Predicting nonlinear genetic relationships between traits in multi-trait evaluations by using a GBLUP-assisted Deep Learning model

Fatima Shokor [1,2]*, Pascal Croiseau [2], Hugo Gangloff [3], Romain Saintilan [1,2], Thierry Tribout [2], Tristan Mary-Huard[3,4], and Beatriz C.D. Cuyabano[2]

[1]Eliance, 149 Rue de Bercy, 75012 Paris, France

[2]Université Paris Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France

[3]Université Paris Saclay, INRAE, AgroParisTech, UMR MIA Paris-Saclay, 91120 Palaiseau, France

[4]Université Paris Saclay, INRAE, CNRS, AgroParisTech, GQE – Le Moulon, 91190 Gif-sur-Yvette, France

*Corresponding author

E-mail addresses:

FS: fatima.shokor@eliance.fr

Abstract

Background

Genomic prediction aims to predict the breeding values of multiple complex traits assumed to be normally distributed, thus imposing linear genetic correlations between traits. However, these statistical methods are unable to model nonlinear genetic relationships between traits, if existent, potentially leading to a decrease in prediction accuracy. Deep learning (DL) is a promising methodology for predicting multiple complex traits, in scenarios where nonlinear genetic relationships are present, due to its capacity to capture complex and nonlinear patterns in large data. We proposed a novel pure DL model, designed to obtain predicted genetic values (PGV) while accounting for nonlinear genetic relationships between traits, and extended this model to a hybrid DLGBLUP model which uses the output of the traditional GBLUP, and enhances its PGV by using DL. Using simulated data, we compared the accuracy of the PGV obtained with the proposed pure DL model, the hybrid DLGBLUP model, and the traditional GBLUP model – the latter being our baseline reference.

Results

We found that both DL and DLGBLUP models either outperformed GBLUP, or presented equally accurate PGV, with a particular greater accuracy for traits presenting a strongly characterized nonlinear genetic relationship. DLGBLUP presented the highest prediction accuracy and smallest mean squared error of the PGV for all traits. Additionally, we evolved a base population over seven generations and compared the genetic progress when selecting individuals based on the additive PGV obtained by either DL, DLGBLUP or GBLUP. For all traits with a nonlinear genetic relationship, after the fourth generation, the observed genetic gain when selection was based on the additive PGV

from GBLUP was always inferior to the observed when selection was based on either DL or DLGBLUP.

Conclusions

The integration of DL into genomic prediction has potential to bring significant advancements in the field. By identifying nonlinear genetic relationships, our DL and DLGBLUP models improved prediction accuracy. It offers an insight to genetic relationship and its evolution over generations, with potential to improve selection strategies in commercial livestock breeding programs. Moreover, DLGBLUP shows that DL can be used as a complement to statistical methods, by enhancing their performance.

Background

Genomic prediction (GP) [1] in genetic evaluations uses DNA marker information, most commonly single nucleotide polymorphisms (SNPs) data, to predict the genetic merit of complex traits, referred to as genetic value (GV). When focusing on additive genetic values, it is commonly termed as breeding values (BVs) in both livestock and plant studies. Traditional statistical methods applied to GP rely on linear mixed models [2], and typically use either the SNP or the genomic best linear unbiased prediction (SNP-BLUP and GBLUP respectively), or Bayesian approaches, with their various prior assumptions and alphabets [1, 3] to obtain the predicted genetic values (PGV), by assuming normally distributed effects. Due to the assumption of normality of the data, when extended to a multi trait (MT) scenario, the relationship between traits is modeled as linear. These statistical methods, combined with the current availability of genomic information are very powerful, and have undoubtedly revolutionized genetic evaluations, but since they fully rely on linear approximations, a higher prediction accuracy is prevented, despite the efforts of improvement by providing highly

informative data, such as denser SNP-chips, or using a single-step strategy that combines the pedigree to the genomic data in a single relationship matrix [4].

