Identifying the Drivers of Protein Environment Classification

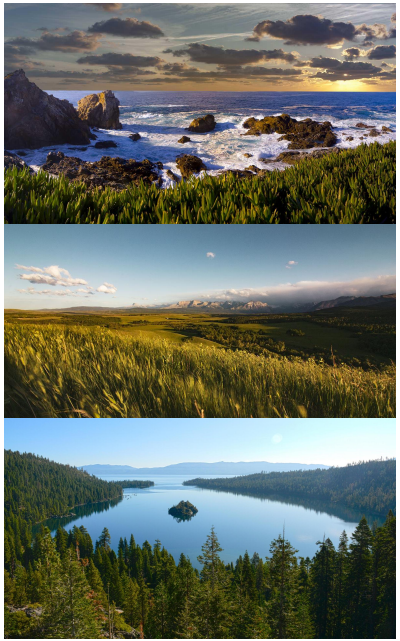
By Sophie Yeh, Edward Kirton, Delaney Scheiern, Haibi Lu

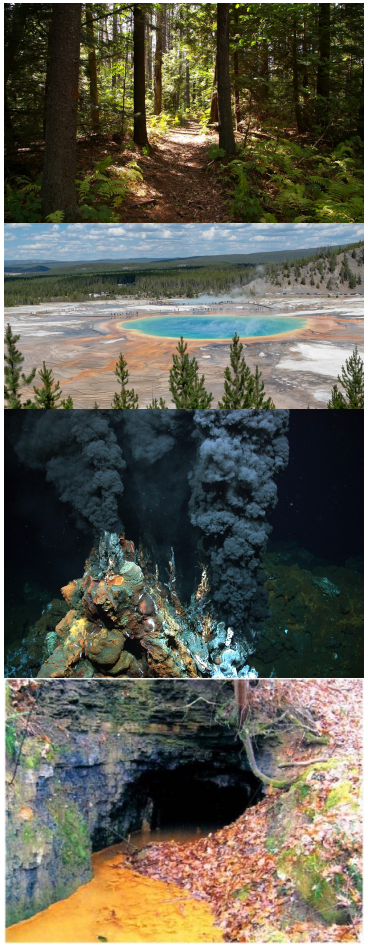
Outline

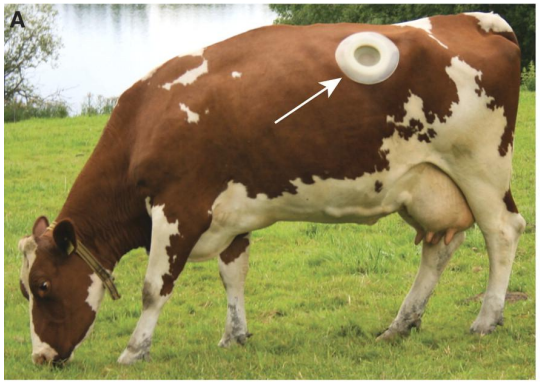
1. Intro - 1.5 min - Ed
2. EDA: class imbalance - 4 min - Haibi
3. Final Model Process - 7 min
    a. Why neural net & other baselines - 1 min - Haibi
    b. Optuna - 2 min - Ed
    c. XGB - 2 min - Sophie
    d. NN - 2 min - Delaney
4. Evaluation with other models - 2.5 min - Sophie
5. Explainability - 3 min - Delaney
6. Conclusions - 2 min - Ed

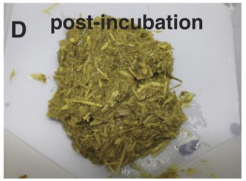
Motivation

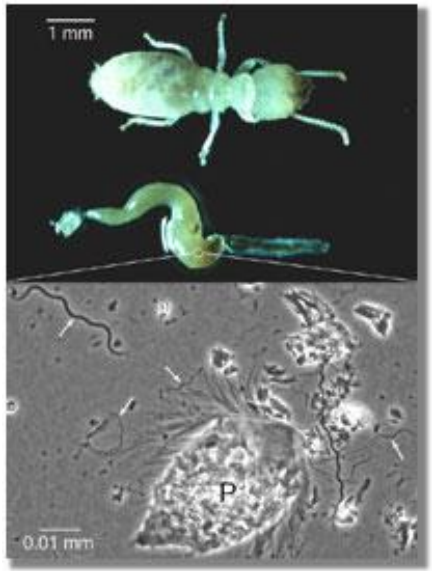
Labels: DNA samples from various environments

