



0 stars 0 forks 1 watching Activity

Private repository

main



Branches Tags



Your main branch isn't protected

Protect this branch from force pushing or deletion, or require status checks before merging. [Learn more](#)

Protect this branch

Dismiss



Haradai updates last minute

1 minute ago 23

[View code](#)

README.md



Bio datathon UiB

Introduction:

We were presented with two files:

abundance_table.csv which contains the percentage of DNA for each species found in different samples which have been given different treatments across several replicas.

env_parameter_sample.csv which contains the environmental conditions for samples.

We had some proposed questions to answer but also were invited to explore our own. Some of the proposed questions are answered but we focused mainly in questions that we found interesting, so we did a bit of our own.

The code of this project you will find it in different notebooks, its not super well organised but everything is there.

- Initial exploration -> research_graphing.ipynb

- how treatments impact the populations project.ipynb
- classical ML in how treatments impact the populations -> learn_sample.ipynb
- experiments in DL to determine treatment by water conditions and regression of the %of dna values -> water2treatment
- experiments in DL to do a regression of dna% slope by treatment, specie, system and water conditions. -> dl.ipynb

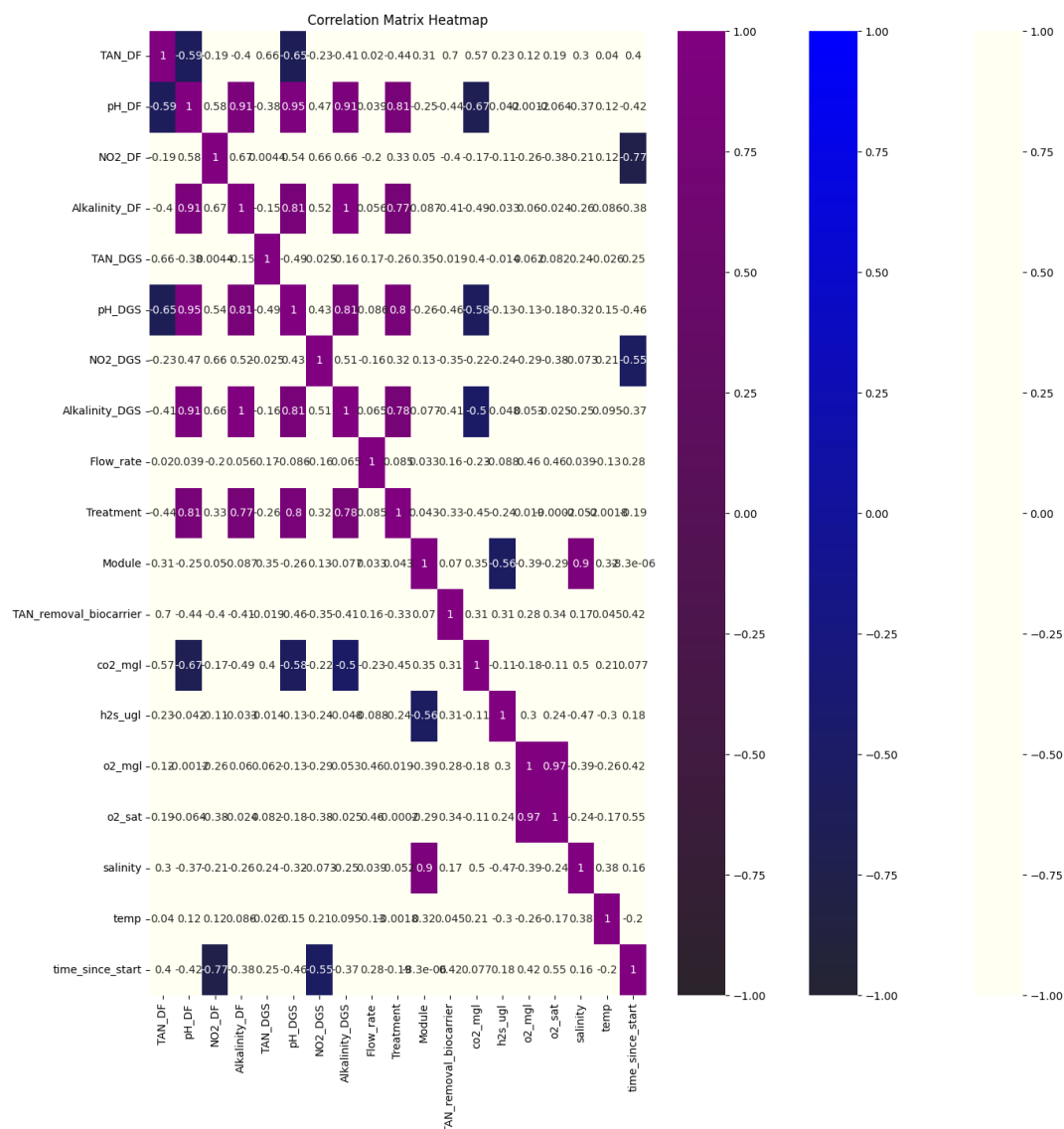
Initial exploration: [🔗](#)

Making a correlation plot over all values in the env_parameter_sample file can show the correlation between variables.

Purple = high positive correlation

White = low correlation

Blue = high negative correlation



We can see that some of the variables in the env_paramter_sample file have high degrees of correlation, but most variables no do not correlate. This initial exploration served us to see some of this environment data. We also noticed the little amount of data here and that timewise

What we want to answer [↗](#)

The instructions were pretty open regarding what we could do with the data. For us, it was interesting to find out how the **different treatments** impact the life thats in it, regarding the **concentrations of dna** found.

Initially, we thought about it broadly, hence we asked ourselves:

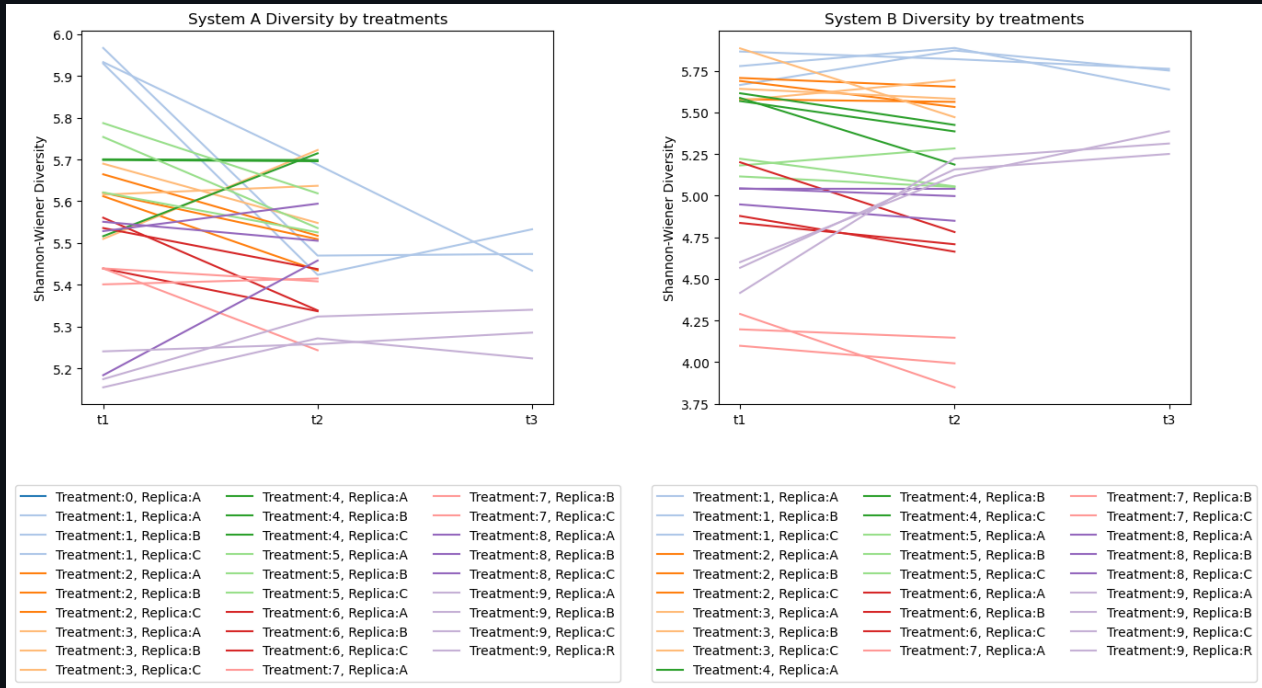
How is the biodiversity impacted by treatment? [↗](#)

