Something old, something new, something borrowed: evaluation of different neural network architectures for genomic prediction

María Inés Fariello[1,2], Lucía Arboleya[1], Diego Belzarena[1], Leonardo de los Santos[1], Juan Elenter[1], Guillermo Etchebarne[1], Ignacio Hounie[1], Gabriel Ciappesoni[3], Elly Navajas[3], Federico Lecumberry[1,2]

Farielberry Lab

(1) Facultad de Ingeniería, Universidad de la República, Uruguay. (2) Institut Pasteur de Montevideo, Uruguay. (3) Instituto Nacional de Investigación Agropecuaria, Las Brujas, Uruguay.

fariello@fing.edu.uy

Introduction and motivation

Genome enabled prediction of complex traits aims to predict a measurable characteristic of an organism using their genetic information. We benchmarked several popular Machine Learning models: Bayesian and penalized linear regressions, kernel methods, and Decision Tree ensembles. Through exhaustive hyperparameter tuning we outperform state-of-the-art results in most datasets. We also explore different Deep Learning architectures for this task such as Convolutional Neural Network (CNN) and Graph Convolutional Network (GCN) architectures and their combination. We show that using residual connections improves performance but that in some cases FCN outperform CNNs. In the GCN trait prediction is formulated as a node regression problem on a population graph. We evaluate the transferability of these graphical models and find that the extent to which they exploit neighborhood information is limited.

A Python library is in process to be released at github.com/farielberry-lab.

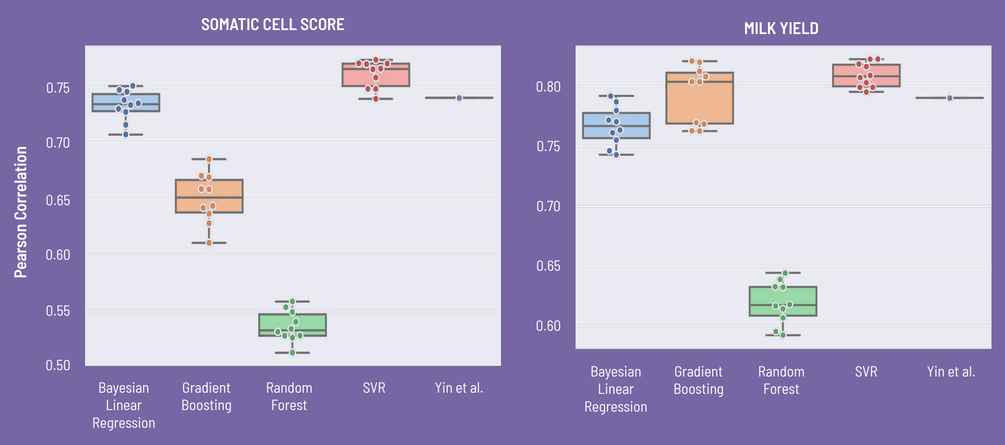
Dataset and experiments settings

- Individuals (*n*): 5024 German Holstein bulls.
- Genotypes (*p*): 42.551 SNPs after quality control filtering.
- Phenotypes: somatic cell score (SCS) and milk yield (MY).
  - SCS is governed by many small effect loci.
  - MY is determined by a few moderate effect loci and many small effect loci
- Experiments were repeated 10x using random splits.
  - Hyperparameter searches and fine tuning were done using randomized five-fold cross-validation.
  - Results can be found on https://www.comet.ml/dna-i

Something old: Evaluation of classical methods for benchmarking

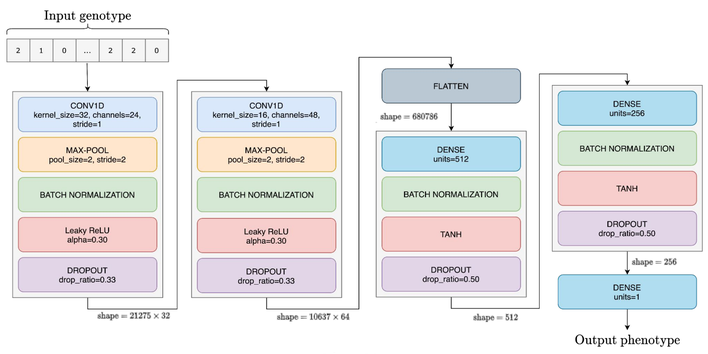
- Hetereogeneity in heritability, genetic architecture, marker and sample. dimensions among datasets enhances model performance analysis.
- Model performance showed high variance with respect to train/test splits.
- Hyperparameter tuning alone enabled surpassing the state of the art in Jersey, Yeast and Wheat datasets (not shown).
- Holstein proved to be more challenging.

Holstein dataset (n: 5000, p: 42k)

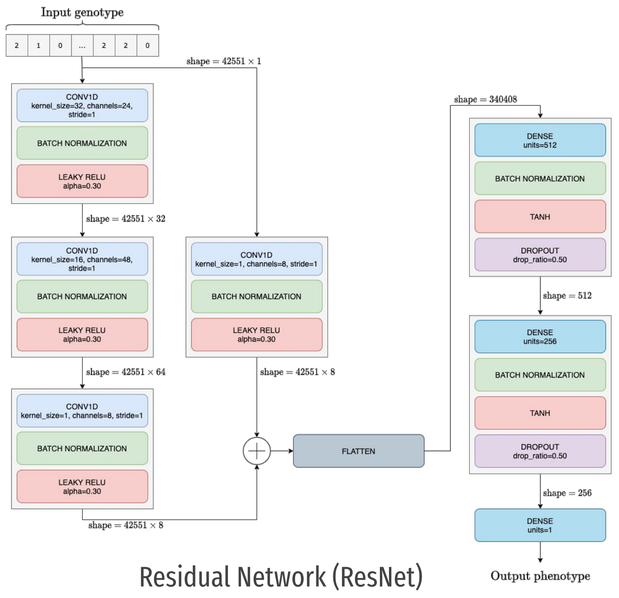


Something borrowed: Convolutional Neural Networks (CNN) + Residual CNN (ResNet)

- CNNs are a classical architecture used in image analysis.
- AlexNet-like CNN, residual CNN, with their corresponding single and multitrait variants were tested.
- The ResNet made the difference in MY prediction.