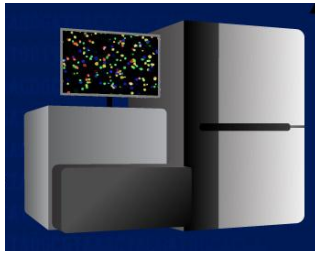




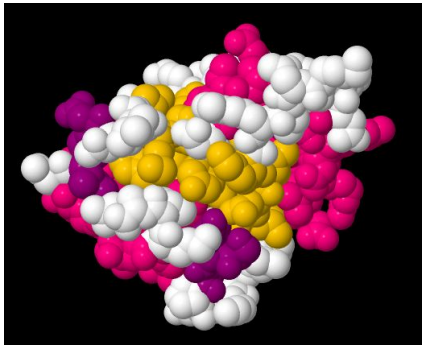


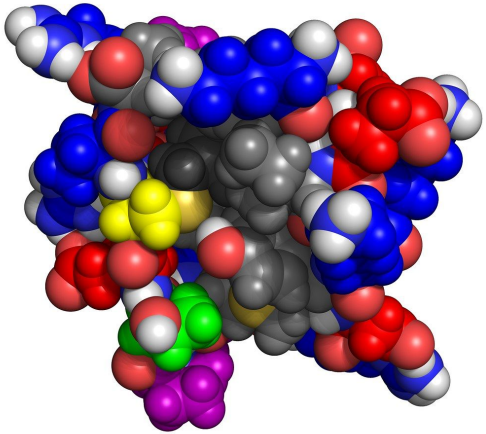


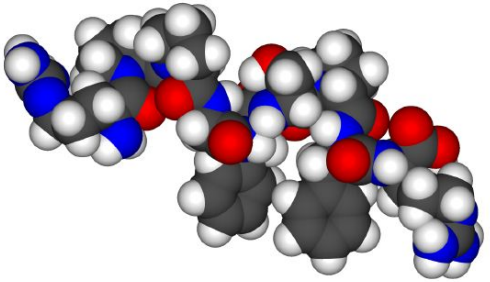


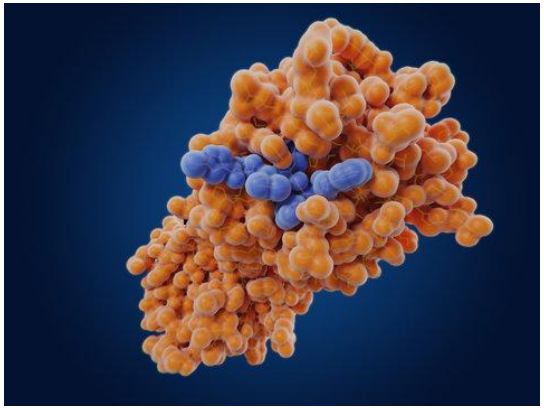


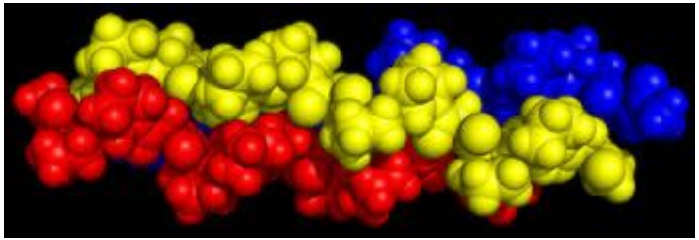
Features: Samples each have 16k protein counts