First we had to define a measure of diversity.

Thanks to chatGPTs recommendation we found the **Shannon-Wiener Diversity Index (H')**. Which is a measure used to measure the diversity of species in a community.

This is defined by the equation $H = -\sum p_i * \ln(p_i)$ where p_i is the proportion of each individual in the community.

This is what we found:



We can see some treatments, such as treatment 1 that seem to have a different behaviour in system A than B. Also it is interesting to see how indeed the replicas have similar values and hence, behave similarly.

How do the different treatments impact the populations?



Again, looking at the dna percentage **by specie** in the samples we attempted to discover if:

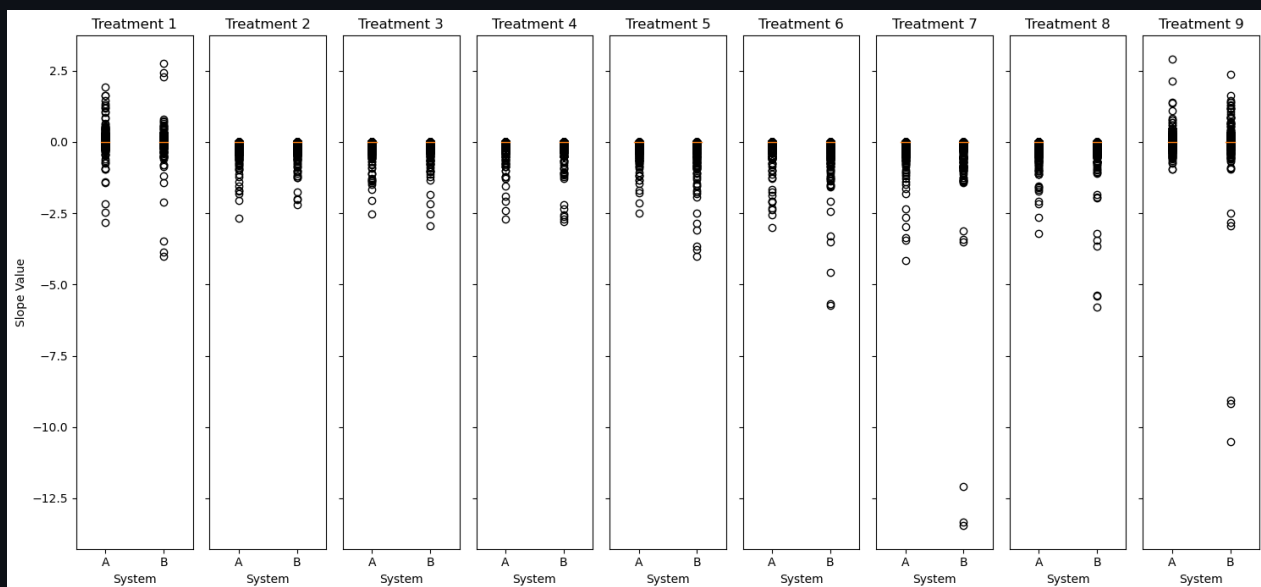
1. Treatments overall affect species population
2. What treatments affect the most

As a simple measure of what each population was doing in each sample we applied linear regression to output a slope, which would serve as an indicator of growth/decrease and its speed. Our dataset looked like this:

	id	Specie	System	Treatment	Replica	t1	t2	t3	slope	var	std	mean
0	Homo sapiens_A1_A	Homo sapiens	A	1	A	0.262329	0.327225	0.831947	0.284809	0.064824	0.254606	0.473834
1	Homo sapiens_A1_B	Homo sapiens	A	1	B	0.318878	0.447284	0.941915	0.311519	0.072147	0.268602	0.569359
2	Homo sapiens_A1_C	Homo sapiens	A	1	C	0.286041	0.518135	0.730994	0.222476	0.033018	0.181708	0.511723
3	Homo sapiens_A2_A	Homo sapiens	A	2	A	0.641026	0.587084	0.000000	-0.320513	0.084277	0.290305	0.409370
4	Homo sapiens_A2_B	Homo sapiens	A	2	B	0.729927	0.543478	0.000000	-0.364964	0.095881	0.309646	0.424468

[1.] Do treatments overall affect species population?

We first analyzed the distribution of slopes by treatment and system.



In order to determine how much were the treatments relevant in the changes of population trend(slope) we performed an ANOVA analysis. This way we can get a statistical significance "confidence" value regarding if a concrete specie is affected by the treatments.

Before starting though, we had to make sure the following, as this is some assumptions this method makes on the data. It is essential in order to be able to consider the results valid.

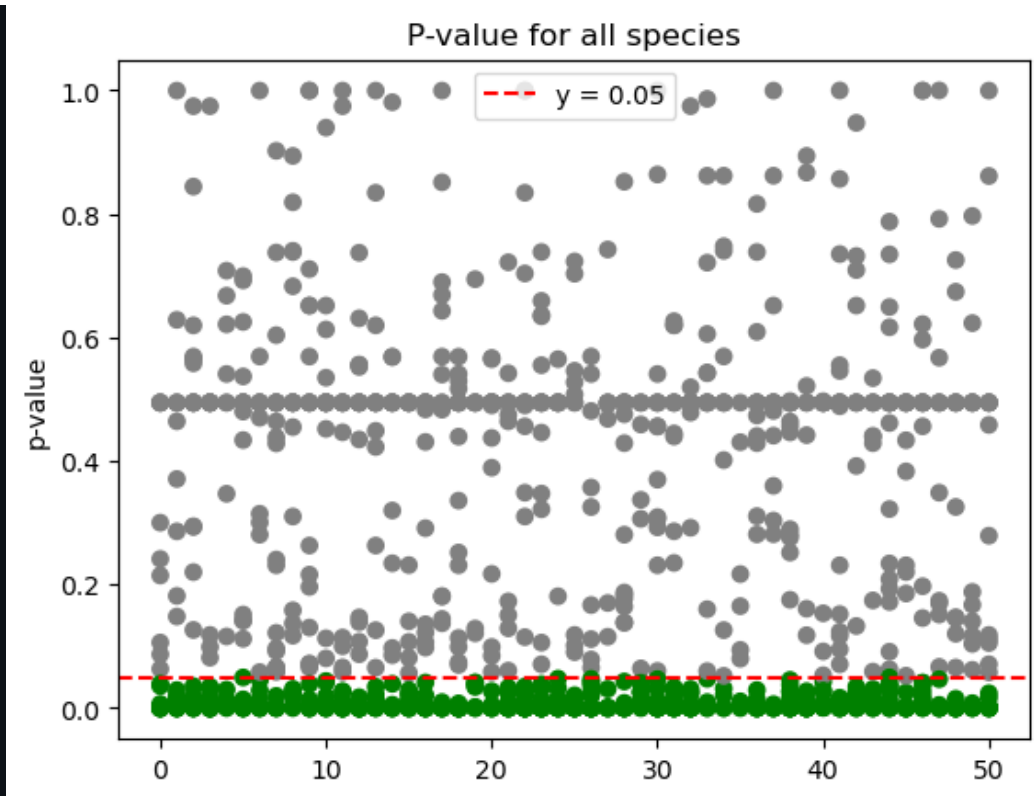
1. The observations are independent within and among groups. 2. The observations within each group are normally distributed. 3. The distributions from which the samples are drawn have the same finite variance.