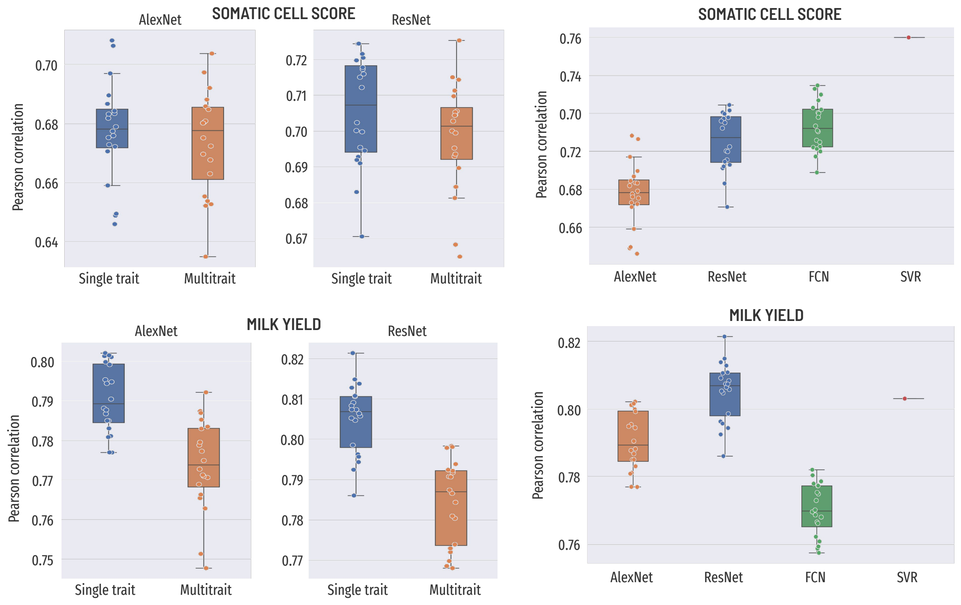


AlexNet



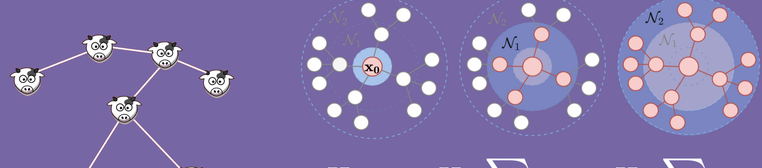
Residual Network (ResNet)

Holstein dataset



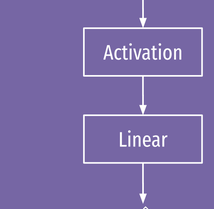
Something new: Graph Convolutional Networks (GCN)

- Build a graph with an individual's parameter in nodes and a similarity measure between nodes as edge's weights.
  - CNN output in each node
- Define a *convolution* supported in the graph for data aggregation.

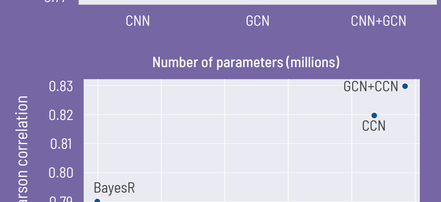


$$H_0\mathbf{x}_0 + H_1 \sum_{i\in\mathcal{N}_1} \mathbf{x}_i + H_2 \sum_{i\in\mathcal{N}_2} \mathbf{x}_i$$

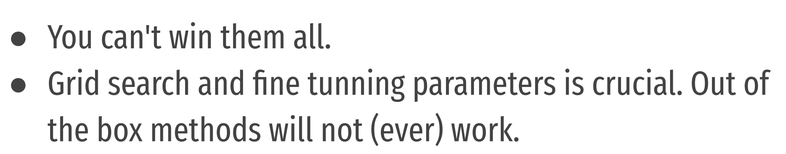
$$H_k \in \mathbb{R}^{d_{\text{in}}\times d_{\text{out}}}$$



Holstein dataset



Conclusions



- You can't win them all.
- Grid search and fine tunning parameters is crucial. Out of the box methods will not (ever) work.
- There is still room for improvement with non-linear deep learning methods and ...
- ... massive amounts of data and computational power.

Acknowledgements

We thank Francisco Peñagaricano, José Crossa, Abelardo Montesinos, Osval Montesinos, Daniel Gianola and Hugo Naya for their valuable discussion, data and proposed experiments and INIA for data access.

This work was partially funded by Universidad de la República and project ANII FSDA 1_2018_1_154364.

References

- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.
- Yin, L., Zhang, H., Zhou, X., Yuan, X., Zhao, S., Li, X., & Liu, X. (2020). KAML: improving genomic prediction accuracy of complex traits using machine learning determined parameters. Genome biology, 21(1), 1-22.
- Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems, 32(1), 4-24.

INIA URUGUAY · Institut Pasteur de Montevideo · FACULTAD DE INGENIERÍA · UNIVERSIDAD DE LA REPÚBLICA URUGUAY