The restriction to linearity, in addition to the continuously increasing amount of recorded and genotyped animals, which posed computational constraints for genetic evaluations with the classical statistical models, were among the reasons that pushed geneticists to explore the use of deep learning (DL) [5] for genetic evaluations and genomic predictions [6]. DL approaches use artificial neural networks, which have the capacity to learn complex patterns and features from very large datasets, in order to map input to output data. Different DL algorithms, also called architectures, can be considered depending on the prediction task such as classification and regression. Among the three most common DL algorithms we have (1) Multi-layer perceptron (MLP), which is particularly effective for tasks that involve non-sequential and non-spatial data; (2) Convolutional neural network (CNN) designed for visual data, such as images and videos; and (3) Transformer [7], the backbone of most large language models (LLM), used in natural language processing (NLP) tasks, suited for any data that can be represented as a sequence, including text, time-series data, and even genomic sequences. DL has been successfully applied in many domains, such as computer vision, speech recognition, text and image generation, and biomedicine [8, 9, 10, 11, 12].

Most studies that use DL for GP have focused on the inclusion of the non-additive genetic effects (epistasis and dominance) to the linear model that uses genotypes to describe phenotypes, by modeling their non-linearity using DL. There was a focus on the use of CNN since they are designed to exploit the local characteristics of data, which can be a manner to describe linkage disequilibrium (LD) between the SNPs in genomic data. Still, some studies have shown that the MLP models gave better results [6]. Although in some studies DL did outperform the conventional models for very specific datasets and traits, there was no clear evidence of an overall superiority of DL over the statistical methods in performing GP [13, 14]. More recently, Lee et al. [15] proposed deepGBLUP,

a combination of DL and GBLUP into a single model. Given input SNP data, the DL networks of deepGBLUP extract the effects of adjacent SNPs using locally-connected layers to estimate a GV through fully-connected layers. In parallel, a GBLUP model accounting for additive, dominance and epistasis effects (through their respective genetic relationship matrices) is fitted. The final estimated GV is the sum of all previous estimated GV (DL, additive, dominance and epistasis). Applied to a real dataset of Korean native cattle, and to simulated data, the results of the proposed deepGBLUP model outperformed those of the traditional GBLUP and Bayesian methods in different single trait prediction scenarios.

In livestock production, beyond predicting GV for individual traits, breeders aim to jointly improve multiple traits of commercial interest, in order to achieve genetic progress for these traits altogether. When working with multiple traits, genetic relationship between traits [16] must be considered, since selection for one trait will affect the other correlated traits [17]. Genetic relationships are thus a relevant factor to account for, on the different stages of a genetic evaluation, from the estimation of variance components to the prediction of GV, in order to optimally improve multiple traits of interest altogether. It has been shown, for example, that it is more advantageous to perform a multitrait model for genetic evaluation, that considers the genetic relationships between traits to improve their predictive ability [18].

Until now, due to the Gaussian nature of the models employed, the genetic relationship between traits has been always considered as linear. Therefore, if two traits present a non-linear genetic relationship, statistical methods will fail to identify this relationship, thus limiting prediction accuracy of the PGV associated to the traits of interest. Such limitation cannot be overcome, unless the assumption of normality of the data is dissociated from the model, in order to allow the possibility of nonlinear relationships. Although the possibility of nonlinear relationships creates a difficulty for the implementation of the traditional statistical models, it opens an opportunity door for the

implementation of DL methods. Few studies explored non-linear genetic relationships between traits [19, 20], and even so, such studies focused merely on how to identify these non-linear genetic relationships, rather than on how to account for them in GP.

This study aimed to use DL to model non-linear genetic relationships between traits and to use these identified nonlinear relationships to predict the GV of multiple traits simultaneously. Using simulated data, we explored (1) how the presence of nonlinear relationships between traits affect the performance of GBLUP with respect to the accuracy of the PGV, (2) how to use DL to model these nonlinear relationships and consider them for selecting individuals, and (3) how nonlinear relationships affect the genetic progress over generations when ignored (GBLUP) or when they are taken into account (DL methods). We proposed a pure and a hybrid DL model. The pure DL model consisted of two steps, a first that uses the SNP data as input to predict the GV of multiple traits accounting only for the genomic effects, and a second that re-predicts a new GV from the initial predicted GV while allowing the flexibility to capture potential nonlinear genetic relationships between traits. The hybrid model, called DLGBLUP, combines both DL and GBLUP, thus benefiting from their strengths while minimizing their pitfalls, and also consists of two steps. A first that predicts the GV using a multi-trait GBLUP and the genomic data, by learning the output of GBLUP using DL, and a second that performs exactly the same second step of the original pure DL model.