What can we say about our data? 1. -> We have replicas, so therefore not really. We should only grab one replica of each. 2. -> We will assume this is true, that the dna is homogeniously distributed in the sample medium. 3. -> Again, as in 2 we will assume homogeneous medium.

After removing replicas, this is the result we got for the *Homo Sapiens Specie*. $p_value = 0.0011488636867922114$

Typically in ANOVA analysis a p_value of 0.05 is the threshold to determine if there is a statistical significance or not. In this case it is below, so we have some! Note: we utilized the function `f_oneway` from the *scipy statistics* package.

After this result, we continued to determine, how does this affect all other species? Hence, we did the exact same calculation for all



In the above plot, each dot represents a different specie, they are randomly plotted on the canvas maintaining a y-value equal to the calculated p-value.

What percent of species are significantly affected by the treatments? --> 4.35%

ML approach to steer formulating an hypothesis:

The purpose of this part is to attempt to create a classifier that given the amount of DNA per species over the three values, can determine which treatment was used.

Example data row:

id	Specie	System	Treatment	Replica	
Rhodocytophaga_rosea_B2_A	Rhodocytophaga rosea	157223	B	2	,

The coloumns that contained alphabetic sata such as "Treatment" being "B" was transformed into numeric values.

Initally, the data was skewed, and this proved to be a big issue. Upon initial prediciting, the model simply guessed the 9th treatment much too often, but it makes sense since this is a valid "strategy".

After removing the 9th treatment to see if the model could improve, it made the model slightly worse. So even with balanced treatment data, the model had a lot of issues predicting the correct treatment from the data.

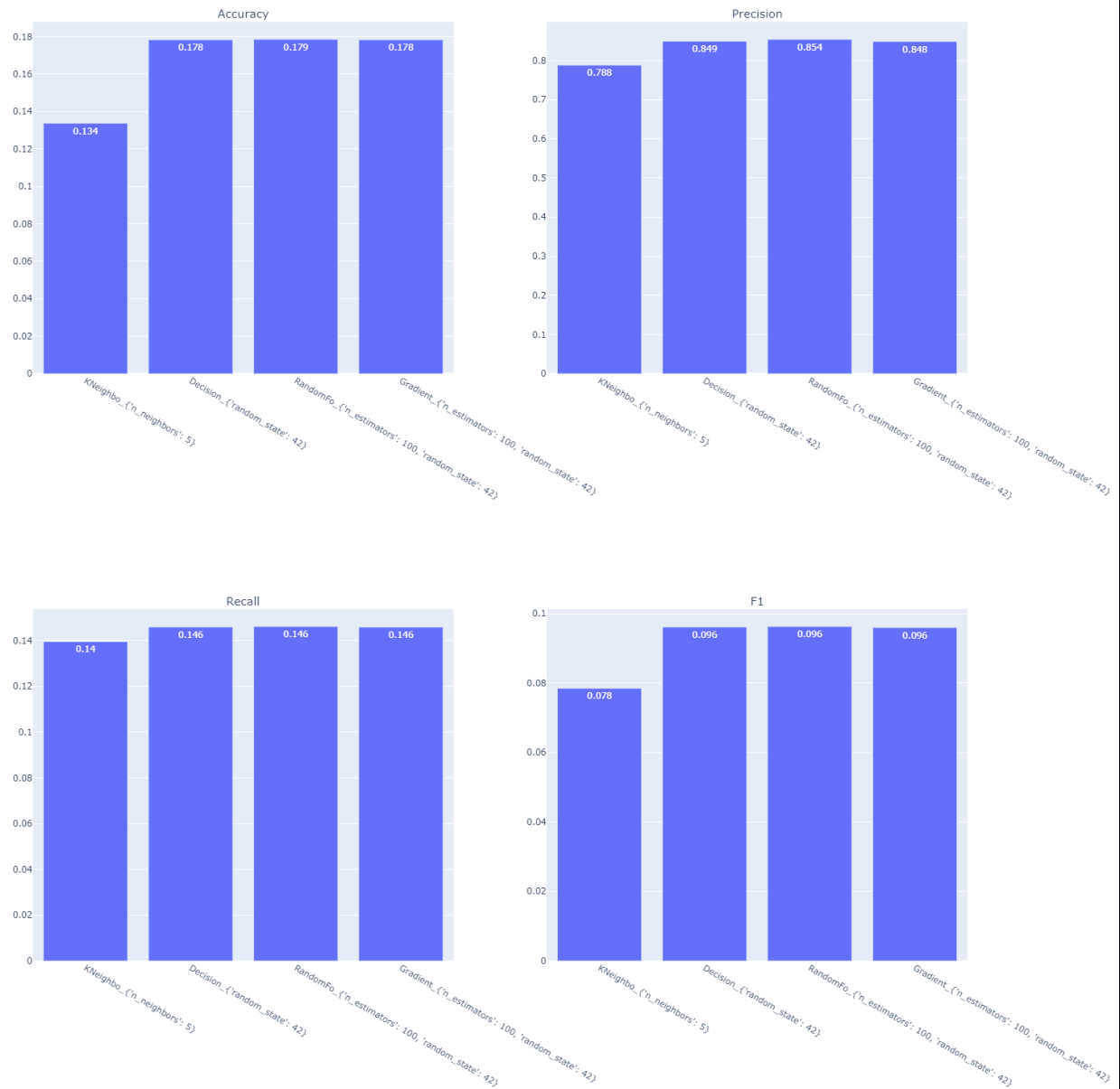
Treatment	Count	Percentage
1	73392	10.714
2	73392	10.714
3	73392	10.714
4	73392	10.714
5	73392	10.714
6	73392	10.714
7	73392	10.714
8	73392	10.714
9	97856	14.285

Treatment 9 is occouring much more often than the other treatments.

After attempting to classify which treatment comes from what values, it proved to be a bit of an issue.