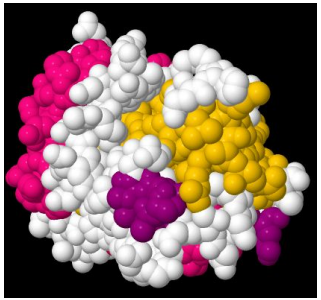












Data

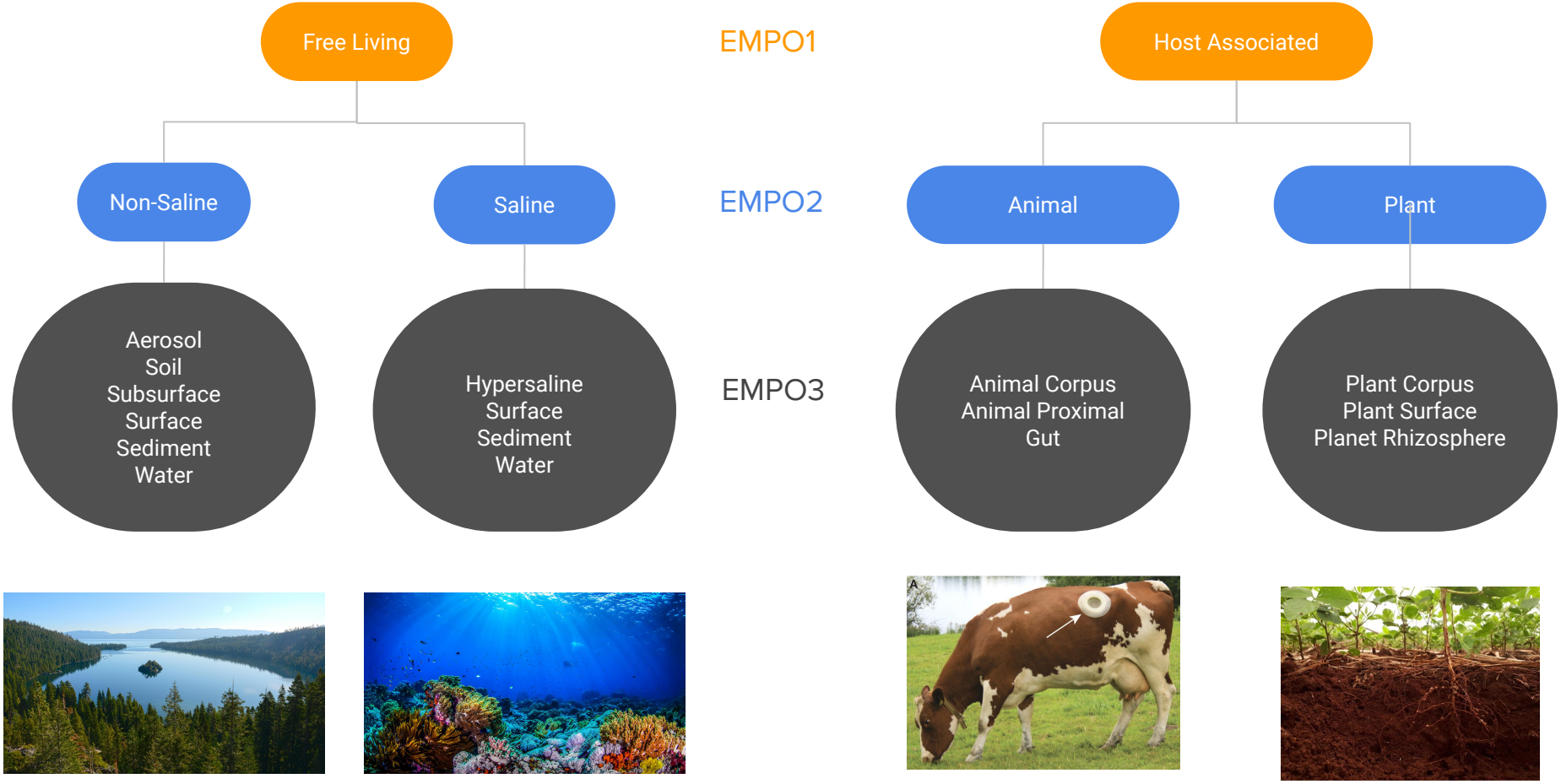
Exploratory Data Analysis

Data Source:

- Joint Genome Institute Online Database
  - https://gold.jgi.doe.gov/index

Data Features and Labels

- Features: Protein Family Ids (Pfam): Known protein sequences
- Labels: Environments
  - EMPO1
  - EMPO2
  - EMPO3


Free Living
EMPO1
Host Associated
Non-Saline
Saline
EMPO2
Animal
Plant
Aerosol
Soil
Subsurface
Surface
Sediment
Water
Hypersaline
Surface
Sediment
Water
EMPO3
Animal Corpus
Animal Proximal
Gut
Plant Corpus
Plant Surface
Planet Rhizosphere

Exploratory Data Analysis

Data Shape

16306 Pfam

1785

PF00001.19	PF00002.22	PF00003.20	PF00004.27	PF00005.25	PF00006.23	PF00007.20	PF00008.25	PF00009.25	...
0	0	0	2649	14350	1225	0	0	2214	...
0	0	0	662	3805	293	0	0	515	...

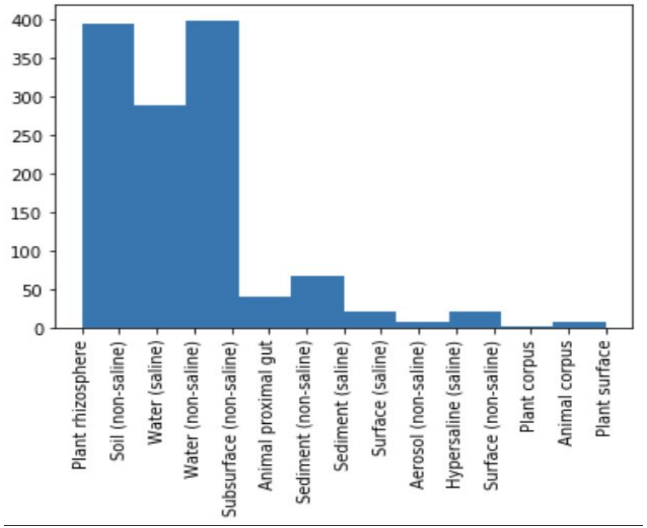
PF00001.19	PF00002.22	PF00003.20	PF00004.27	PF00005.25	PF00006.23	PF00007.20	PF00008.25	PF00009.25
0.000000e+00	0.000000e+00	0.0	0.004020	0.006243	0.001039	0.0	0.000000	0.003265
0.000000e+00	0.000000e+00	0.0	0.003575	0.013739	0.001026	0.0	0.000000	0.002235