Methods

Simulated data

The complete simulated genomic data consisted of 10,000 SNPs distributed across 29 chromosomes, with an average LD pattern resembling that of a cattle population. From the 10,000 simulated SNPs, 512 were assigned as quantitative trait loci (QTL) to be shared between all traits. The non-centered

SNPs were coded as 0, 1, and 2, referring to homozygous for reference alleles, heterozygous, and homozygous for alternate alleles, respectively.

A quantitative reference trait was simulated using the 512 QTL for a total of 25,000 individuals, following the model: $\mathbf{y_{ref}} = \mathbf{M}\,\boldsymbol{\alpha}_{\mathbf{ref}} + \mathbf{e_{ref}} = \mathbf{g_{ref}} + \mathbf{e_{ref}}$, where $\mathbf{M}$ is a matrix of which element $M_{i,j}$ corresponds to the centered genotype of individual i at QTL j, $\boldsymbol{\alpha}_{\mathbf{ref}} = [\alpha_{ref,1} \ldots . \alpha_{ref,q}]$ is the vector of the q=512 additive QTL effects, such that $\alpha_{ref,j} \sim N(0, \sigma_{\alpha_{ref}})$ i.i.d. for every $j = 1, \ldots, q$, $\mathbf{g_{ref}} = \mathbf{M}\boldsymbol{\alpha}_{\mathbf{ref}}$ is the vector of the true genetic values (TGV) and $\mathbf{e_{ref}} \sim N(0, \sigma_{e_{ref}}\mathbf{I})$ is the vector of random errors. The genotypes and the reference trait were simulated using the GenEval R package [19].

Five dependent traits, were simulated conditional to the TGV of the reference trait as $\mathbf{y_t} = \mathbf{g_t} + \mathbf{e}_t$, for every $t = 1, \ldots, 5$, such that $\mathbf{g_t} = f_t(\mathbf{g_{ref}}) + \mathbf{g_{t_2}}$, where f_t is the function describing the (potentially) nonlinear relationship between the TGV of two traits. $\mathbf{g_{t_2}}$ is the vector of the genetic value specific to each dependent trait, simulated as $\mathbf{g_{t_2}} = \mathbf{M}_t\boldsymbol{\alpha}_t$, where $\mathbf{M_t}$ is the genotype matrix of specific QTLs different from the common QTLs and different between the dependent traits; $\boldsymbol{\alpha}_\mathbf{t} \sim N(0, \sigma_{\alpha_t} I)$ is the vector of the corresponding additive QTL effects. $\mathbf{g_t}$ was normalized such as $\sigma^2_{g_t} = \sigma^2_{y_t} \times h^2$, which h^2 is the heritability. The error vector $\mathbf{e_t} \sim N(0, \sigma_{e_t}\mathbf{I})$ with $\sigma_{e_t} = \sigma^2_{y_t} - \sigma^2_{g_t}$. We fixed a $\sigma^2_y = 20$ for all traits. We choose a default h^2 of 0.3 and 50 specifc QTLs. Then we tested different levels of h^2 for the traits, from high to low (0.6, 0.15, 0.05), with the 50 specific QTLs, and different numbers of specific QTLs (0, 10, 250), that defined the level of genetic relationships between traits, with h^2 of 0.3. For each of the dependent traits, the following different relationships were considered: (1) linear, (2) quadratic, (3) sinusoidal, (4) logistic and (5) exponential. While different levels of h^2 and number of specific QTL were tested, each set of simulated reference and dependent traits had the same h^2 and number of specific QTLs.

To train and evaluate the DL models, the dataset was split into three sets: training (80%), validation (10%), and test set (10%), being the test set that for which PGV are to be obtained in the absence of phenotypic records, while both the training and validation sets are those for which individuals have genotypes and phenotypic records. Different from statistical models, DL requires an internal validation set with the complete information to fine-tune the model parameters. For the GBLUP model, the training and validation sets were combined to fit the model, and the same test set (10% of individuals) was kept for evaluation. To assess the repeatability of our findings, we simulated 20 replicas of the complete data set under each combination of different levels of h^2 and number of QTL.

