RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

GenIALearn

EXPLORATORY PROJECT

2021-2023

Application of machine learning and deep learning to improve genomic selection in animals

Coordinators
Eric Barrey, JRU GABI
eric.barrey@inrae.fr
Didier Boichard, JRU GABI
didier.boichard@inrae.fr

Keywords
Genomics,
Gene interactions,
Statistical learning, Automatic learning,
Deep learning

Participating INRAE Units
GABI JRU
MIA – Paris Saclay

External partners
UEVE, Paris-Saclay University
IBISC

© kjpargeter, CS - freepik

Context, key challenges and goals

The development of genomic selection – and of other 'omics' fields of research – has made it possible to characterize animals using many thousands of measurements. These very large datasets are integrated into models that seek to predict animal production traits with the highest possible degree of accuracy. The most used models in genomic prediction (additive genetic models such as GBLUP[1]) are very efficient in predicting the genetic value of animals based on a small number of genetically correlated traits. However, this type of model does not allow the integration of very large numbers of heterogeneous measurements, nor can it predict multiple output traits without knowing their genetic correlations. What is more, such models struggle to accommodate the many non-linear interactions that occur between different regions of the genome and between environmental factors.

The GenIALearn project set out to evaluate the performance of statistical and deep learning methods in the joint prediction of multiple complex traits in dairy cattle through the integration of massive genotyping data. Two complementary approaches – ensemble methods (machine learning) and neural network methods (deep learning) – were implemented separately and in a hybrid model to predict 33 phenotypic traits (associated with production, morphology, fertility, lameness, etc.) based on a single genotype. The various methods were then compared, with the GBLUP model as the reference.

The GenIALearn teams were able to use a very large database of dairy cattle genotypes and phenotypes (based on 113 599 Holstein females), built and managed by the GABI Joint Research Unit, which led the project.

[1] Genomic Best Linear Unbiased Prediction
[2] Graphics processing unit (server equipped processors)

Results

Access to appropriate calculation technology and the assessment of machine learning methods

Métaprogramme
DIGIT-BIO

digitbio@inrae.fr
www.inrae.fr/digitbio/

The project allowed a Graphics Processing Unit to be acquired, along with an archiving server, both of which were integrated with the INRAE Data Center's CoLab.IA in Toulouse, a digital platform devoted entirely to AI. During the project, Masters 2 internships were used to test different methods. Some turned out to be unsuited to the context of application, while others showed real potential (although they will require improvements to be fully effective for such applications).

In particular, the project enabled the following to be established:

- Of the 33 phenotypic traits studied, around a dozen were better predicted by the AI models tested, which were also faster to run than 33 single trait GBLUP models. The GBLUP reference model nevertheless remained sufficient for a majority of traits.
- The Deep Learning (neural network) models appeared, overall, to be more adaptable and performed better than ensemble models.
- Of the models based on neural networks, generative models of the WGAN-GP type produced highly realistic artificial genotypes in our tests (PCA analysis, distance metrics). These show promise for the improvement of learning in predictive models for genome selection and are worthy of further exploration.

Perspectives for the future

Research theme explored further in two theses

Project partners continue to collaborate on the development of shared resources (large datasets, CoLab.IA digital platform). The project has also resulted in the funding of two theses on the application of AI to genomic selection within the GABI unit:

- **Sihan Xie (deepSelectGene, 2024-26):** thesis funded by Metaprogramme DIGIT-BIO, aiming to develop machine learning methods for species where genotype-phenotype data are available for only a few thousand animals.
- **Fatima Shokor (2022-2025):** thesis funded by APIGENE, aiming to develop AI models for the prediction of phenotypes produced by bovine cross-breeding.

Advanced computing infrastructure for AI

The Colab.IA platform for AI applications is a long-term experimental engineering project. It is maintained and developed through the active collaboration of the GenIALearn project members with the EPIA Epidemiology unit in Clermont-Ferrand. Additional investment in 2024 made it possible to boost its GPU computing and storage capacities, enabling the exploration of more complex models based on larger learning datasets.

The GenIALearn interdisciplinary project has encouraged working partnerships between different unit teams, between INRAE departments, and with external partners, in particular the IBISC lab (a joint venture between INRAE, the Université d'Évry Val-d'Essonne and the Université Paris-Saclay). This dynamic collaboration continues through ongoing development of the platform and doctoral supervision. Additionally, the units involved in GenIALearn contributed to the creation of the DATA AI cluster at the l'Université Paris-Saclay as part of the France 2030 program run by the ANR (French national research agency), and are currently stakeholders in the cluster.

Publications

- Xie, S., Tribout, T., Boichard, D., Hanczar, B., Chiquet, J., & Barrey, E. (2025). Deep Generative Models for Discrete Genotype Simulation. BioRxiv, 2025.08.08.669289. https://doi.org/10.1101/2025.08.08.669289
- Shokor, F., Croiseau, P., Gangloff, H., Saintilan, R., Tribout, T., Mary-Huard, T., & Cuyabano, B. C. D. (2024) Deep Learning and GBLUP Integration: An Approach that Identifies Nonlinear Genetic Relationships Between Traits. bioRxiv https://doi.org/10.1101/2024.03.23.585208
- Shokor, F., Croiseau, P., Gangloff, H., Saintilan, R., Tribout, T., Mary-Huard, T., & Cuyabano, B. C. D. (2025). Deep learning and genomic best linear unbiased prediction integration: An approach to identify potential nonlinear genetic relationships between traits. Journal of Dairy Science, 108(6), 6174–6189. https://doi.org/10.3168/jds.2024-26057

See the full list of publications and papers resulting from the project on https://digitbio.hub.inrae.fr