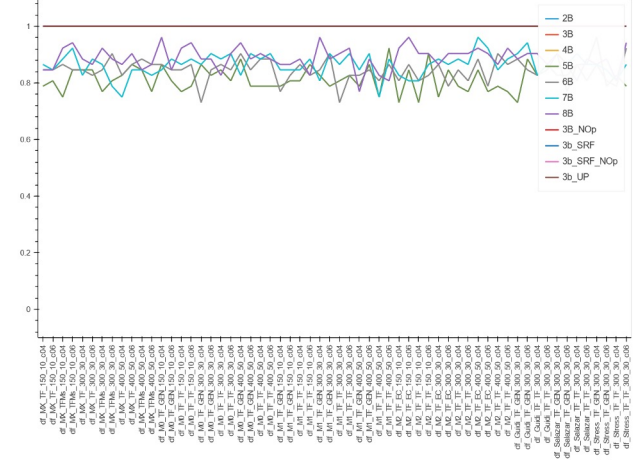
Variable: Mean Flux at 150m



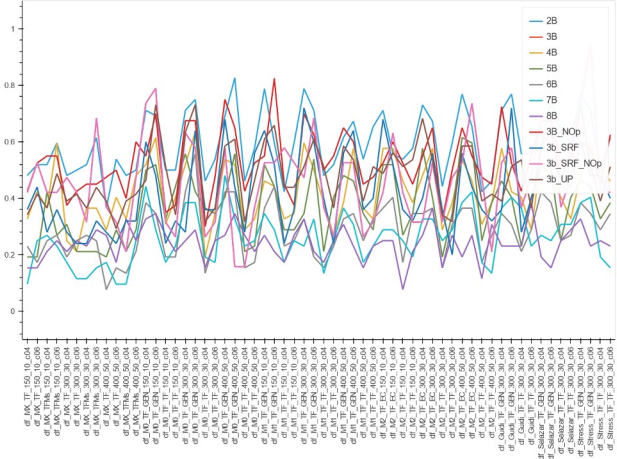
Variable: CO3



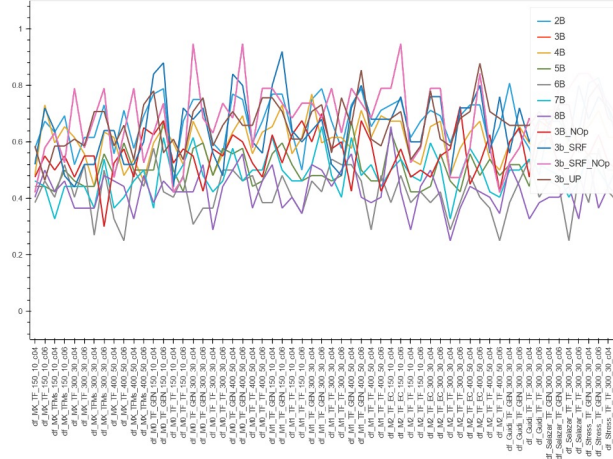
Variable: HCO3



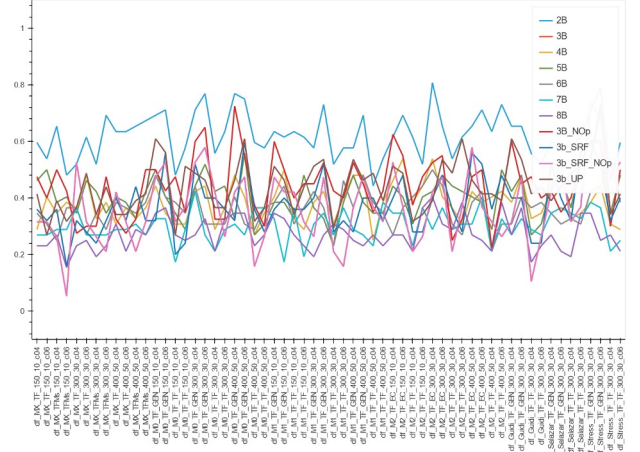
Variable: Iron.5m



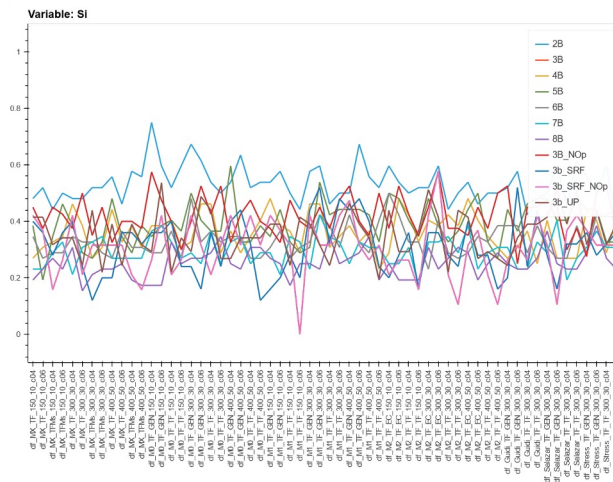
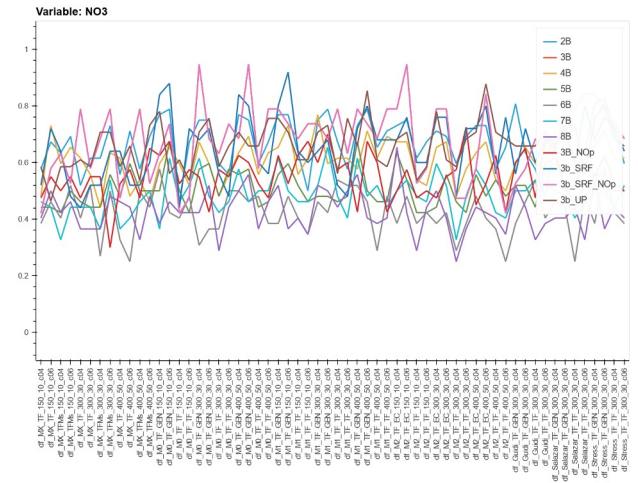
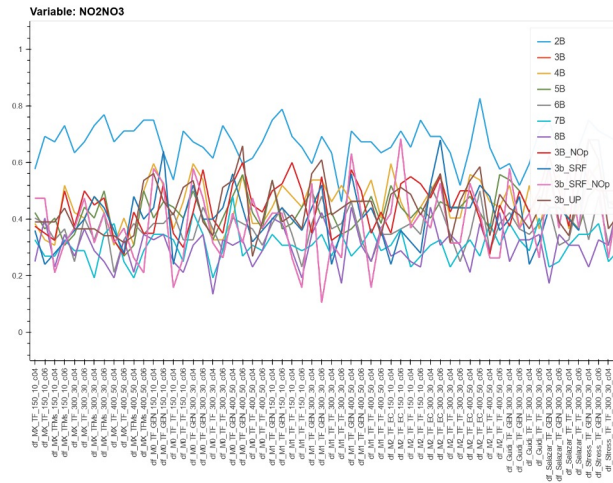
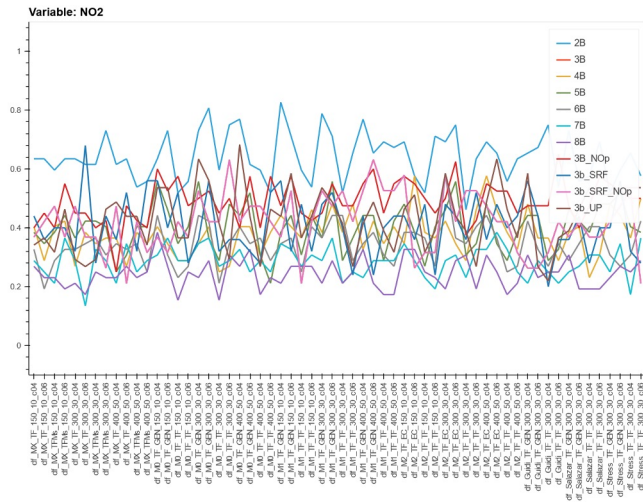
Variable: NO3