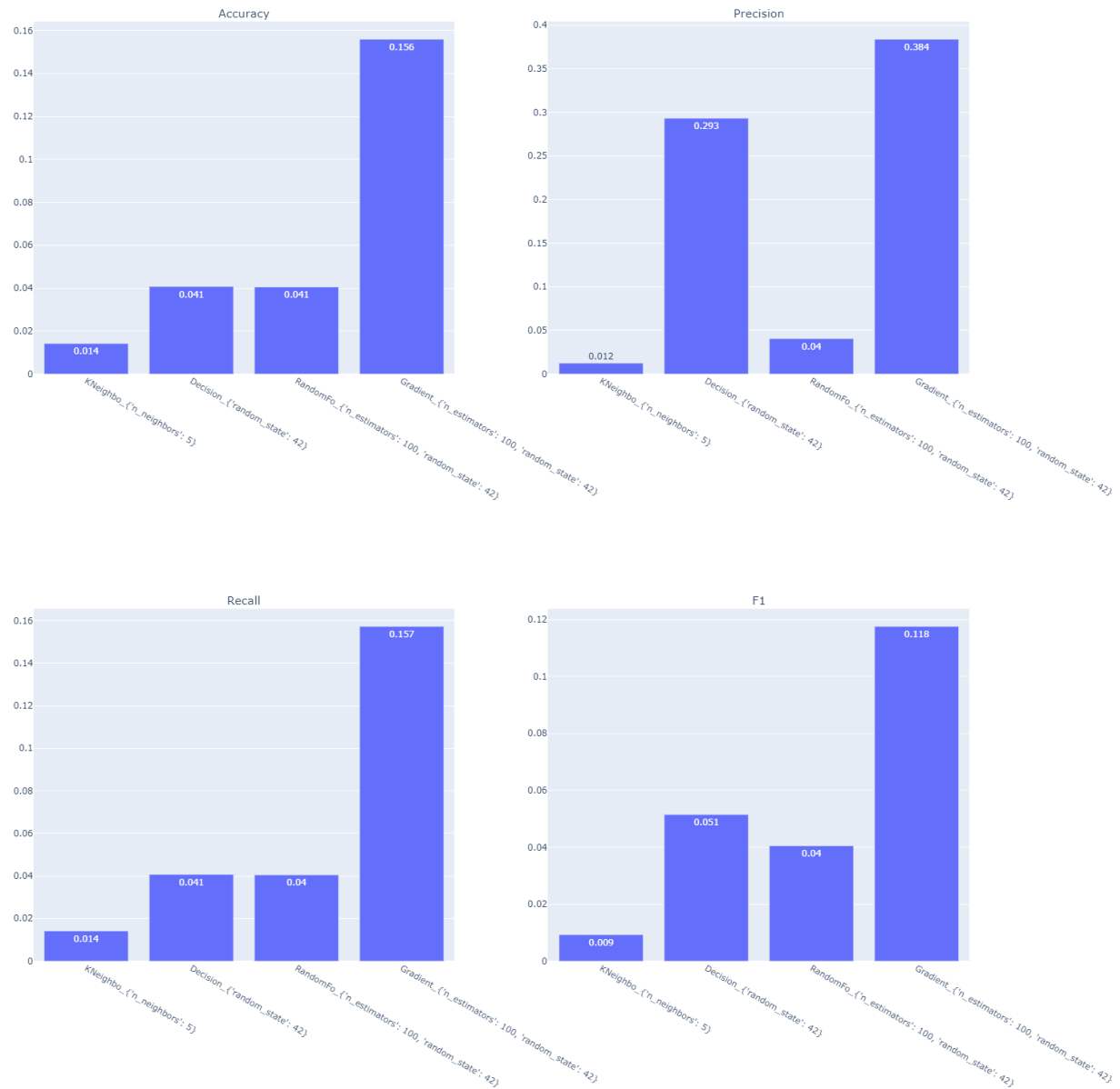
Generally, the models came up with an accuracy around 18%, this is a reflection of when the model guesses, what is the chance it guesses right? An accruacy of 18% is only slightly better than simply guessing treatment 9 every time.

Baseline model, 1/5 of data used, most data columns dropped



After removing treatment 9, the accuracy fell to 0.156. This means that the model is able to sometimes guess the treatment from the data, however the accuracy falling when removing treatment 9, simply shows the model guesses treatment 9 too often.

100 percent of data used, all data columns kept, treatment 9 removed



Overall, these models proved that simply determining treatment from the amount of dna from each species, is a difficult task, and perhaps since DNA percent differs so much between each species, there may not be a clear correlation.

How does treatment affect water conditions? [🔗](#)

In order to answer this question we attempted to create some simple ML classifiers, SVM and KNN. They attempted to classify the applied treatment by using the water conditions as features. The idea is that if we are able to make a model with a decent accuracy that **could** mean that there is some relation between them.

Results: We didn't manage to classify. We cannot conclude from this that the inability to classify means that there is no relation so we could consider it non conclusive. Anyhow, talking to *Natacha* she mentioned that the scientists taking the samples were trying to make the conditions stable, which aligns with having no relation between them.

What water conditions are related to the increase or decrease of certain species? [↗](#)

Note: What water conditions?

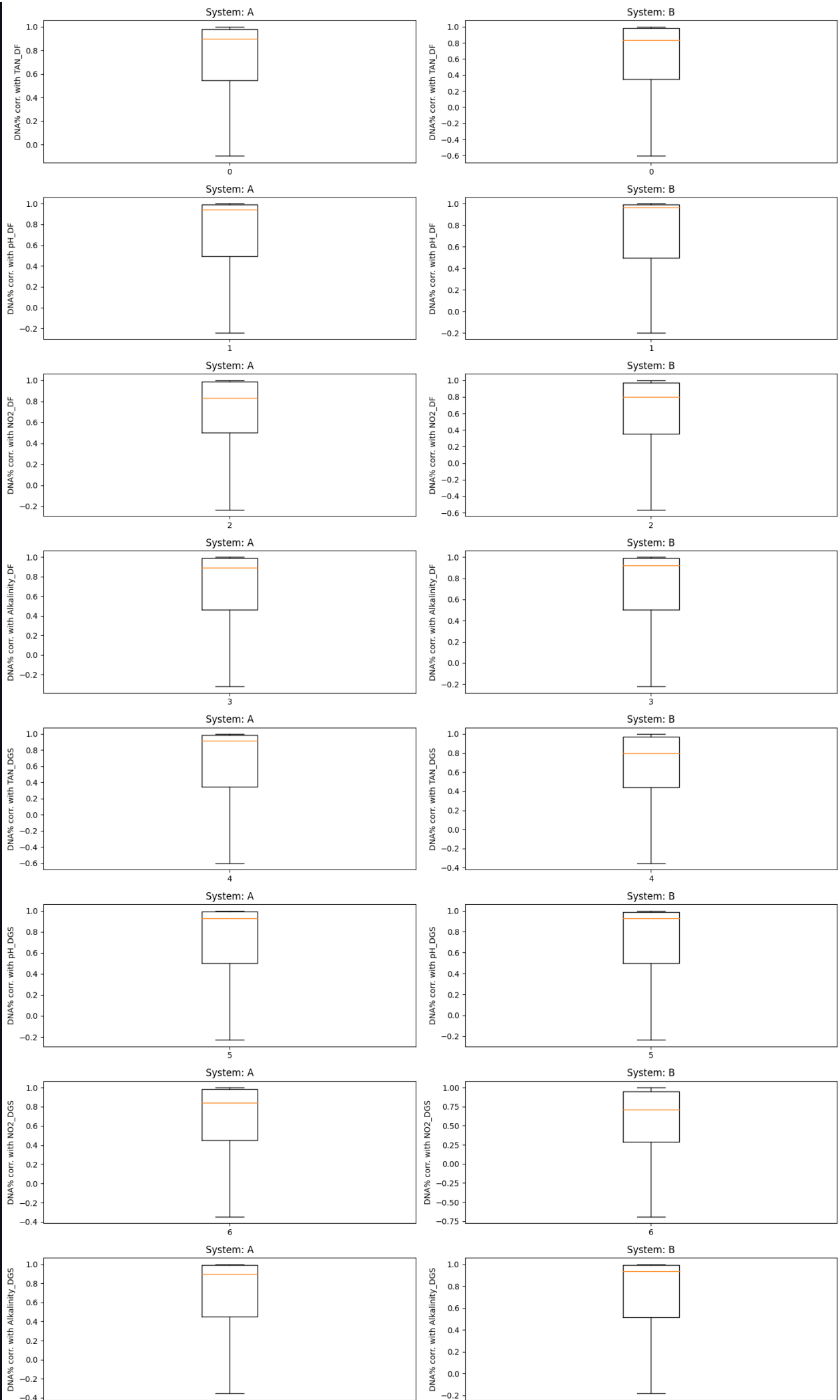
- TAN_DF
- pH_DF
- NO2_DF
- Alkalinity_DF
- TAN_DGS
- pH_DGS
- NO2_DGS
- Alkalinity_DGS
- Flow_rate
- TAN_removal_biocarrier
- co2_mgl
- h2s_uql
- o2_mgl
- o2_sat
- salinity
- temp

