Deep learning for genomic prediction and tasks learned on the way

María Inés Fariello[1,2], Lucía Arboleya[1], Diego Belzarena[1], Graciana Castro[1], Leonardo de los Santos[1], Juan Elenter[1], Guillermo Etchebarne[1], Romina Hoffman[1], Ignacio Hounie[1], Mateo Musitelli[1], Federico Lecumberry[1,2]

Farielberry Lab

(1) Facultad de Ingeniería, Universidad de la República, Uruguay. (2) Institut Pasteur de Montevideo, Uruguay

fariello@fing.edu.uy

Introduction and motivation

- Genome enabled prediction of complex traits aims to predict a measurable characteristic of an organism using its genomic information.
- Deep Learning architectures: CNNs, GCNs and CNN+GCN.
- Increasing number of SNPs, increases number of parameters and memory needed very fast.
- Uses of AEs and VAEs in population genomics.
- Python library is in process to be released at github.com/farielberry-lab

Datasets and experiments settings

German Holstein bulls:

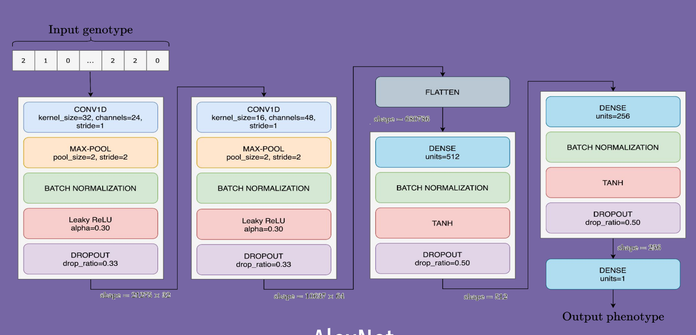
- Individuals (*n*): 5024
- Genotypes (*p*): 42.551 SNPs after quality control filtering.
- Phenotypes: somatic cell score (SCS) and milk yield (MY).
  - SCS is governed by many small effect loci.
  - MY is determined by a few moderate effect loci and many small effect loci
- Experiments were repeated 10x using random splits.\
- Hyperparameter searches and fine tuning: five-fold X-val.

Jersey bulls:

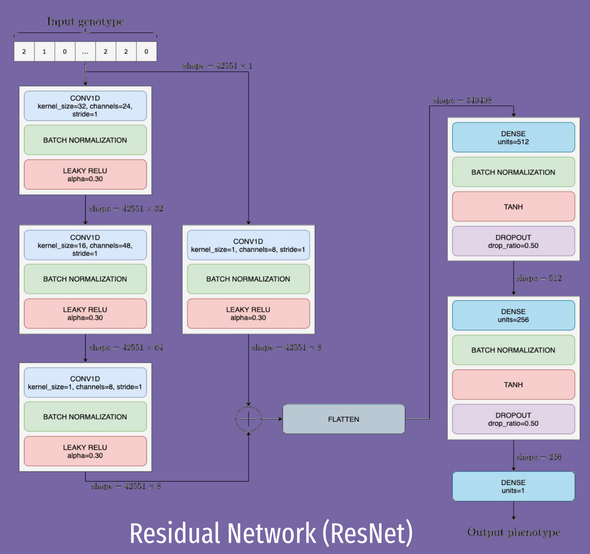
- Individuals (*n*): 1569 (from 2008 to 2018)
- Genotypes (*p*): 95.434 SNPs after quality control filtering.
- Phenotype: Sire Conception Rate (SCR)
  - Expected difference between the conception rate (CR) of a bull compared with the mean of the rest of the population in a certain year. CR: amount of successful inseminations as a fraction of the total inseminations attempted

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.