Multi-Trait Genomic Prediction:

We used three models to perform GP: a baseline GBLUP, a pure DL model, and a hybrid DLGBLUP model that combines the two previous ones.

Genomic Best Linear Unbiased Prediction (GBLUP)

The GBLUP model is one of the most popular statistical methods used to predict the GV of genotyped individuals using their genomic relationship matrix. The model considered for multi-trait GBLUP is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{e} \quad (1)$$

where $\mathbf{Y} = [\mathbf{Y}_1 \dots \mathbf{Y}_k]$ is a matrix of the phenotype vectors for all trait included in the model; $\mathbf{X}$ is a design matrix for the fixed effects, $\boldsymbol{\beta} = [\boldsymbol{\beta}_1 \dots \boldsymbol{\beta}_k]$, such that $\boldsymbol{\beta}_k$ is the vector of fixed effects for each of the $k = 1, \dots, K$ traits; $\mathbf{Z}$ is an incidence matrix linking the phenotypic records to the breeding values $\mathbf{g} = [\mathbf{g}_1 \dots \mathbf{g}_k]$, such that $\mathbf{g}_k$ is the vector of breeding values for each of the $k = 1, \dots, K$ traits, with distribution, $\mathrm{vec}(\mathbf{g}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_g \otimes \mathbf{G})$ in which $\boldsymbol{\Sigma}_g$ is the additive genetic (co)variance matrix of the traits, vec($\mathbf{g}$) is the vector version of the matrix $\mathbf{g}$, $\mathbf{G}$ is the genomic relationship matrix (GRM), and

$\otimes$ represents their Kronecker product; and $\mathbf{e} = [\mathbf{e}_1 \ldots \mathbf{e}_k]$ is the matrix of normally distributed random errors with $\mathrm{vec}(\mathbf{e}) \sim \mathrm{N}(\mathbf{0}, \mathbf{R})$, in which $\mathbf{R}$ is the error (co)variance matrix between the traits. We used the R (4.1.1) package BGLR package (1.0.8) [22] to perform the multi-trait GBLUP, here we did not consider a fixed effect so $\boldsymbol{\beta}$ is just the vector of the intercepts.

Deep Learning model

The MLP [23] is a feedforward network that maps an input to an output using learned parameters. It consists of several layers, including hidden layers with nonlinear activation functions, each layer having multiple nodes. The model takes the input layer and passes it into a network of multiple connected layers (hidden layers) up to the output. In a general execution, each node is characterized by the following equation:

$$y_j = f(\Sigma_{i=1}^{I} w_{j,i} x_i + b_j) \quad (2)$$

where y_j is the scalar output of node j, b_j is the bias, x_i represents the output of node $i \in \{1, \ldots, I\}$ of the previous layer, $w_{j,i}$ symbolizes the weight assigned to this x_i, and f is a nonlinear activation function. During training, the optimization process adjusts the weights and biases of the MLP by minimizing a loss function between the predicted and true outputs of the training set. The Mean Squared Error (MSE) and Mean Absolute Error (MAE) losses are commonly used for regression, while cross entropy loss is used for classification. The backpropagation algorithm is used to efficiently compute gradients, and to understand how changes in model parameters can affect the output error. Stochastic gradient descent is used to update these parameters in the direction that minimizes the loss.

The proposed DL model, illustrated in Figure 1, comprises three distinct steps (also called modules hereafter): SNPs2Trait, Trait2Trait and MeanTrait2Trait. The first step, SNPs2Trait, takes the SNP data as input, and then extracts the important information in the hidden layer. The prediction of each trait was made separately in the output layer by processing all information from the hidden

layer related to each trait. However, this separation does not allow the model to capture the relationships between traits. This module will output a first predicted GV ($PGV_{_1}$) for multiple traits simultaneously, by capturing the relationship between the genomic data and the phenotype, which represents both the additive and some non-additive effects. Then the module Trait2Trait uses the $PGV_{_1}$ of all traits as input to capture the relationship between all traits. This provides a new predicted GV ($PGV_{_2}$) for all traits simultaneously, that will be considered as the final prediction of the model. Lastly, in parallel to Trait2Trait, the module MeanTrait2Trait uses PGV_{ref_1} to output predictions of the dependent traits which we call $PGV_{_3}$. As such, MeanTrait2Trait can capture the mean relationship between reference trait and the other dependent traits but not the dispersion around this mean. This last module is used for illustration purpose only.

