TRANSFORMERS FOR GENOMIC PREDICTION:
working with Yeast and Wheat traits

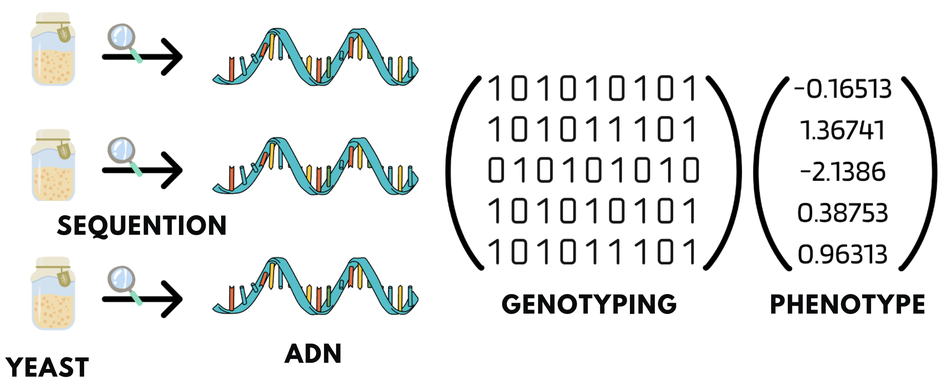
Graciana Castro[1], Romina Hoffman[1], Mateo Musitelli[2],
María Inés Fariello[2], Federico Lecumberry[1]

[1]Instituto de Ingeniería Eléctrica, [2]Instituto de Matemática y Estadística Rafael Laguardia
Facultad de Ingeniería, UDELAR - Montevideo, Uruguay
gcastro@fing.edu.uy

ABSTRACT

AI is becoming state-of-the-art across scientific fields, giving novel solutions to age-old problems. In genomic prediction, Machine Learning methods could not outperform linear regressions in a general way yet, but are becoming closer. An important feature when working with genomic data, which is non other than a long sequence of information, is to account for the linkage disequilibrium, i.e. dependencies between genome variations that do not need to be close in the genome, and variate with respect to the reference genome. To explode this feature, we evaluate Transformers, known for their great performance with long sequences. We worked with two databases: the first one composed of Yeast SNPs seeking to predict the growth of each individual in two different environments and the second one composed of Wheat SNPs seeking to predict four phenotypes. We compare the results with different linear models (BRR, BayesA, BayesB, BayesC and BayesL) typically used for genomic prediction and also with XGBoost, commonly known to have well performance in the area. We conclude that Transformers have shown to be a competitive model for genomic prediction, even tho it does not achieve the state-of-the-art yet.

DATASETS



Yeast:[3]

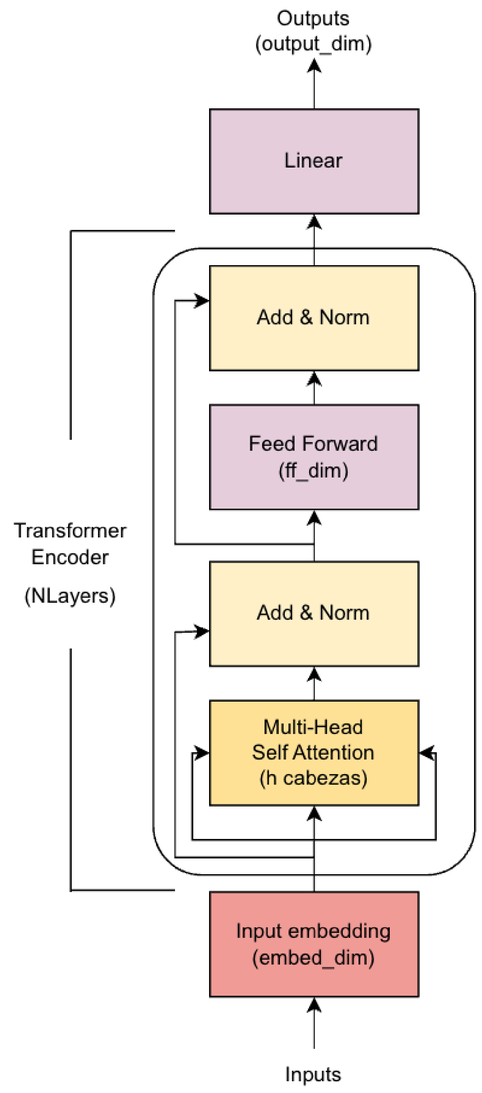
- 1008 individuals with 11623 SNPs
- Phenotype: yeast growth in different environments. Working with Lactate and Lactose.