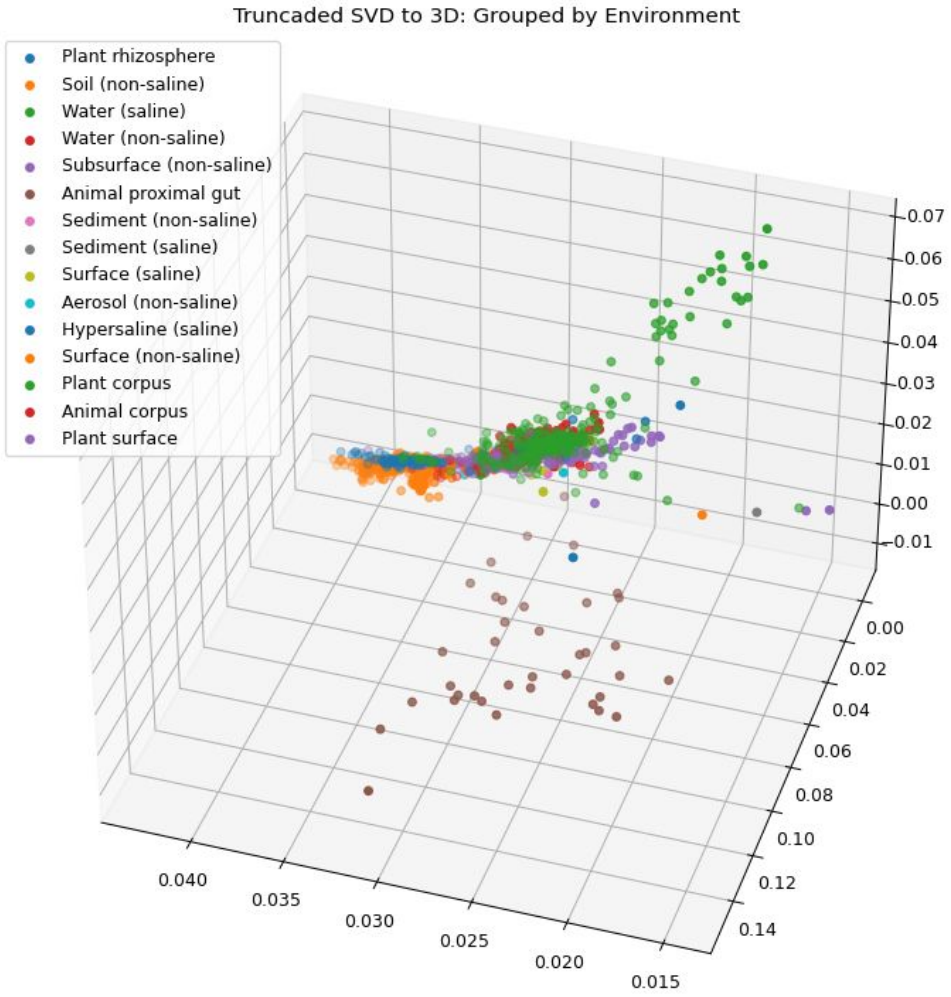
Exploratory data analysis

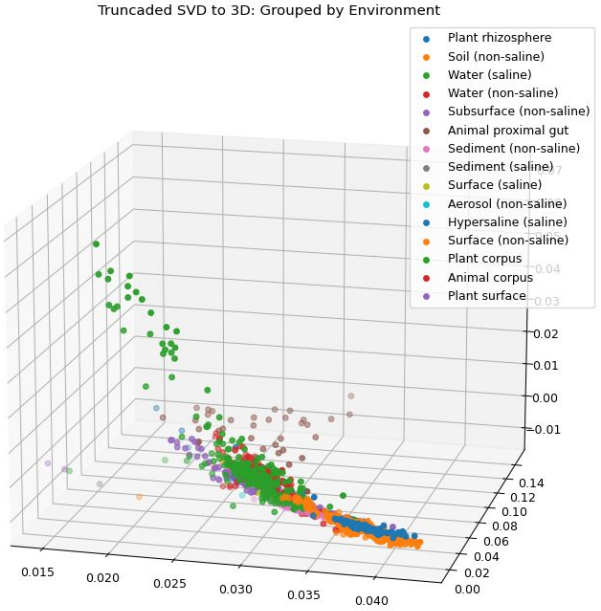
Training-Test Data Split

- 70/30 Split Training - Test
- 80/20 Split Training - Validation
- Label Distribution In Training Dataset

EMPO_1	EMPO_2	EMPO_3	
Free-living	Non-saline	Aerosol (non-saline)	8
		Sediment (non-saline)	73
		Soil (non-saline)	303
		Subsurface (non-saline)	144
		Surface (non-saline)	5
		Water (non-saline)	245
	Saline	Hypersaline (saline)	15
		Sediment (saline)	14
		Surface (saline)	6
		Water (saline)	290
Host-associated	Animal	Animal corpus	4
		Animal proximal gut	39
	Plant	Plant corpus	1
		Plant rhizosphere	96
		Plant surface	6







Truncated
Singular Value Decomposition (SVD)

Approach

Modeling Process Overview

- Trained Models
  - Decision Tree (Baseline)
  - XGBoost
  - Neural Networks

- Metric of Interest
  - Weighted F1 Score

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})}$$

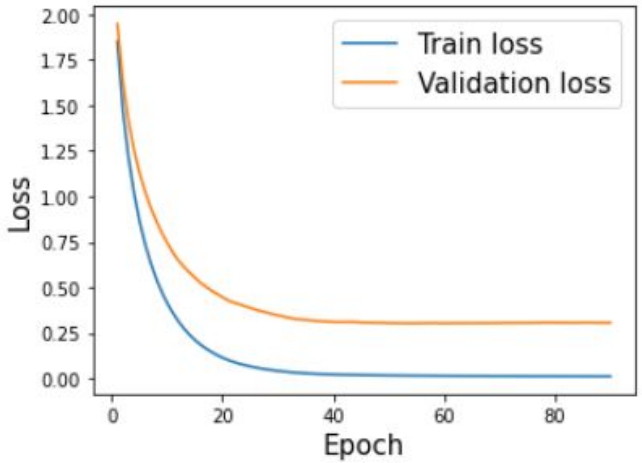
Hyperparameter tuning

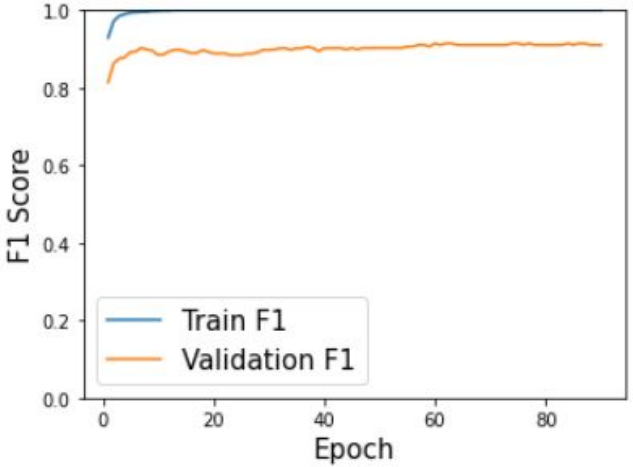
OPTUNA

```
def train_and_evaluate {...}

def objective(trial):
  learning_rate = trial.suggest_float('learning_rate', 0.0001, 0.01)
  layer_size = trial.suggest_int('layer_size', 20, 1024)
  dropout = trial.suggest_float('dropout', 0, 0.1)
  F1 = train_and_evaluate(learning_rate, layer_size, dropout)
  return F1

study = optuna.create_study(
  direction="maximize",
  sampler=optuna.samplers.TPESampler(),
  pruner=optuna.pruners.MedianPruner(
    n_warmup_steps=20)
)
study.optimize(objective, n_trials=100)
print(study.best_params)
```

XGBoost

```
{'reg_lambda': 0.01505853587641787,
 'reg_alpha': 0.03476744563778924,
 'colsample_bytree': 0.7,
 'learning_rate': 0.14665385125779304,
 'max_depth': 6,
 'min_child_weight': 2}
```





1. Baseline model is very similar but Optuna helped to fine-tune the parameters.
2. Curves are ideal with some overfitting to ensure model complexity is maximized.
3. Pros of XGB: fast setup and training, simple model, high evaluation metrics.