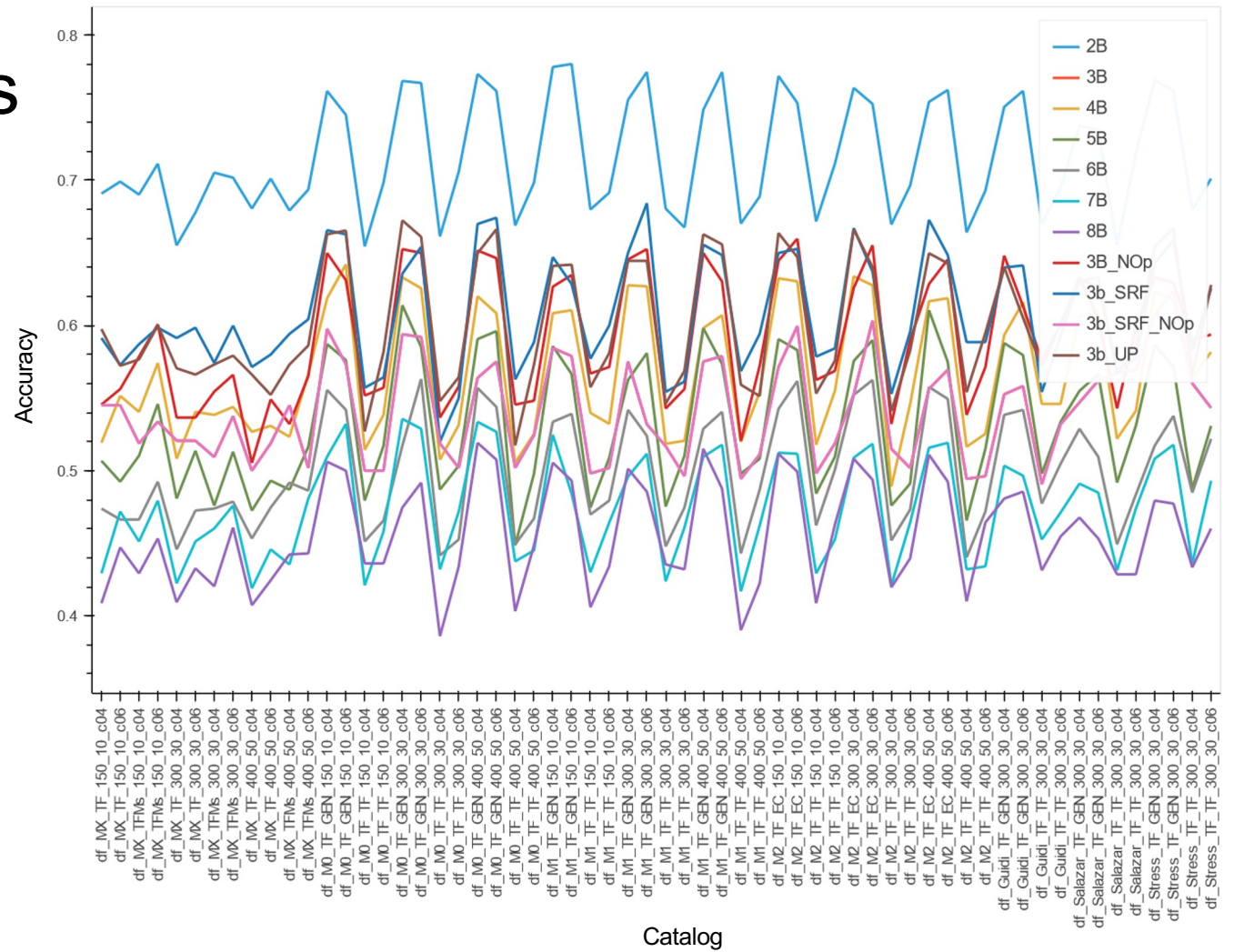
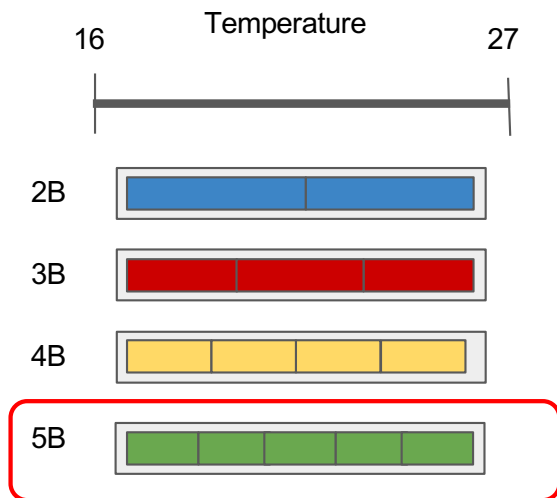
Variable: PO4



Physicochemical data



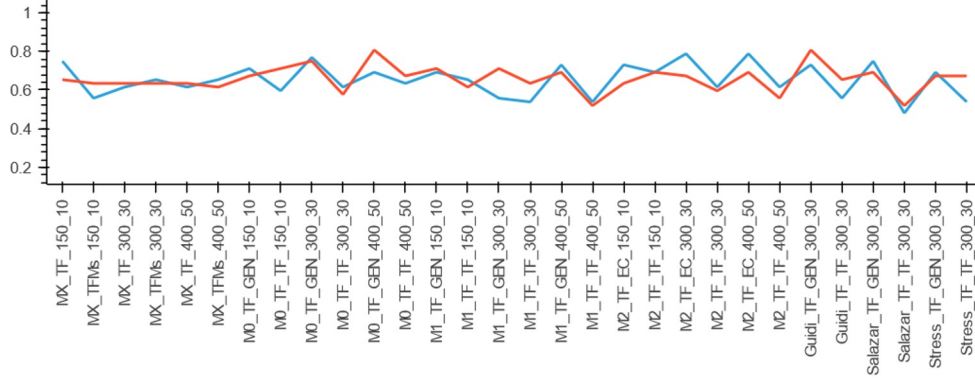
Intervals / Bins



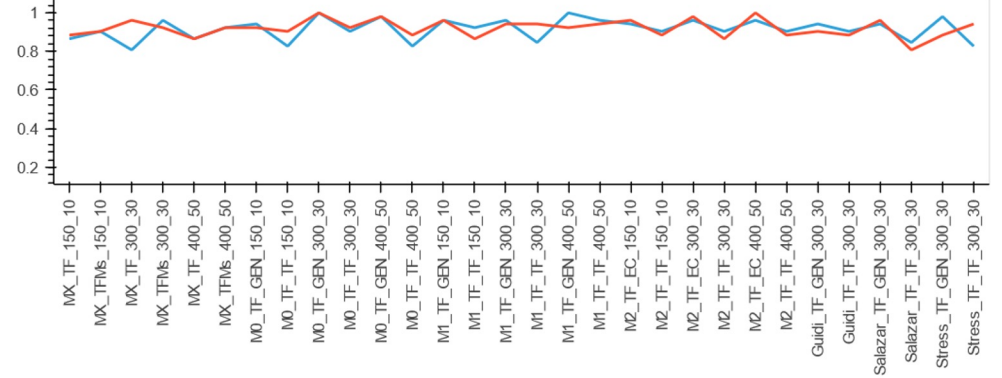
E-value cutoff of hit (TFs)
on
intergenic region