Wheat:[4]

- 599 individuals with 1280 SNPs
- Four phenotypes to predict

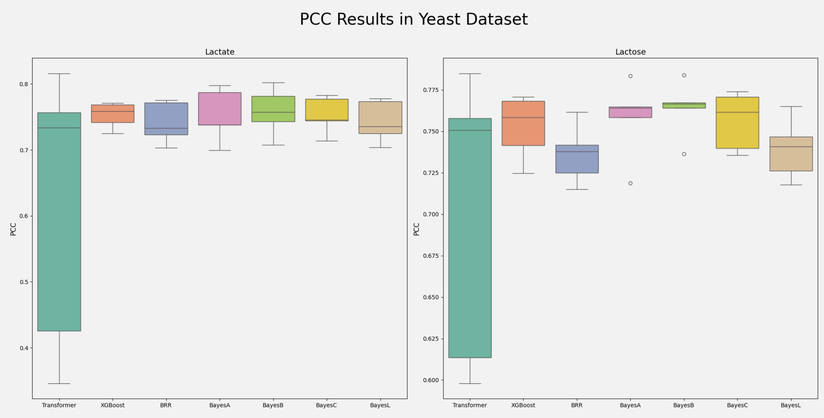
Both datasets divided into 5 folds each.

ENCODER BASED MODEL



- Based on Jubair et Al.[4]
- Input embedding: fully connected feed-forward network
- Encoder: one or two layers with the classical encoder structure.
- Linear layer: output dimension is two.
- Trained to optimize PCC.

RESULTS: IN GENOMICS, NO MODEL OUTPERFORMS THE REST

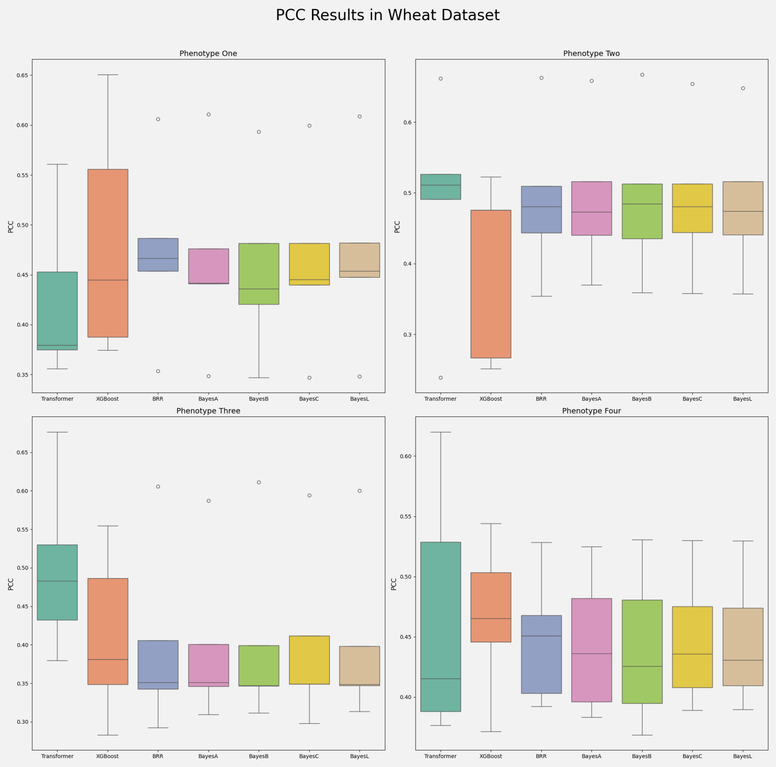


YEAST RESULTS

Transformers show competitive results for Yeast growth prediction in both environments (Lactate and Lactose) compared with linear models and XGBoost.
However, a large variance can be observed. This can result in unstable predictions, indicating that the model could lack robustness.

WHEAT RESULTS

As expected, linear models once again demonstrate robustness and provide consistent results across all phenotypes. However, when focusing on phenotypes Two and Three, Transformers exhibit superior performance, surpassing the results achieved by linear models. Despite this outcomes, it is important to note that, as highlighted in the Yeast results, Transformers are associated with high variance when applied to Phenotype Four, which may impact the reliability of their predictions in this specific context.



CONCLUSIONS AND FURTHER WORK

Since they present competitive results to those of the linear models, we can conclude that Transformers are a model worth investigating further for genomic prediction.
Next steps in our research include:
- Larger databases to achieve better generalization.
- Longer sequences, to better exploit the virtue of Transformers.
- Adding environmental information.

REFERENCES

1 Elenter, J., Etchebarne, G., Hounie, I.: DNAI: Machine learning for genome enabled prediction of complex traits in agriculture. Master's thesis (2021)
2 Gill, H.S., Halder, J., Zhang, J., Brar, N.K., Rai, T.S., Hall, C., Bernardo, A., Amand, P.S., Bai, G., Olson, E., et al.: Multi-trait multi-environment genomic prediction of agronomic traits in advanced breeding lines of winter wheat. Front. Plant Sci. 12, 709545 (2021)
3 Grinberg, N.F., Orhobor, O.I., King, R.D.: An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. Mach. Learn. 109, 251–277 (2020)
4 Jubair, S., et al.: Gptransformer: A transformer-based deep learning method for predicting fusarium related traits in barley. Front. Plant Sci. 12, 2984 (2021)

IMERL
FACULTAD DE INGENIERIA
UNIVERSIDAD DE LA REPUBLICA URUGUAY