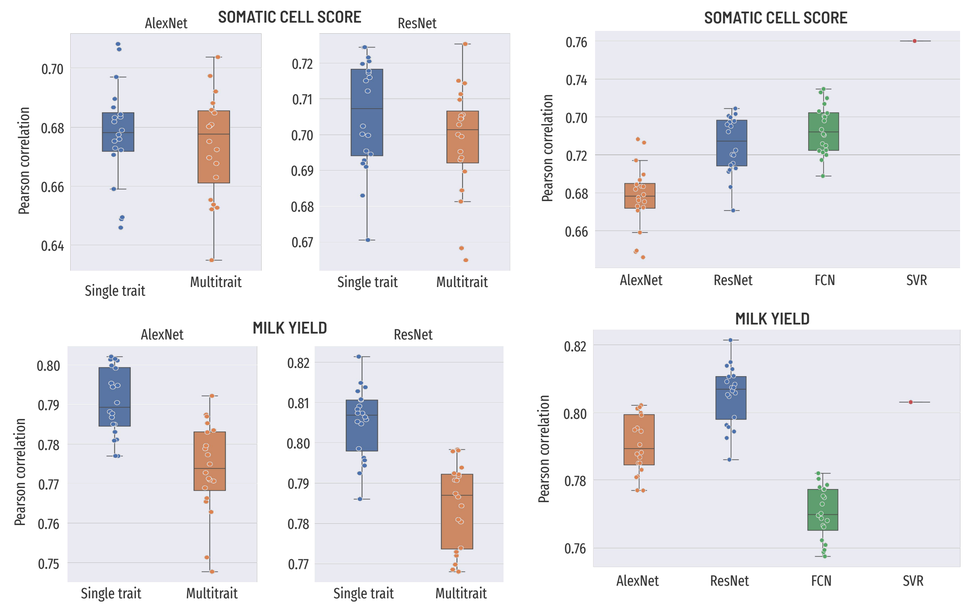


AlexNet

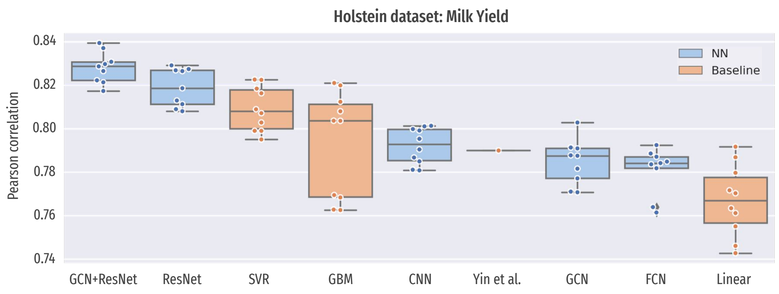


Residual Network (ResNet)

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



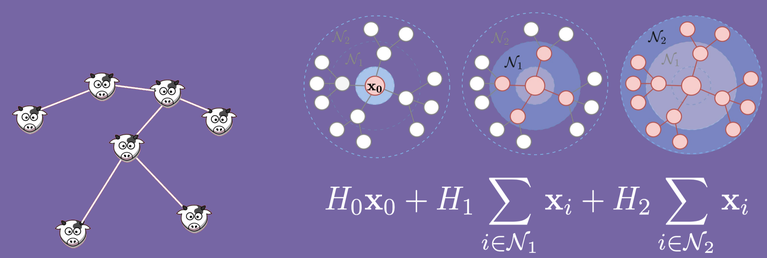
Holstein overall results



- Grid search and fine tuning parameters is crucial. Out of the box methods will not (ever) work.
- Adding information of relatives improves the predictions.
- Number of parameters grows with the model complexity.
- More results on www.comet.ml/dna-i

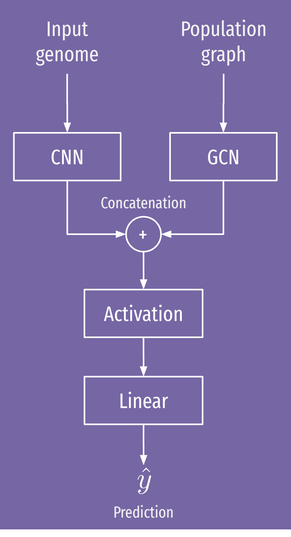
Graph Convolutional Networks (GCN)

- 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
- *Convolution* supported in the graph for data aggregation.