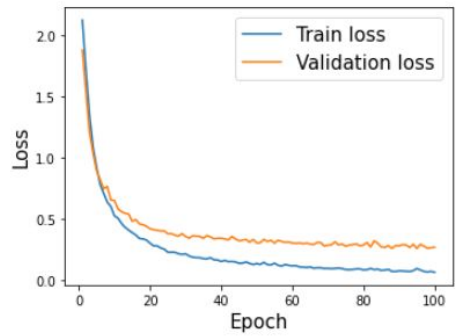
Feedforward Neural Network

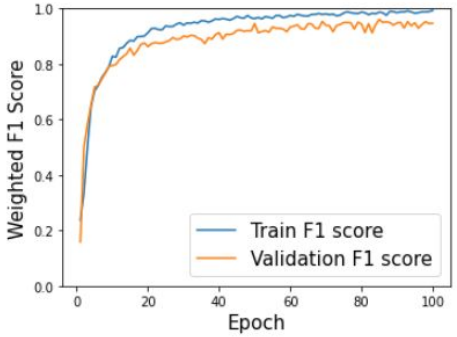
```
Validation F1 Score of 1 Layer: 0.9466294050216675
Validation F1 Score of 2 Layer: 0.9490211009979248
Validation F1 Score of 3 Layer: 0.9133524894714355
```

Optuna hyperparameters for 2 hidden layers:

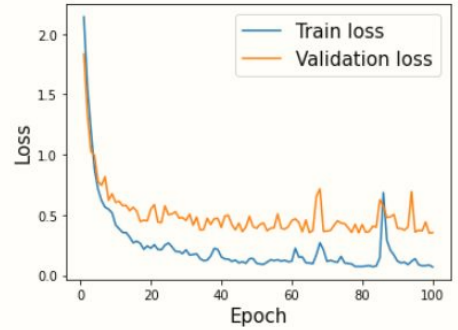
```
{'learning_rate': 0.006729391788774663,
 'hidden_layer_size1': 315,
 'hidden_layer_size2': 826,
 'kernel_regularizer': 4.523985086552334e-05,
 'bias_regularizer': 4.851063379560055e-05,
 'activity_regularizer': 3.62120961768956e-05,
 'dropout': 0.06862369288330496}
```