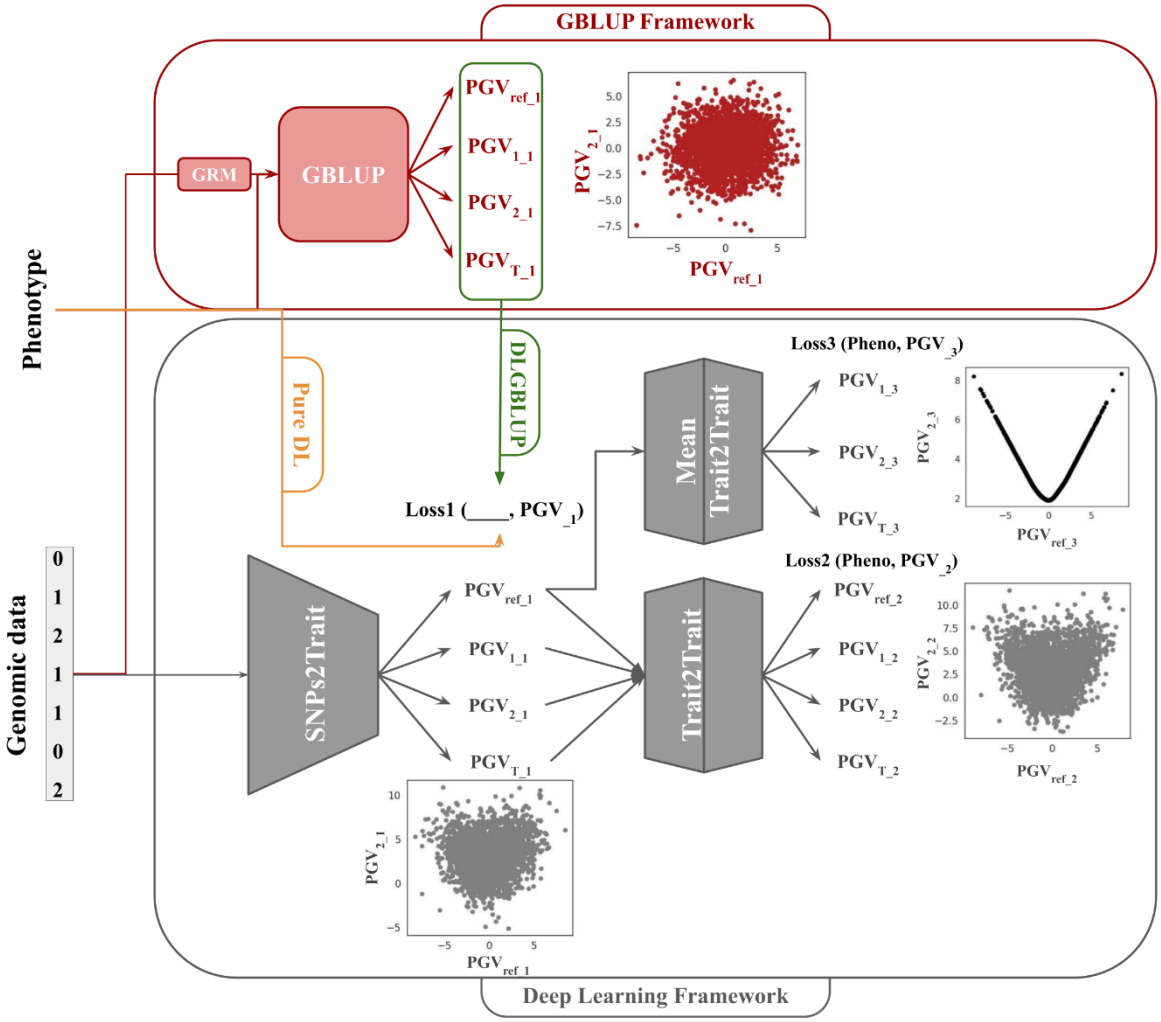


Figure 1: GBLUP vs. Deep Learning Genomic Prediction Frameworks. The GBLUP in red and DL in gray; with the second separated as pure DL in yellow and DLGBLUP in green.