e-05 vs e-06

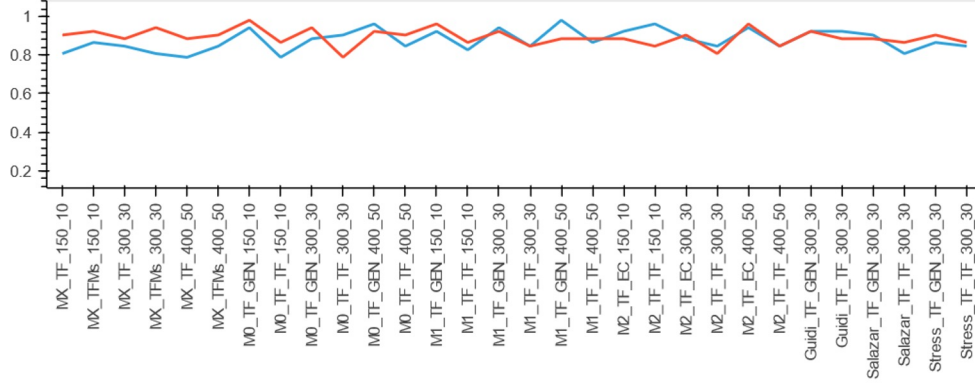
Variable: Layer



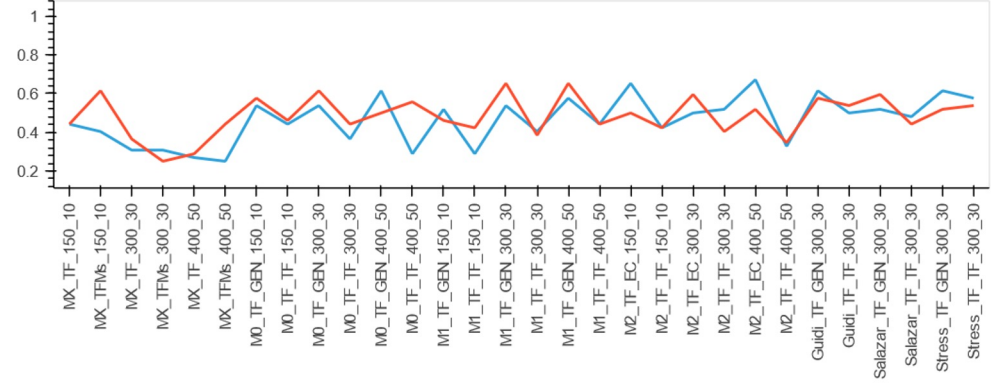
Variable: D_layers



Variable: polar



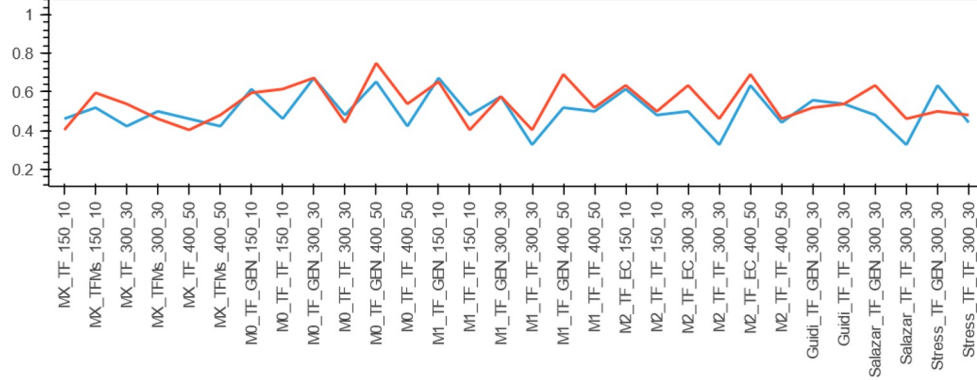
Variable: Latitude



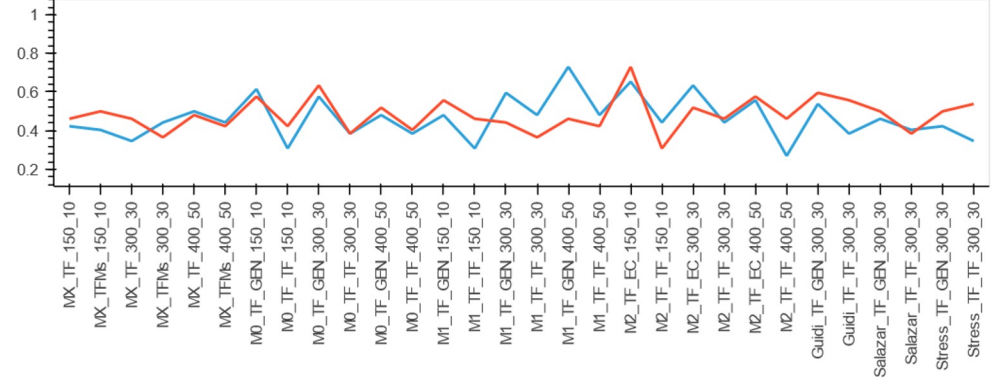
● e05 ● e06

e-05 vs e-06

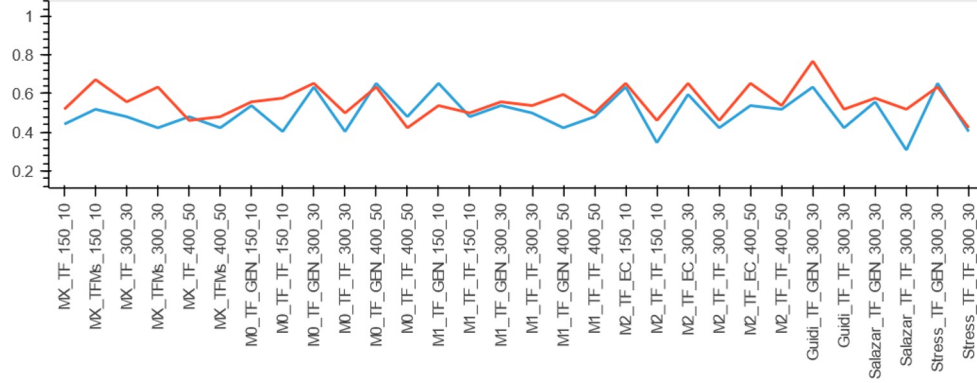
Variable: Temperature



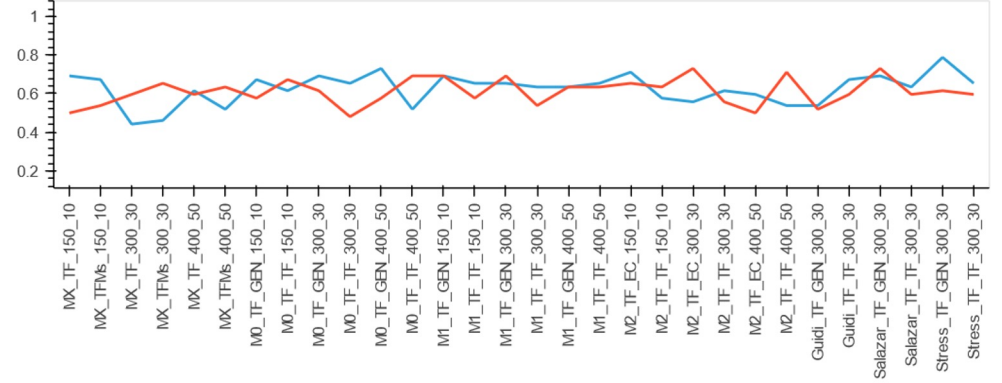
Variable: Oxygen



Variable: ChlorophyllaA



Variable: Mean Flux at 150m

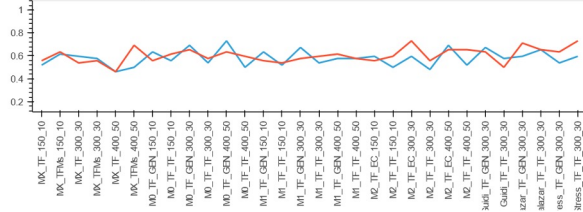


e05 e06

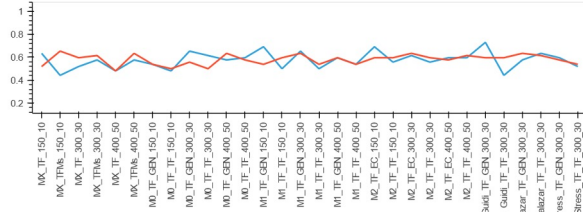
General view

e05 e06