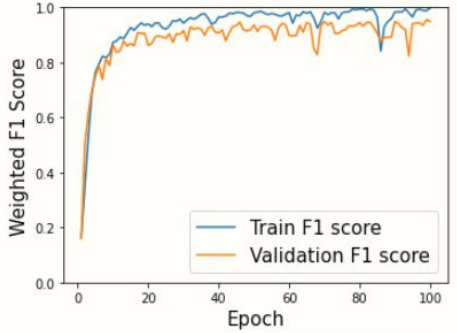
1 Layer



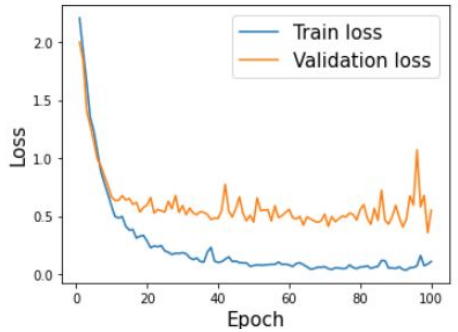


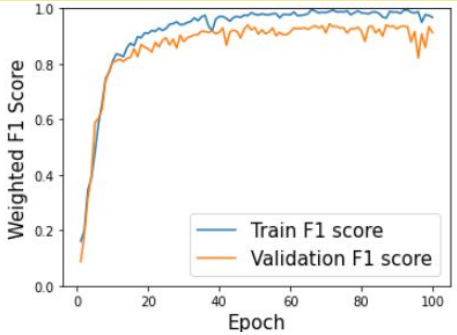
2 Layers





3 Layers





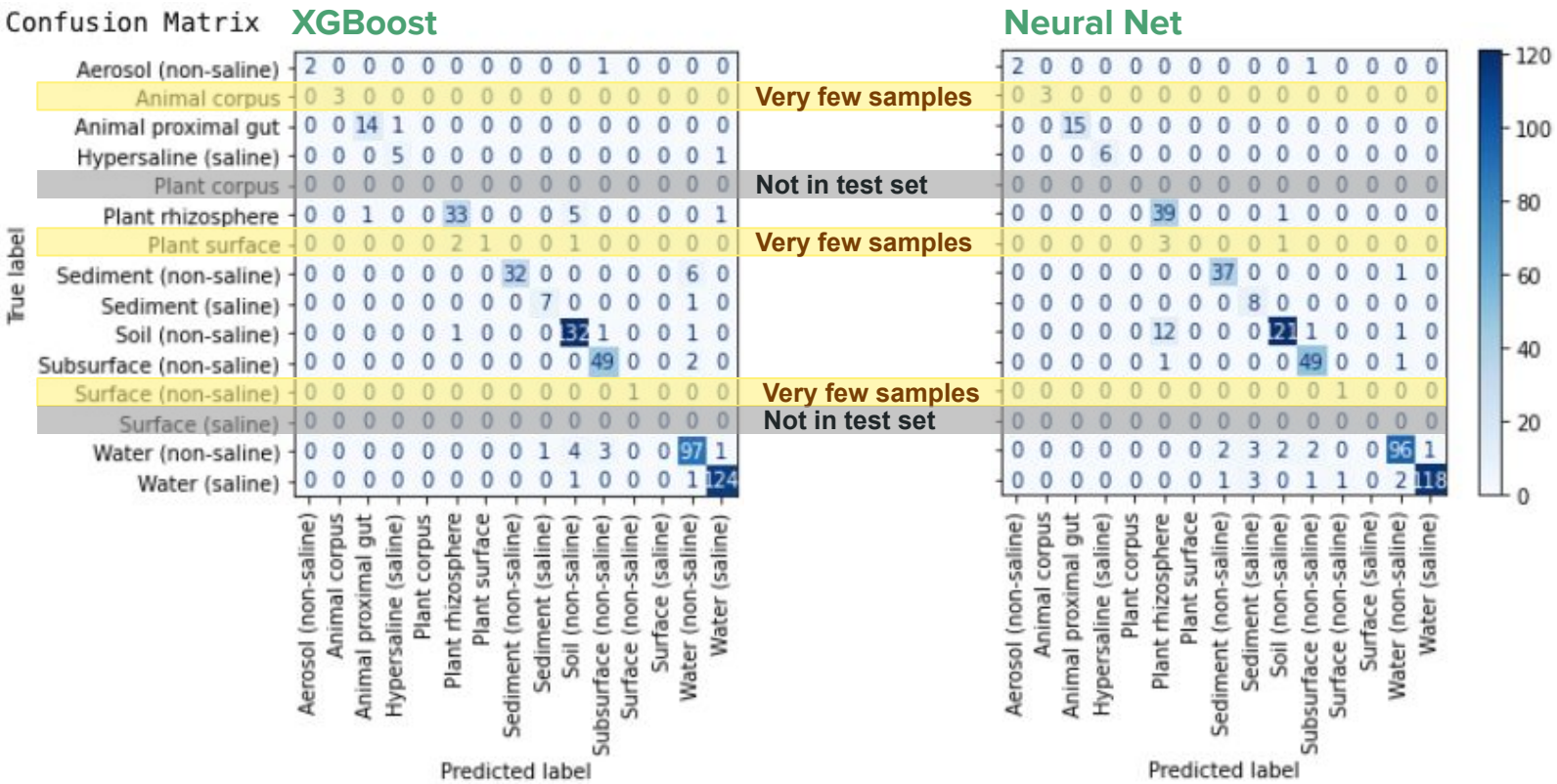
Model Evaluation

Overall performance across models on test set

	Decision Tree	XGBoost	Neural Network
F1 weighted avg	**0.8569**	**0.9308**	**0.9230**
Pros	Simple & fast	Simple & fast, performs better than DT	Complex, opportunity for better explainability

Test Set Confusion Matrix

Metrics can be skewed by
lack of training samples in a label for the training data
& no missing labels in the test set. **More data is needed.**



Classification Report: F1-Score

XGB and **NN** performed similarly well across labels, but did not attempt to predict small classes like **Decision Tree.**

Decision Tree

Classification Report

	precision	recall	f1-score	support
Aerosol (non-saline)	0.0000	0.0000	0.0000	3
Animal corpus	1.0000	1.0000	1.0000	3
Animal proximal gut	0.9286	0.8667	0.8966	15
Hypersaline (saline)	0.7143	0.8333	0.7692	6
Plant corpus	0.0000	0.0000	0.0000	0
Plant rhizosphere	0.7895	0.7500	0.7692	40
Plant surface	1.0000	1.0000	1.0000	4
Sediment (non-saline)	0.9091	0.7895	0.8451	38
Sediment (saline)	0.3077	0.5000	0.3810	8
Soil (non-saline)	0.9084	0.8815	0.8947	135
Subsurface (non-saline)	0.8913	0.8039	0.8454	51
Surface (non-saline)	0.3333	1.0000	0.5000	1
Surface (saline)	0.0000	0.0000	0.0000	0
Water (non-saline)	0.8716	0.8962	0.8837	106
Water (saline)	0.8615	0.8889	0.8750	126
accuracy	0.8526	0.8526	0.8526	0
macro avg	0.6343	0.6807	0.6440	536
weighted avg	0.8641	0.8526	0.8569	536

XGBoost

Classification Report

	precision	recall	f1-score	support
Aerosol (non-saline)	1.0000	0.6667	0.8000	3
Animal corpus	1.0000	1.0000	1.0000	3
Animal proximal gut	0.9333	0.9333	0.9333	15
Hypersaline (saline)	0.8333	0.8333	0.8333	6
Plant rhizosphere	0.9167	0.8250	0.8684	40
Plant surface	1.0000	0.2500	0.4000	4
Sediment (non-saline)	1.0000	0.8421	0.9143	38
Sediment (saline)	0.8750	0.8750	0.8750	8
Soil (non-saline)	0.9231	0.9778	0.9496	135
Subsurface (non-saline)	0.9074	0.9608	0.9333	51
Surface (non-saline)	1.0000	1.0000	1.0000	1
Water (non-saline)	0.8981	0.9151	0.9065	106
Water (saline)	0.9764	0.9841	0.9802	126
accuracy	0.9328	0.9328	0.9328	0
macro avg	0.9433	0.8510	0.8765	536
weighted avg	0.9343	0.9328	0.9308	536