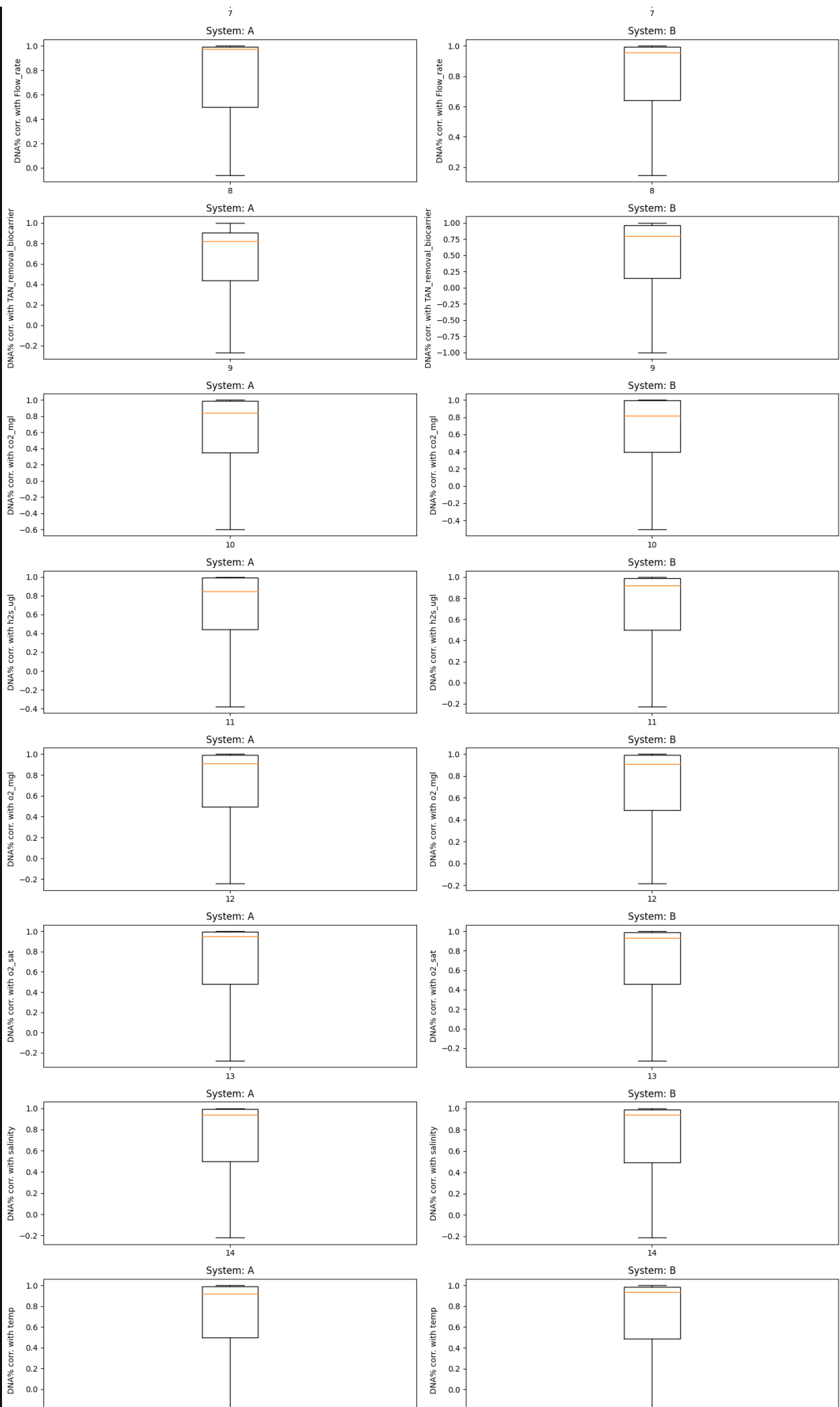
We computed the **correlation** for each of the **water conditions** with the **percentage of dna** per sample, per species. A lot of data. We wanted to do this in order to later on, be able to analyze the correlation of each water condition with the dna %:

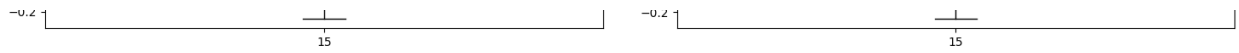
- per system
- per treatment

We will look per water condition the **distribution** of these correlations in each **system group** and **treatment group**. We **could** even analyse for **specific species** we were interested on* We could even analyse per sample the different replicas distribution, in theory they should be pretty packed together.

Per System:

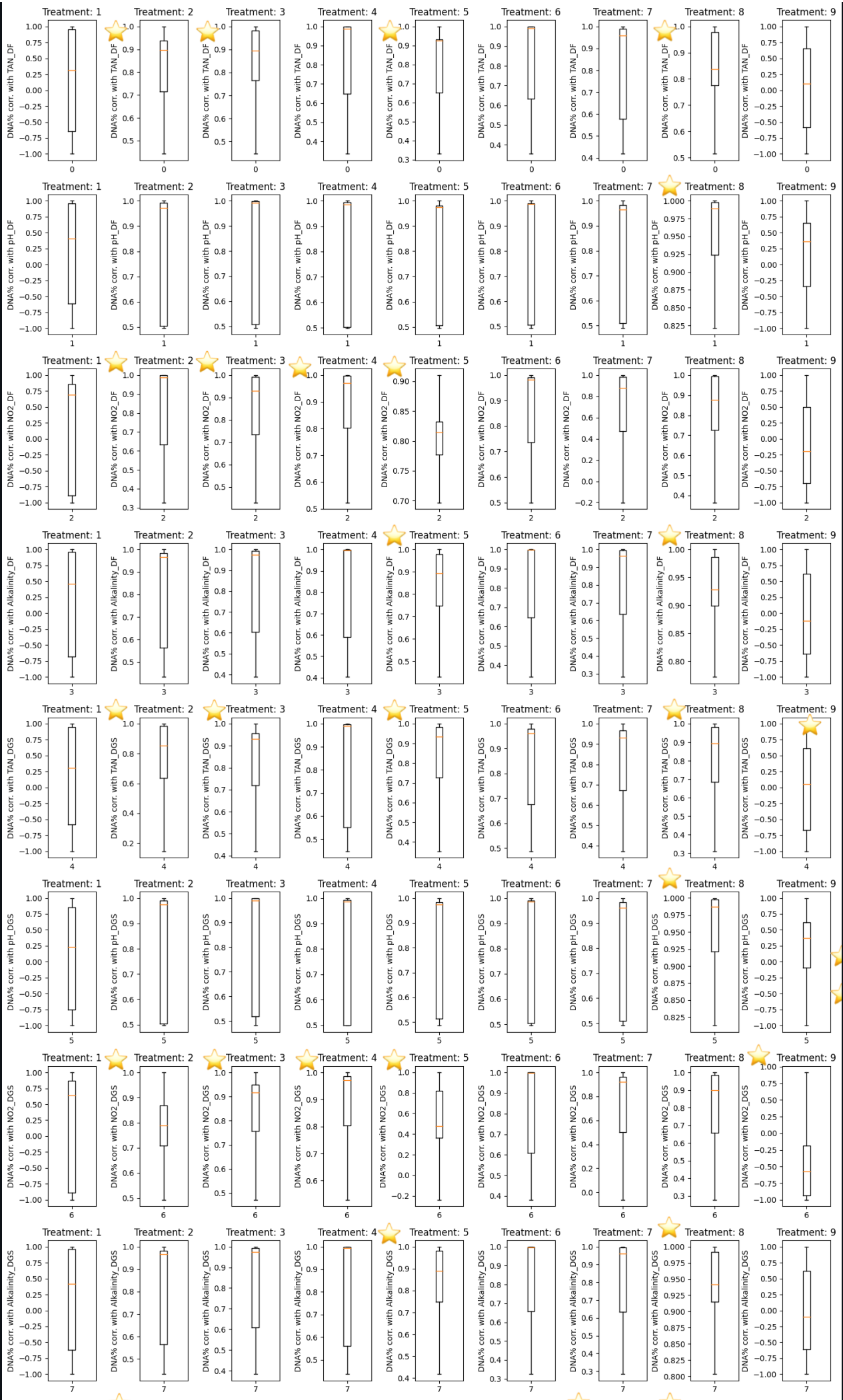


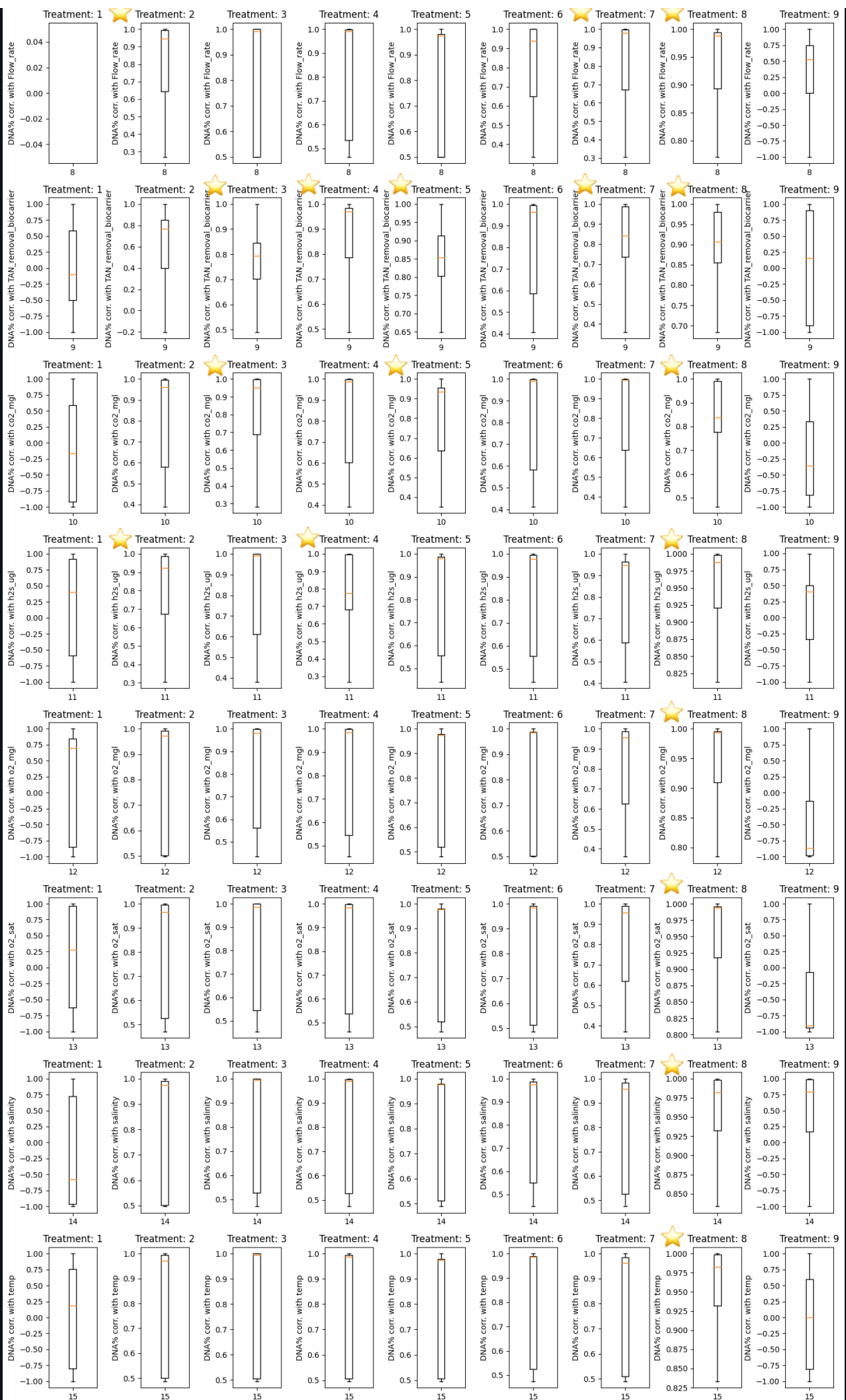




From this plot we cannot take a lot of conclusions due to none of them seem to have a clustered distribution, they all variables seem to have a positive correlation with the % of dna when grouping by system. Anyhow this is useful, because it serves as a baseline for the next plot, where we look by treatment.

Per Treatment:





This plot is already quite interesting, we find on some treatment groups that some variables have a pretty clustered correlation distribution with % of DNA. We marked with stars the ones that we considered most clustered. This could be more precisely detected by analyzing the std of the distribution but didn't find it necessary. Furthermore, note that these difference in clusterization of distributions are not due to difference of quantity of samples, as before we saw that the amount of samples per treatment is except for treatment 9 which is a bit more equal among 1-8.