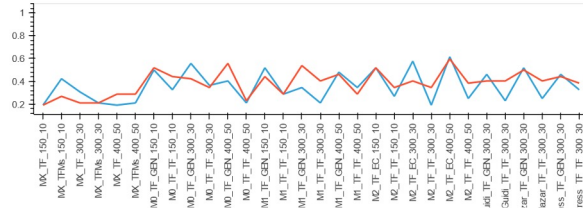
Variable: Alkalinity.total



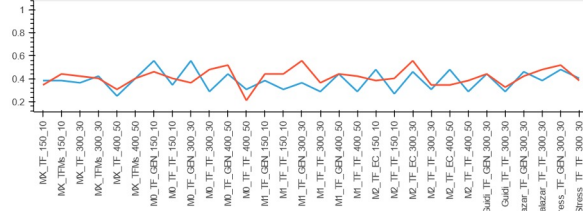
Variable: CO3



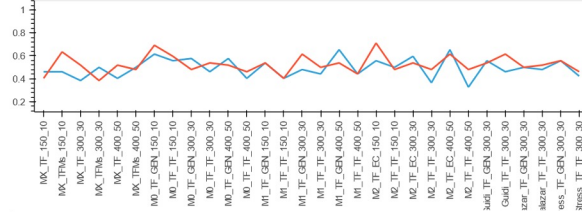
Variable: Iron.5m



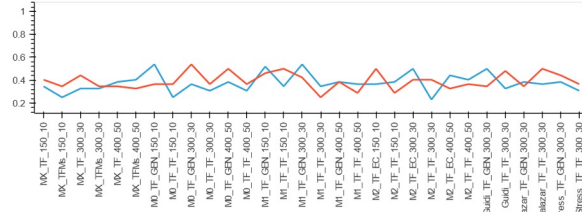
Variable: NO2



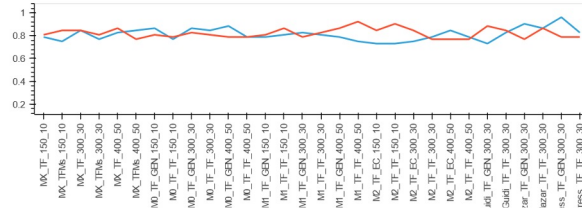
Variable: Province



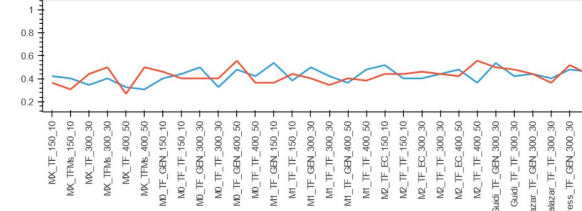
Variable: Fluorescence



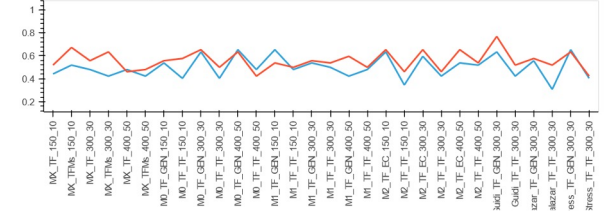
Variable: HCO3



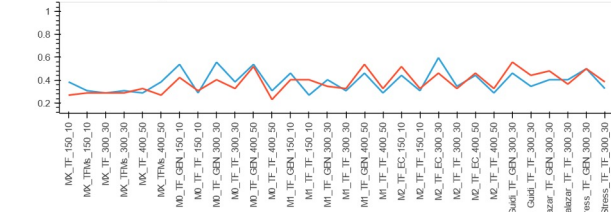
Variable: NO2NO3



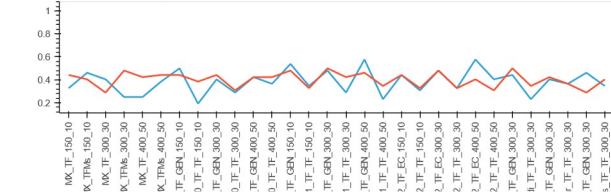
Variable: ChlorophyllA



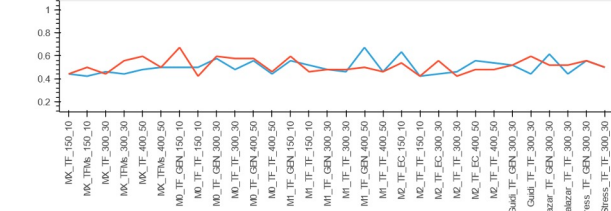