NN

Classification Report

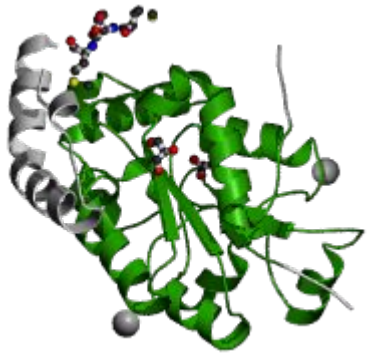
	precision	recall	f1-score	support
Aerosol (non-saline)	1.0000	0.6667	0.8000	3
Animal corpus	1.0000	1.0000	1.0000	3
Animal proximal gut	1.0000	1.0000	1.0000	15
Hypersaline (saline)	1.0000	1.0000	1.0000	6
Plant rhizosphere	0.7091	0.9750	0.8211	40
Plant surface	0.0000	0.0000	0.0000	4
Sediment (non-saline)	0.9250	0.9737	0.9487	38
Sediment (saline)	0.5714	1.0000	0.7273	8
Soil (non-saline)	0.9680	0.8963	0.9308	135
Subsurface (non-saline)	0.9074	0.9608	0.9333	51
Surface (non-saline)	0.5000	1.0000	0.6667	1
Water (non-saline)	0.9505	0.9057	0.9275	106
Water (saline)	0.9916	0.9365	0.9633	126
accuracy	0.9235	0.9235	0.9235	0
macro avg	0.8095	0.8704	0.8245	536
weighted avg	0.9295	0.9235	0.9230	536

Explainability

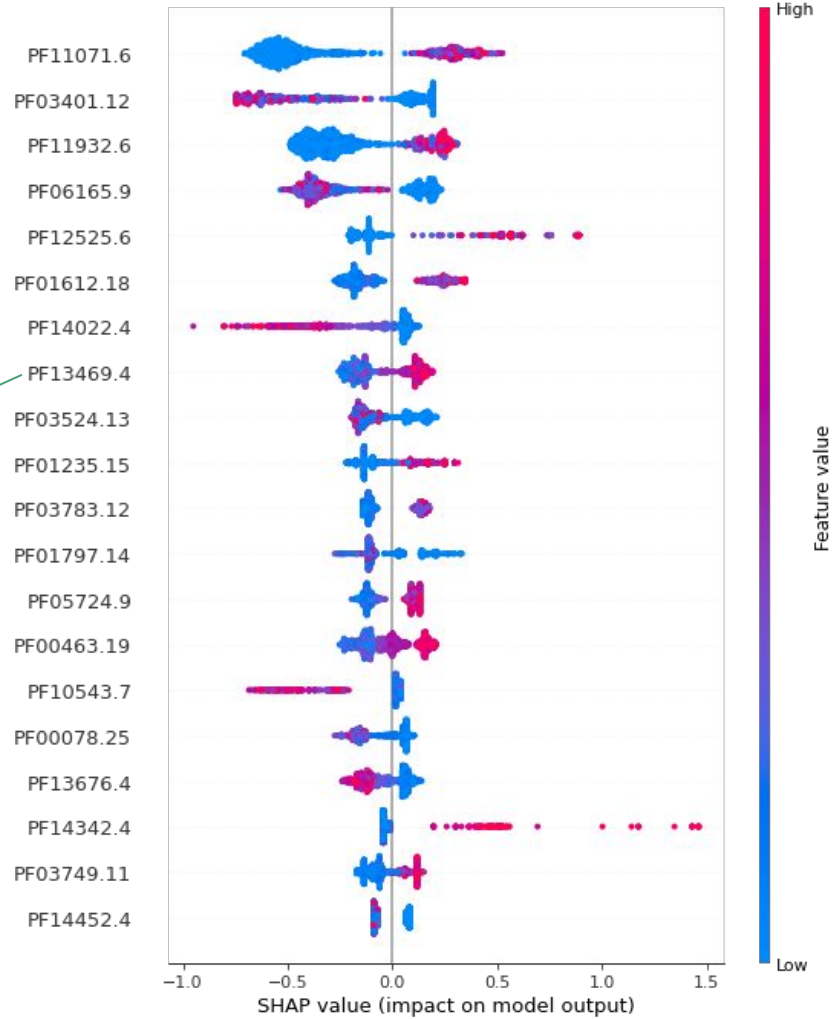
SHAP values - Water (saline)