We believe this way of seeing plot that we used to view the different variables correlation with % of dna by treatment could be very useful when studying specific species. This plot could show indications of what conditions might be changing due to some treatments together with the population of some specie. e.g. We wanted to apply a treatment to

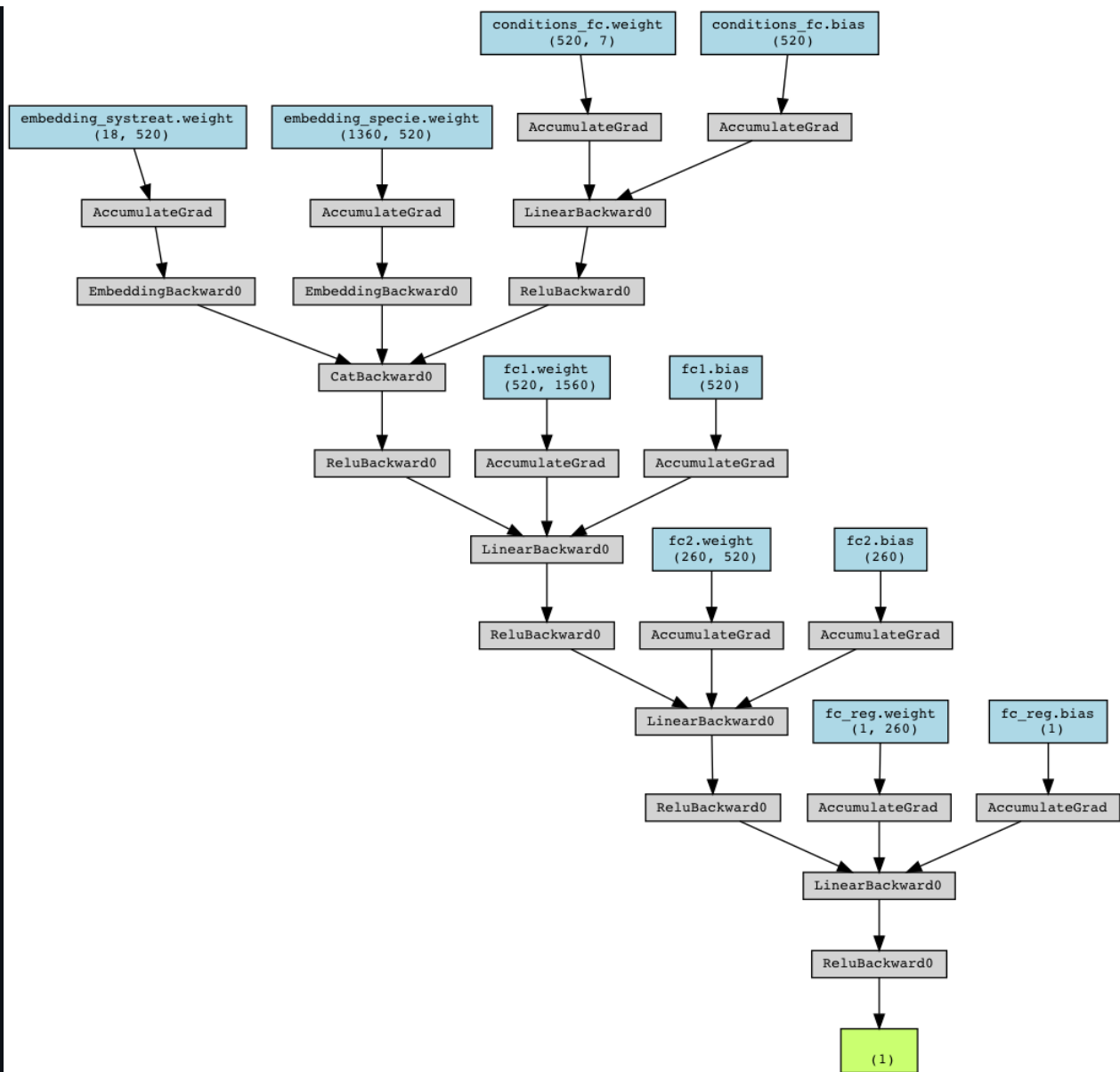
A model to determine if given a specie, treatment and environmental data if the population (% dna) will increase or decrease. Regression of the slope.

Instead of giving the model raw the environmental data, we will give its slope, this makes sense as we are already trying to predict a slope.

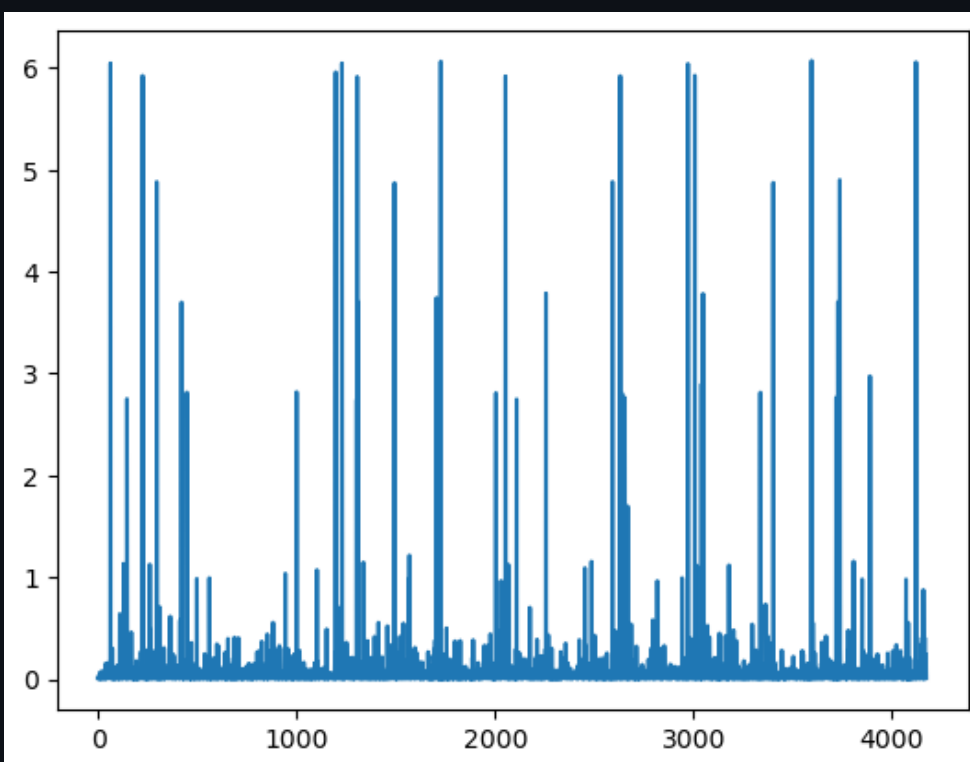
The environmental data that we will feed in, was manually selected after seeing the correlations to %dna by treatment that seemed to be quite clustered among different groups.

The model that I attempted initially was quite simple, two different embeddings, one for different species and one for the different combinations of system and treatment. Fully connected layers following that up to 1d to do regression. The loss was Mean Squared error.

Note: I understand that these diagrams are quite verbose and not so clear. Because of the lack of time we couldn't properly draw them.