Variable: Density



Variable: Salinity



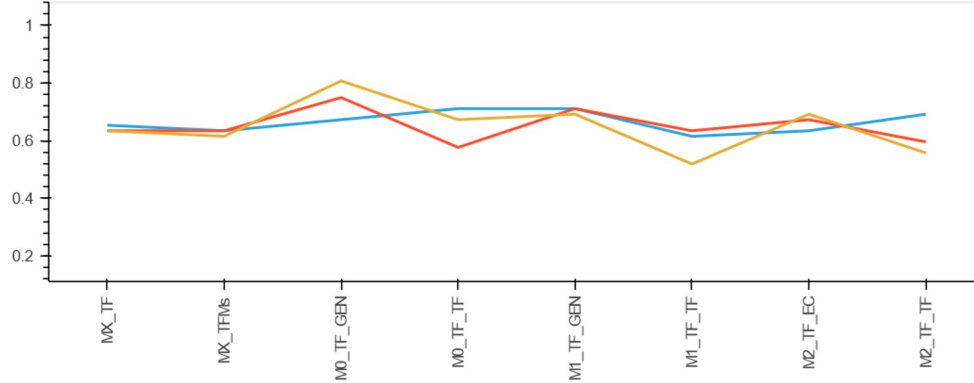
Variable: NO3



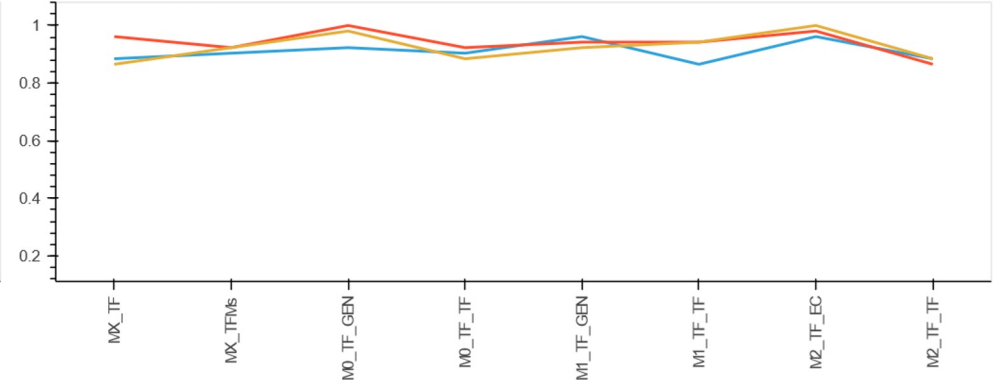
Intergenic size zones

Intergenic Size

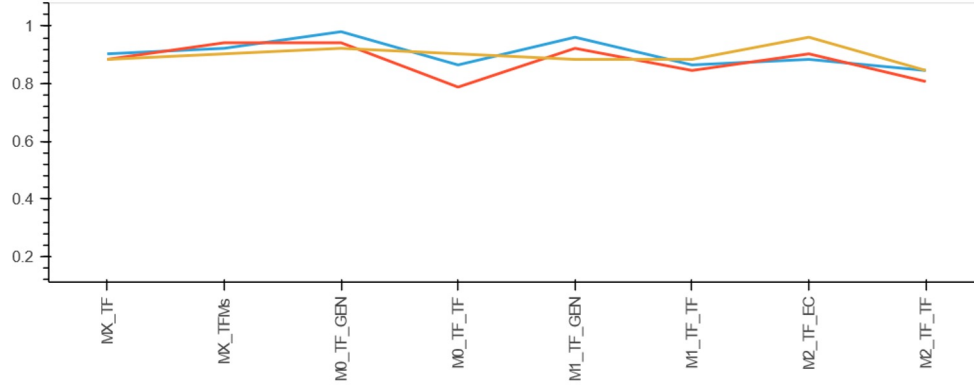
Variable: Layer



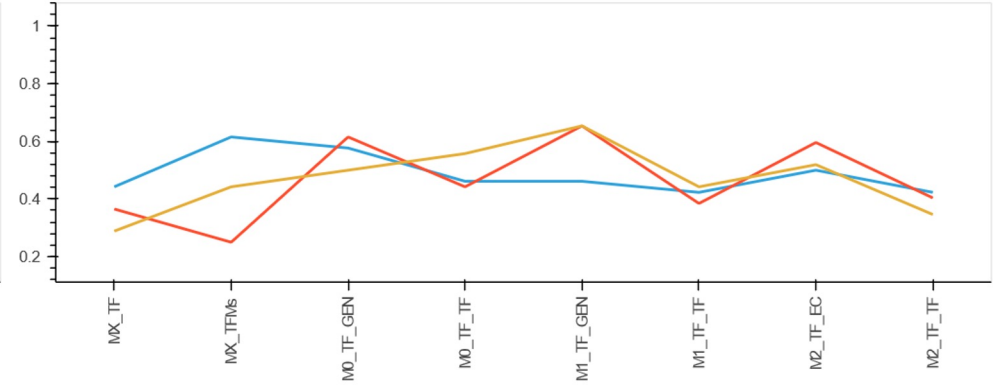
Variable: D_layers



Variable: polar



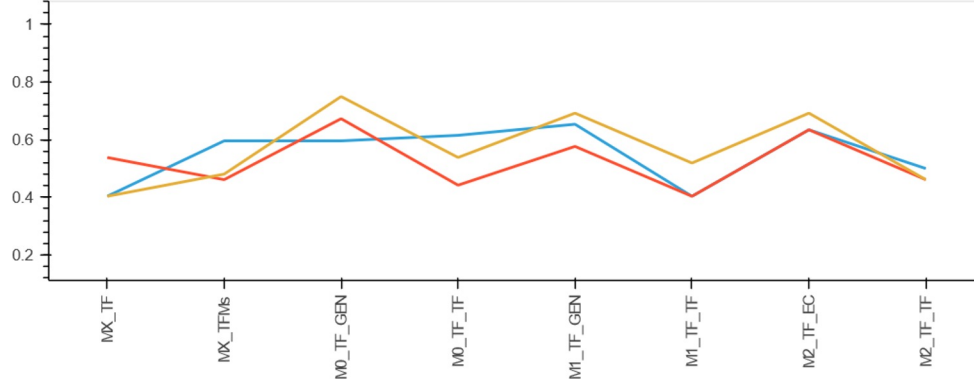
Variable: Latitude



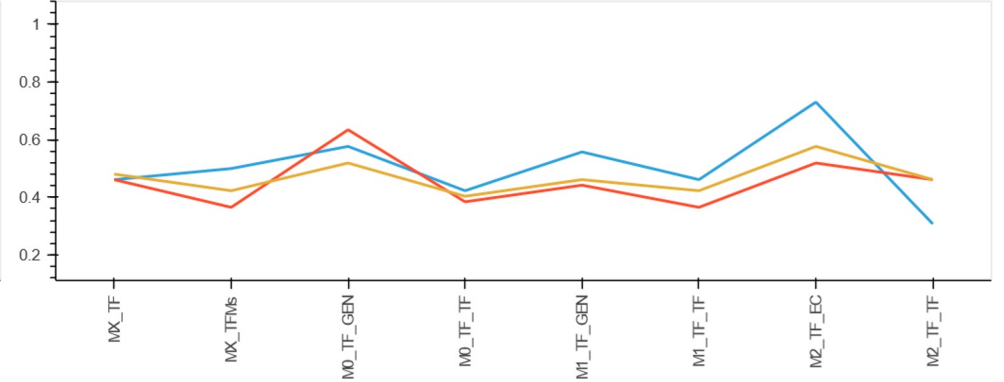
150-10 300-30 400-50

Intergenic Size

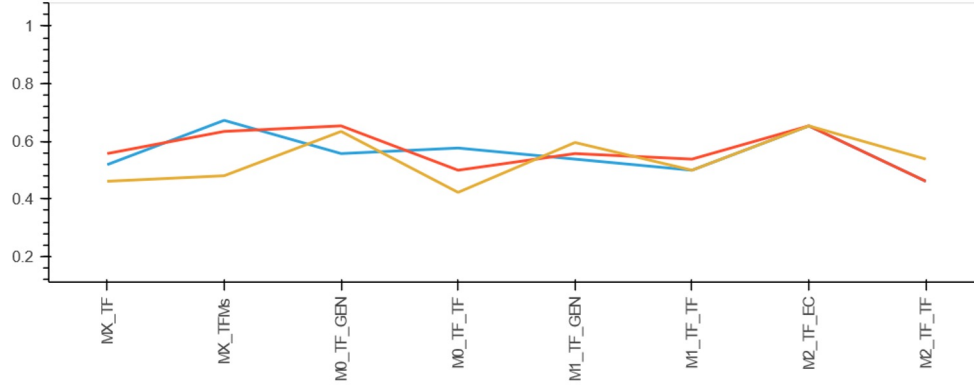
Variable: Temperature



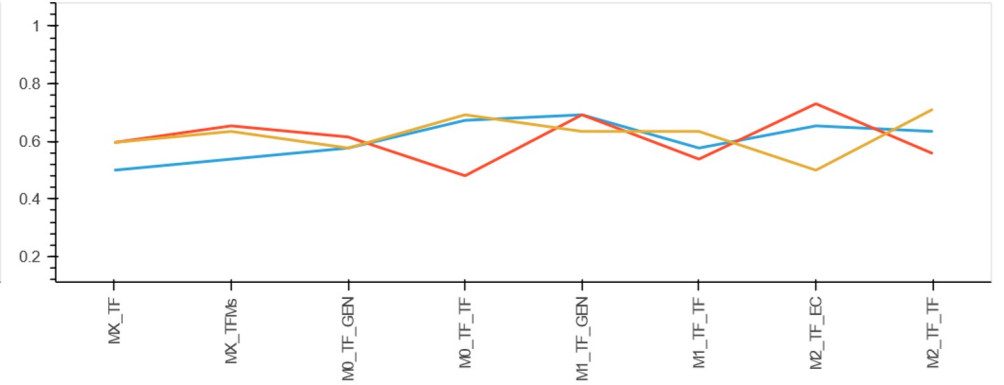
Variable: Oxygen