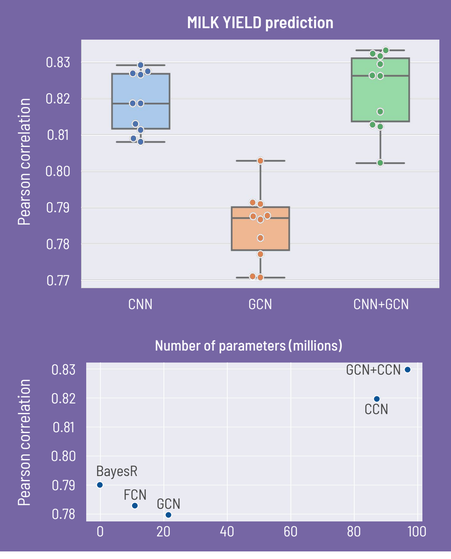


$$H_0\mathbf{x}_0 + H_1 \sum_{i\in\mathcal{N}_1} \mathbf{x}_i + H_2 \sum_{i\in\mathcal{N}_2} \mathbf{x}_i$$

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



Holstein dataset



(Variational) Autoenconders (VAE and AE)

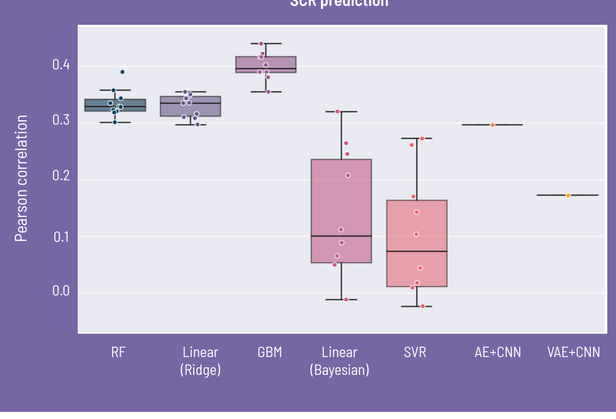
AEs and VAEs are capable of learning dense representations of the input data. They can be used for:

- Dimensionality reduction for visualization or memory reduction through its latent representations.
- Imputation: train the data without missing data, then use the autoencoder on data with missingness.

Algorithm for genomic prediction:

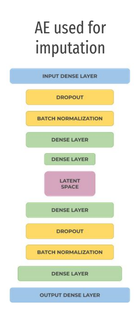
1. Split genotypes by chromosome
2. Fit a dimensionality reduction network (AE, VAE) for each one. The latent dimension is a fixed proportion of the input dimension.
3. Concatenate the latent representations to build the low-dimensional genotype.
4. Train a neural network (eg. CNN) on the low-dimensional genotype for prediction.

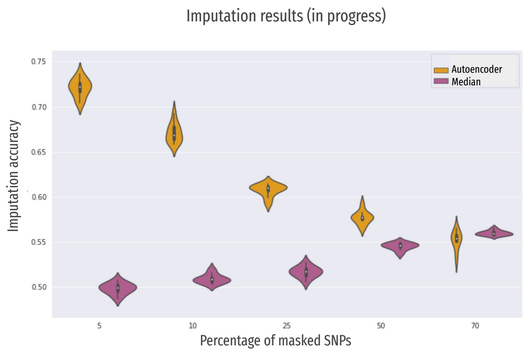
Jersey dataset



Imputation

The AE can be trained including missing data in the data set.





Conclusions and future work

- Same network architectures can be useful for different tasks (imputation, dimensionality reduction).
- Understand how missing data affects the latent representations.
- Good latent representations could overcame imputation.
- As datasets (features and individuals) grow, dimensionality reduction could be as important as regularizations.

Acknowledgements

Francisco Peñagaricano, José Crossa, Abelardo and Osval Montesinos, Daniel Gianola, Hugo Naya, Elly Navajas and Gabriel Ciappesoni 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 and IA_1_2022_1_173411

Some references

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks.

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..

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks.

Battey, C. J., G. C. Coffing, and A. D. Kern (2021). Visualizing population structure with variational autoencoders.

Islam,T.,C. H.Kim,H.Iwata,H.Shimono,A.Kimura,H.Zaw,C.Raghavan,H.Leung,R.K.Singh (2021). A Deep Learning Method to Impute Missing Values and Compress Genomewide Polymorphism Data in Rice.