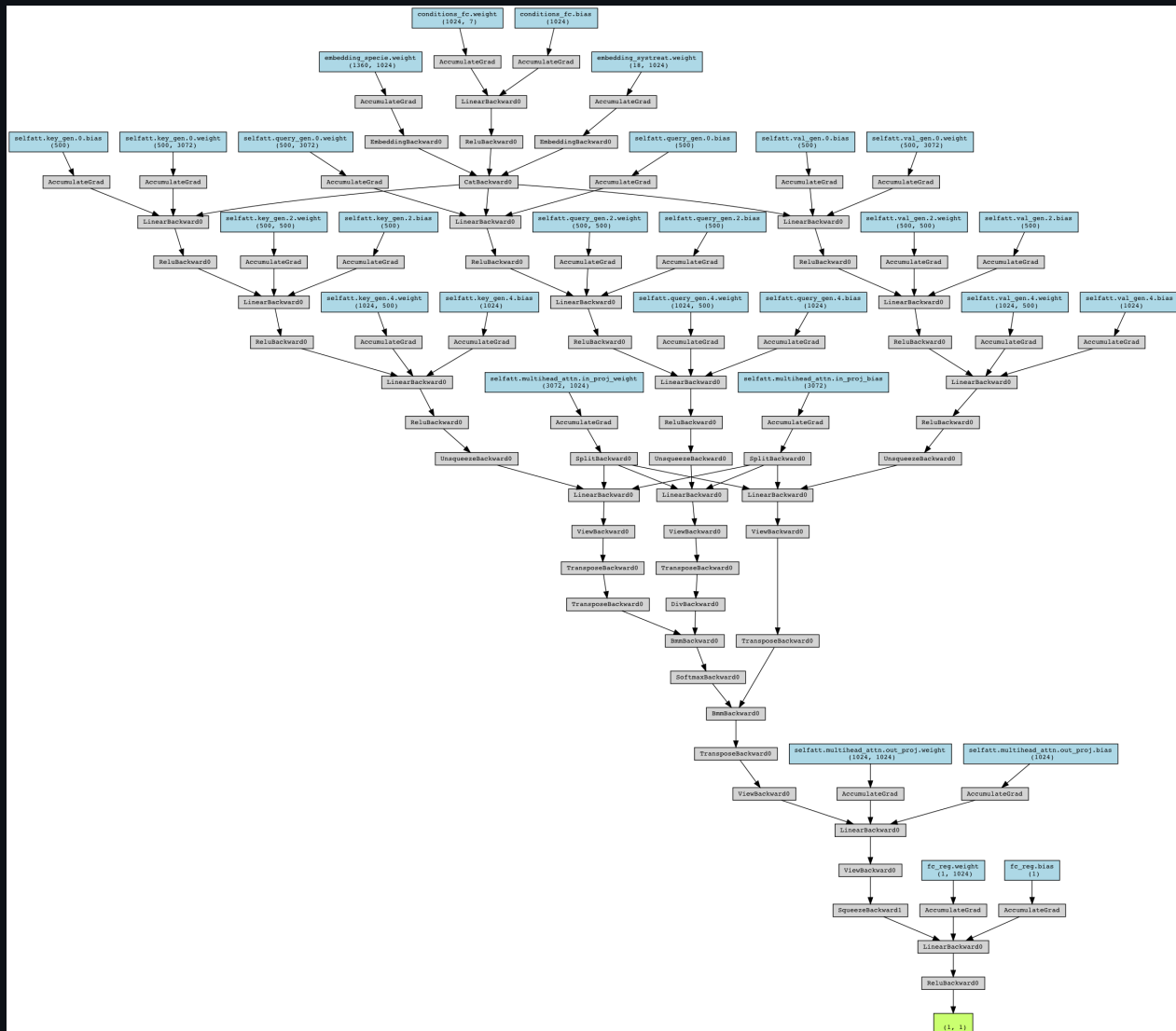


The results using this architecture were not satisfactory as it is visible in the following loss plot.

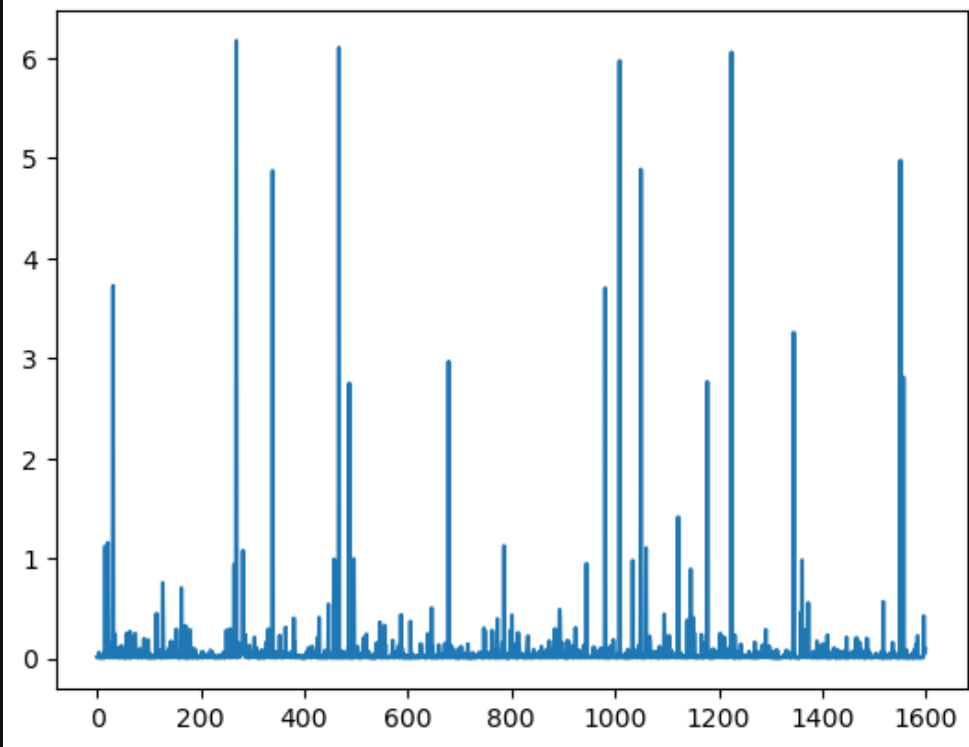


Going for a more complex model, **self attention block!**

The previous architecture was modified to incorporate a selfattention head that hopefully would be able to learn more complex patterns.



Results: Better, but still quite poor, doesn't seem to learn much at all. Some batches have very low error but it doesn't seem able to perform better every epoch for all batches.



Conclusion / findings: [🔗](#)

From the experiments we runned and the analysis of the data we can determine some insights regarding the questions we asked ourselves:

How is the biodiversity impacted by treatment? *We found it non conclusive, there might be some insight in the plot about a few treatments but its not very clear. Might be useful when trying to study specific species though.*

How do the different treatments impact the populations?

- *We clearly found that treatments impact on the dna % of species*
- *We found that the impact is on a number of specific species. Could be interesting to see if these species could be from similar or very diverse taxonomic groups*
- *We found some quite relevant correlations between the water conditions variables and the %of dna over time. Grouping by treatment showed some key water conditions that might play a role in different the treatments, not sure about if its a factor or a result of the treatment but that is details from the experiment that we don't know.*
- *In our attempts to build DL models that related the treatments,species and water conditions to the trend of the %of dna for each specie we found it was very hard for the model to learn. We also think that in further exploration, one could train a model for specific species or a subset that we know might be relevant as in the training process that we did we probably had too much noise from nan values that we filled with 0 and modified a lot the slope and simply species that were not affected at all by the treatments and were in very low quantities.*

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Contributors 2



Haradai



Freskoko Henrik B

Languages



Suggested Workflows

Based on your tech stack



Python package

Create and test a Python package on multiple Python versions.

Configure



Django

Build and Test a Django Project

Configure



Python Package using Anaconda

Create and test a Python package on multiple Python versions using Anaconda for package management.

Configure

[More workflows](#)

[Dismiss suggestions](#)