Variable: Chlorophylla



Variable: Mean Flux at 150m



150-10 300-30 400-50

General view



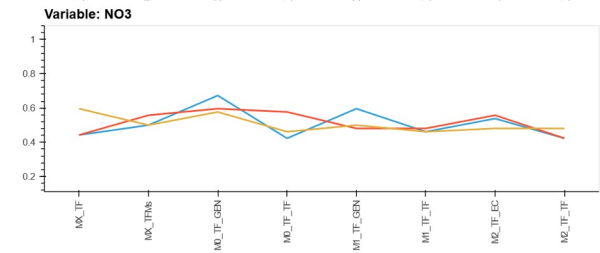
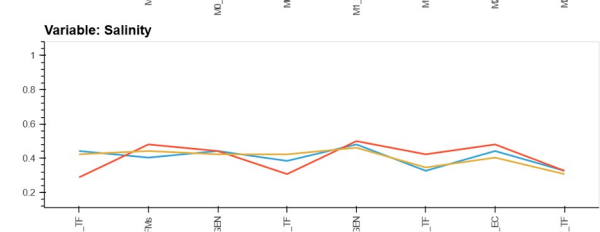
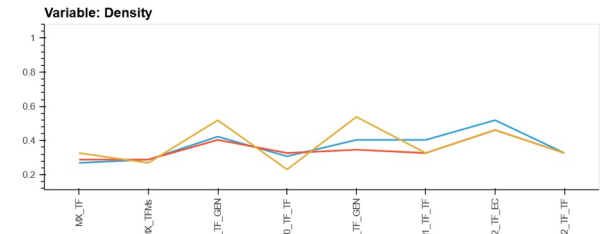
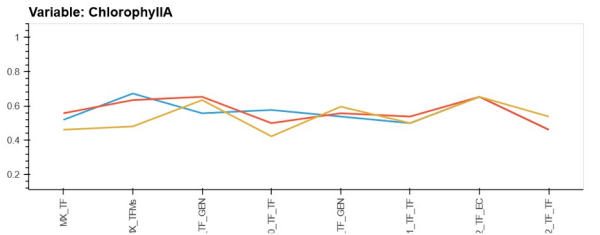
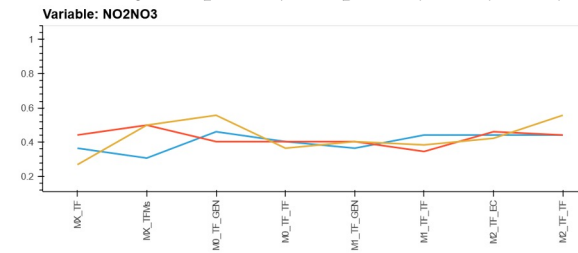
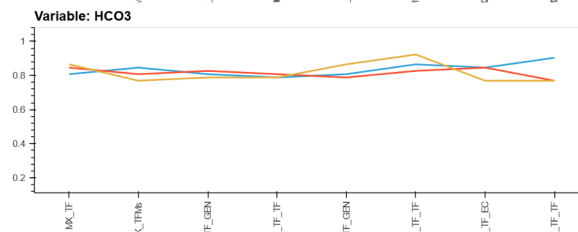
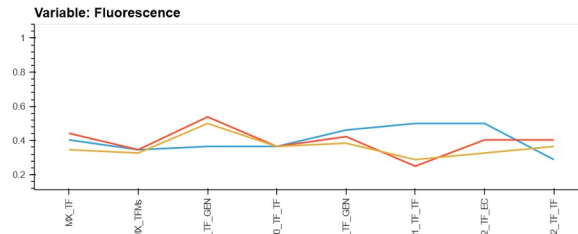
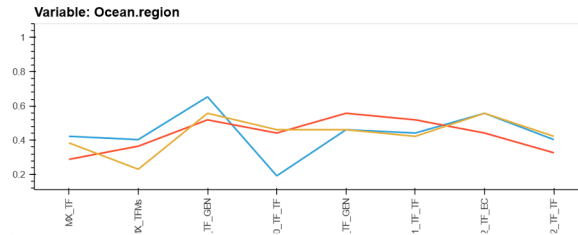
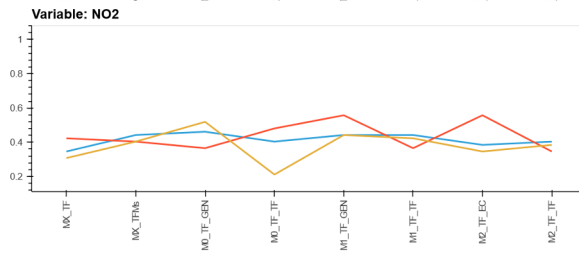
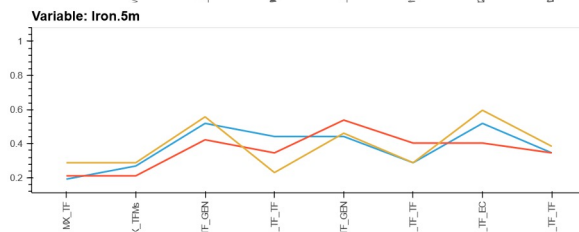
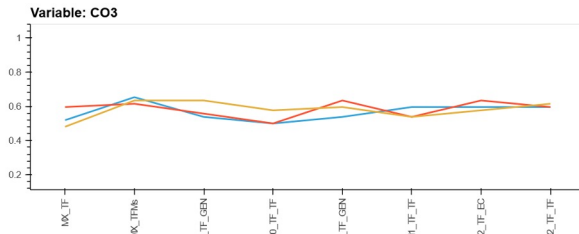
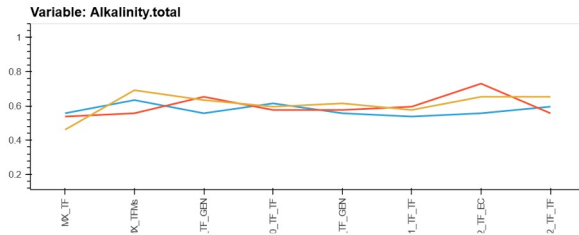
150-10