from XGBoost Model

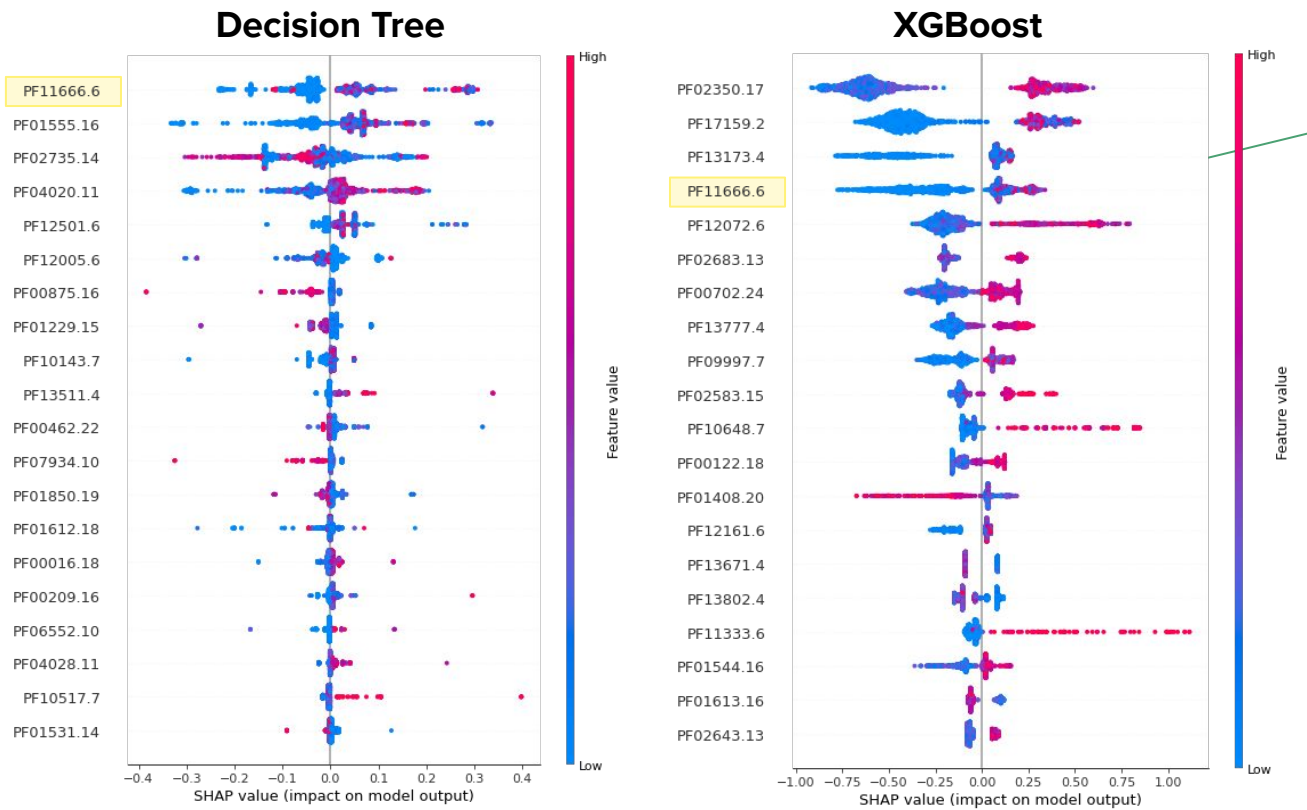
PF13469 Sulfotransferase



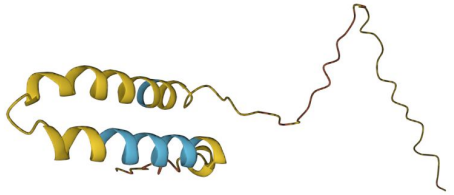
Functional activity: Transfers Sulphur to protein or glycopeptide
Purpose: Unknown



SHAP values - Subsurface (non-saline)



PF11666 is a protein of unknown function



Structure predicted by DeepMind's AlphaFold

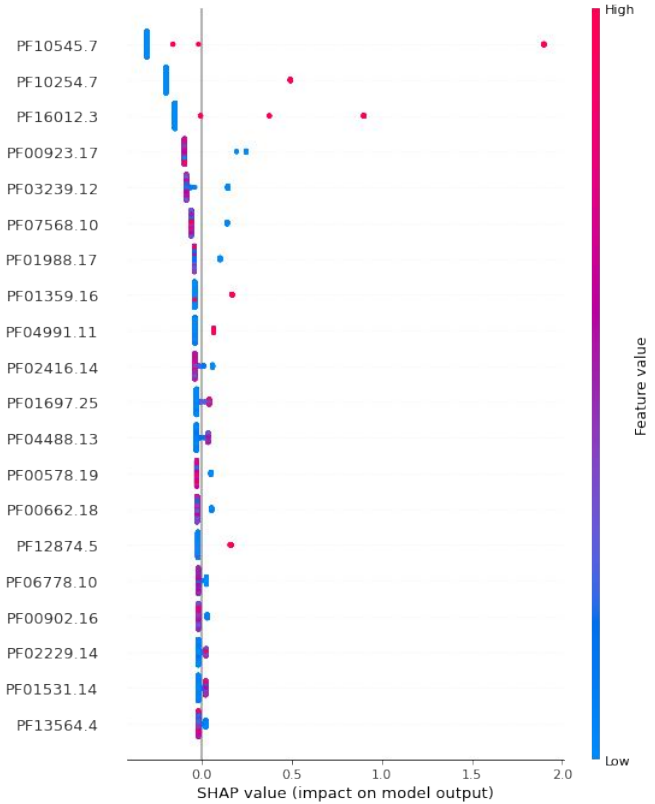
SHAP consistent with our knowledge

Animal proximal gut examples:

- PF00923 : Transaldolase (carbohydrate metabolism)
- PF02416, PF00902 : Twin-arginine translocation pathway (cellulase export)

This supports the idea that SHAP values are useful for identifying proteins of interest.

Other proteins whose function is not clear warrant further consideration.



Conclusions

Key Results: Biological Significance

- XGB SHAP values indicate which proteins have +/- association with particular environments
  - SHAP did not work out-of-the-box on the NN due to the large number of parameters (required >32GB RAM)
- Automatic classification labels can be added to many thousands of samples that do not have EMPO labels

Learnings

- Train-Test splits and stratified sampling significantly impacts model performance for high dimensional, small sample datasets.

Future work

- Must collect more data on classes with fewer samples to address class imbalance
- Figure out how to run SHAP in parallel in the cloud for NN model (promises to annotate more proteins than tree-based models)

Thank you!

Contributions:

Sophie:	XGBoost, Label encoding, Model evaluation
Edward:	Data collection, Neural Network (EMPO labels), Optuna
Delaney:	Repo organization, F1 metric, Decision Tree, Dim. reduction, Hyperparameter tuning, SHAP, `Main` notebook
Haibi:	Exploratory data analysis, Data preprocessing, Neural network (GOLD labels)

Appendix

Decision Tree Test Set Confusion Matrix

XGB outperforms DT with less incorrect predictions, but only a small difference (1 to 3 samples).