300-30



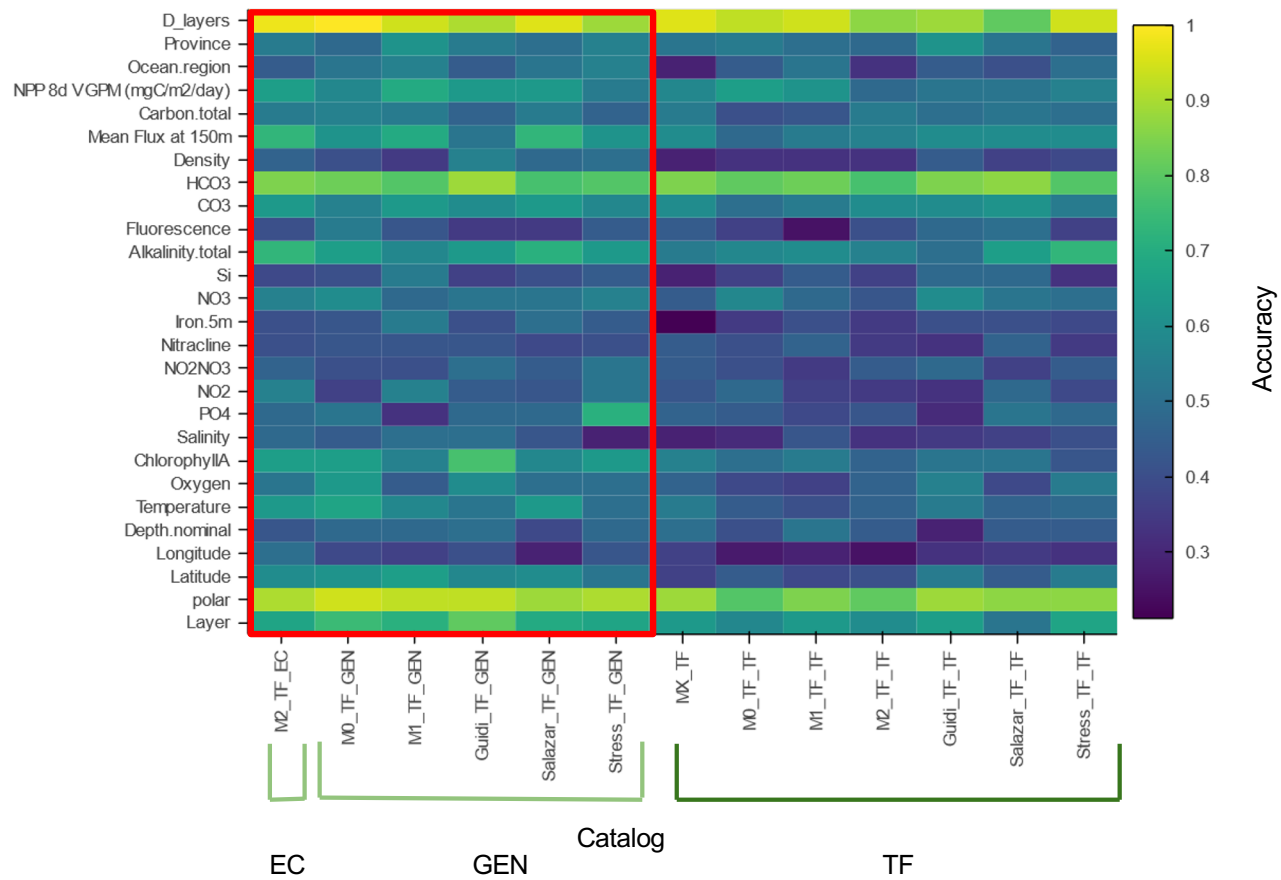
400-50

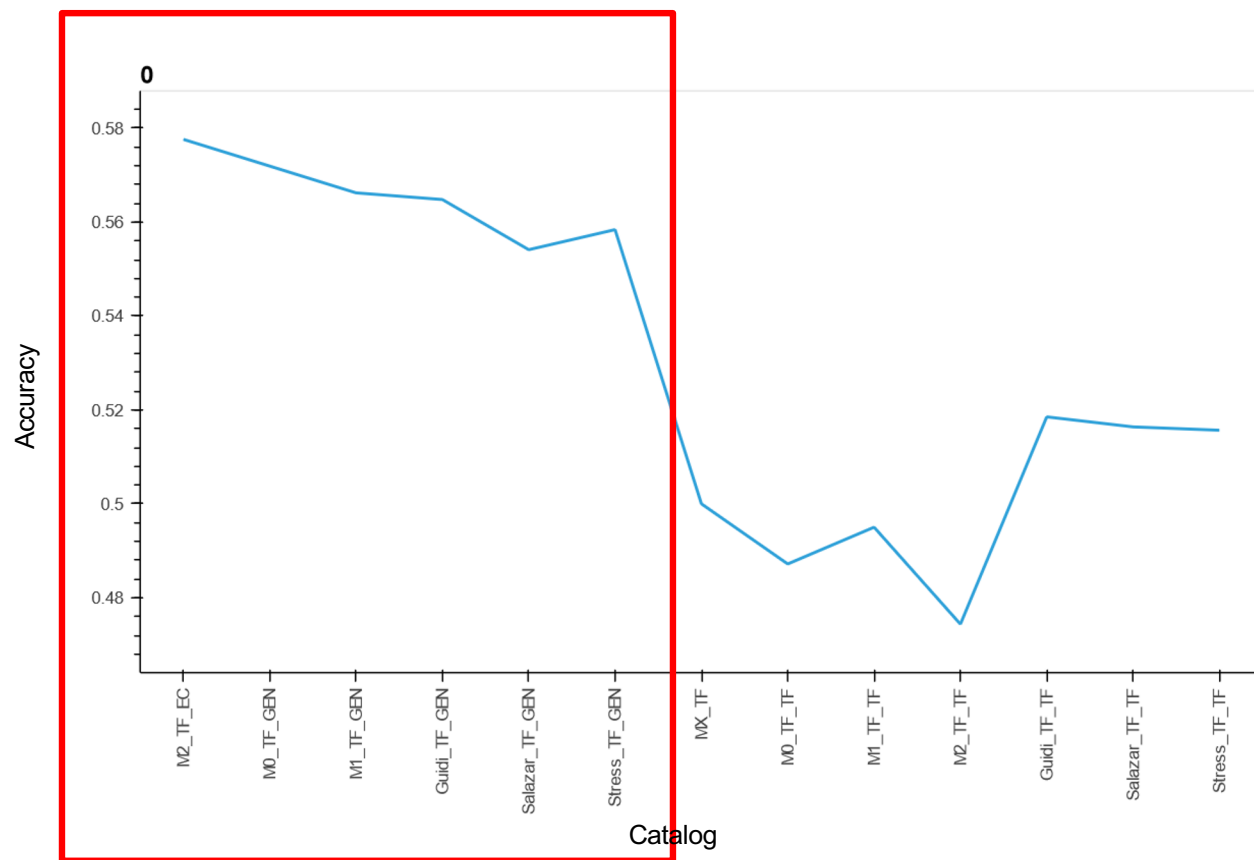


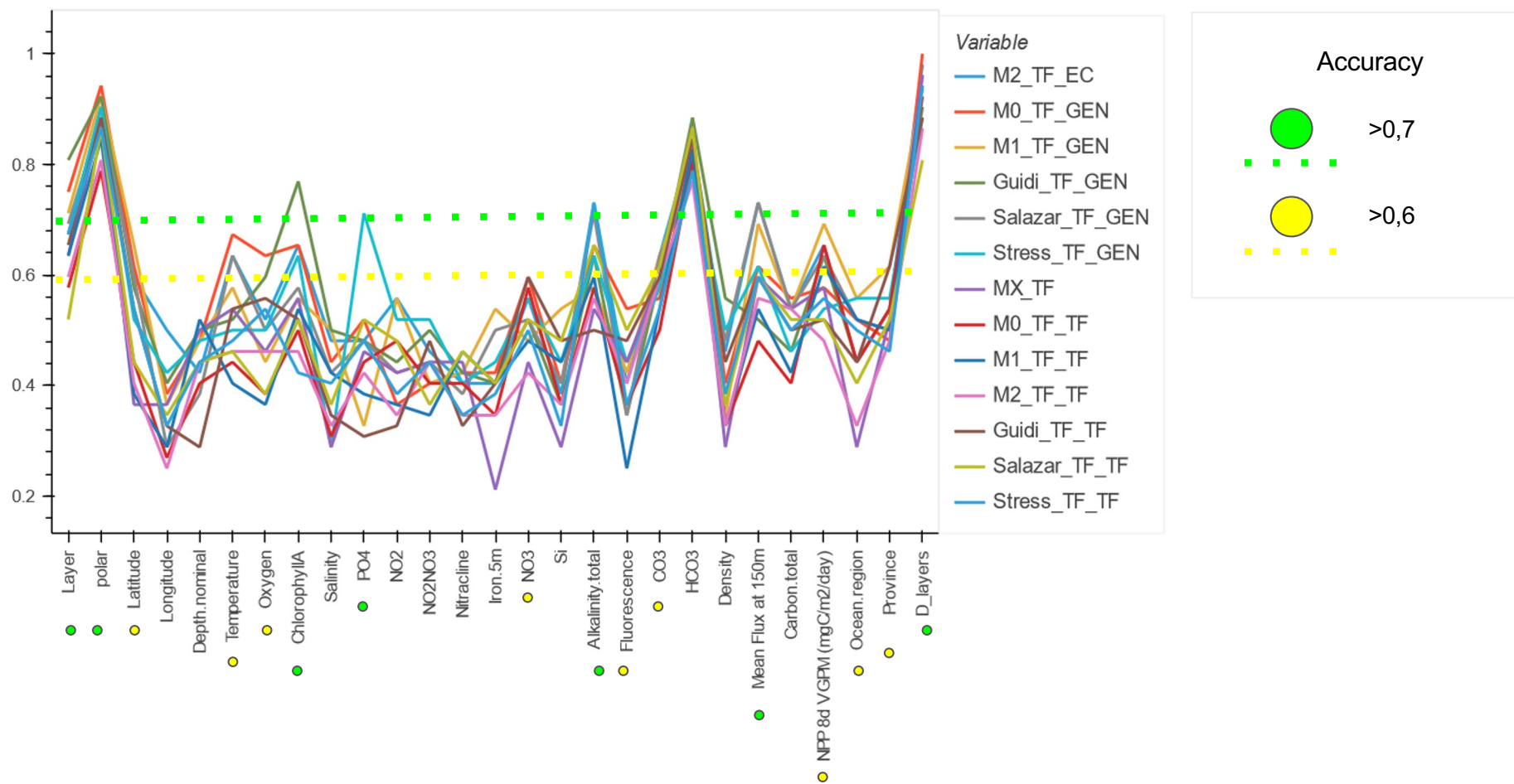
Prediction

300-30-e06

“the 89 TFs are a good resource”

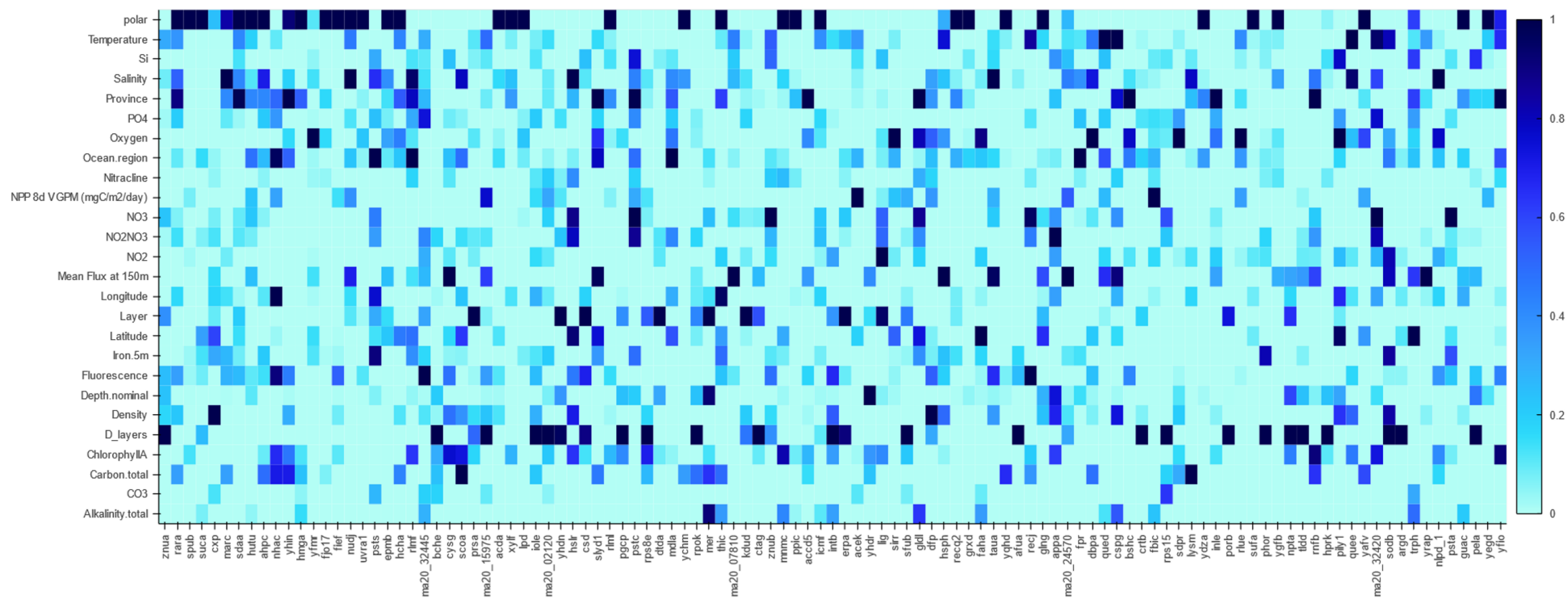




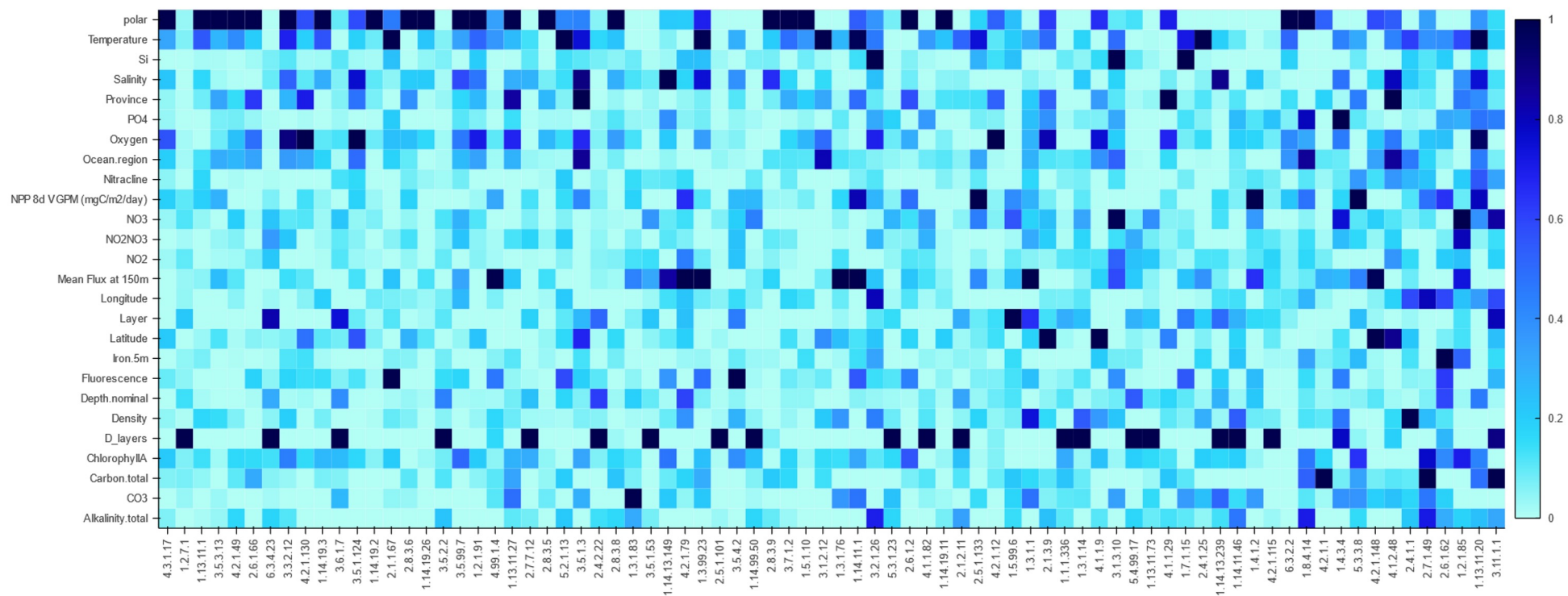


At the level of ocean descriptors

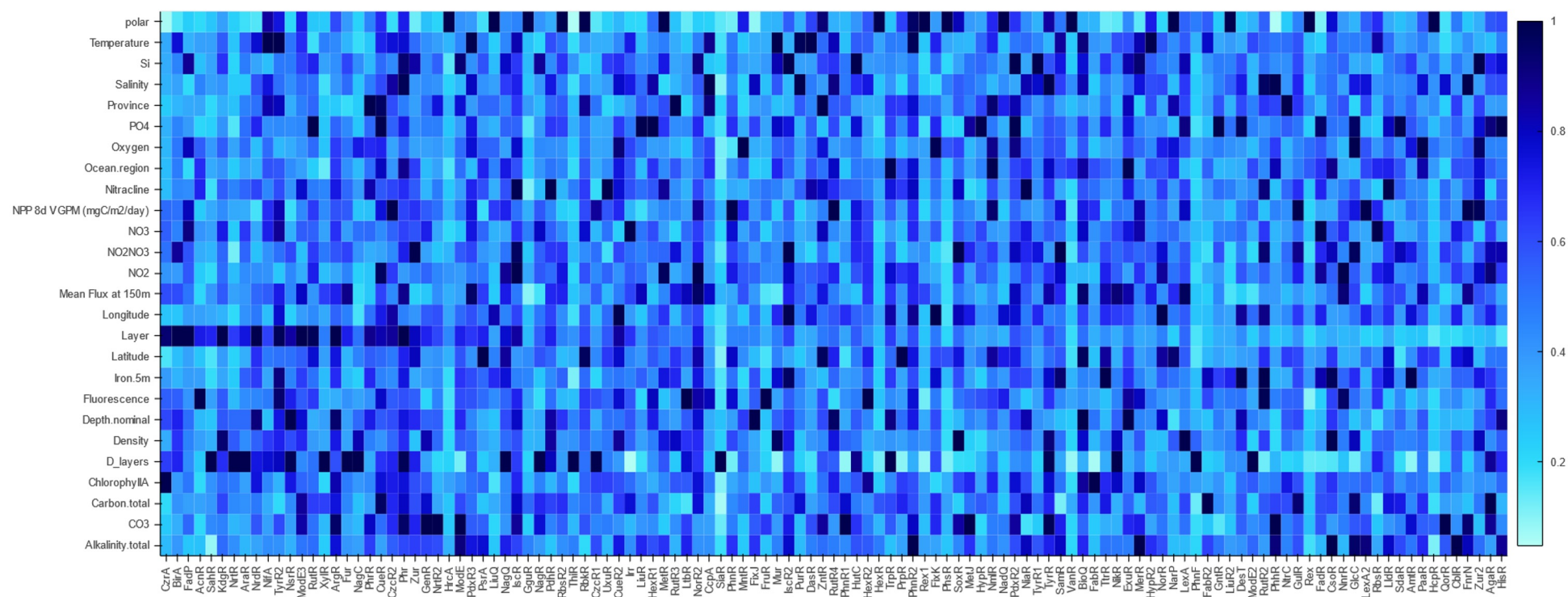
M0-TF-GEN Best 1% n = 32k Normalized



M2-TF-EC Best 3% n = 2,6k Normalized



Guidi-TF-TF all n = 89 Normalized





Thanks !

CEODOS
Chile