All modules consist of two linear layers (with 400 and 256 neurons) and use the LeakyReLU activation functions with a fixed negative slope of 0.1. The model was trained for 100 maximum epochs, with an early stopping after 10 epochs without improvement of the loss in the validation, using a batch size of 200. As loss function, we used Huber loss [24], a combination of the MSE and the MAE, described as:

$$L(\mathrm{z}) = \begin{cases} \frac{1}{2}(\mathrm{z})^2 & \text{for } |\mathrm{z}| \leq \delta, \\ \delta|\mathrm{z}| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases} \quad (3)$$

where z is the error and δ is a threshold parameter that dictates the transition between the two losses. We used Adam optimizer with a learning rate of 10^{-4}. The choice of optimizer, learning rate, activation function and number of neurons in hidden layers was made after a preliminary grid search. Two training approaches were used: (1) training the three modules as one model so the loss function was the sum of loss from each module, i.e.,

$$\mathrm{L}oss = \sum_{\mathrm{t}} \mathrm{L}(\mathrm{y_t} - \mathrm{PGV_{t_1}}) + \mathrm{L}(\mathrm{y_t} - \mathrm{PGV_{t_2}}) + \mathrm{L}(\mathrm{y_t} - \mathrm{PGV_{t_3}}), \quad (4)$$

where y is the true phenotype value for multiple traits; (2) training modules sequentially in terms of gradient flow. The first module was trained initially, and once this was completed, its parameters were fixed by stopping the gradient propagation from a module to another. The trainings of Trait2Trait and of MeanTrait2Trait were independent. The model was implemented using pytorch (1.10.2) in python and trained on a single GPU with 48 GB memory (NVIDIA A40).

DLGBLUP model

We proposed a hybrid model called DLGBLUP, where the first module is trained on the GBLUP output instead of y. The motivation for this replacement was to guide the DL training: using the

GBLUP output may stabilize the optimization in DL when the 3 modules are jointly optimized. Indeed, this may help the convergence of SNP2trait toward a solution close to the GBLUP one that is known to performs well in predicting the additive GV. Note that in the case of a sequential training SNP2Trait will exactly reproduce the GBLUP predictions (on the training set), up to convergence tolerance. The Trait2Trait and MeanTrait2Trait are left unchanged.

Inclusion of Non-QTL SNPs

To develop the DL models properly, we first used uniquely the true QTL to minimize the errors arising from the data. While a first comparison of the accuracy of the PGV from GBLUP to our proposed DL and DLGBLUP models was performed using this approach, we did evaluate the effect of having SNPs that were not QTL in the input genomic data on the prediction accuracy. Therefore, for this evaluation, our baseline prediction data consisted of the QTL only, and then we gradually added non-causative SNPs to the genomic data used for prediction, so that the whole genomic data consisted of a total of non-QTL percentage of 10, 25, 50, 75, and finally a final scenario consisted of using only non-causative SNPs, excluding all the QTL. A comparison of the results from a model excluding non-causative mutations, but including only half of the QTL was also included in this study.

Genomic Selection

After performing the multi-trait genomic prediction, we computed a selection index (SI) as:

$$\mathrm{SI} = \sum_{t=1}^{T}(w_t + \sum_{j=1,j\neq t}^{T} \rho_{t,j} w_j) g_t, \quad (5)$$

in which w_t's are the weights given to traits $t = 1, \ldots, T$, $\rho_{t,j}$ is the empirical correlation between the GVs of traits t and j, and g_t's are the GVs of trait $t = 1, \ldots, T$. For this study, we assigned equal weights to all traits in order to compare the general evolution of the genetic gain based on the breeding values obtained with the different models. Based on the SI, at each generation the top 10% eligible males were selected and mated with all females from the latest generations, in order to generate the

subsequent generation, maintaining at each new generation a female-to-male proportion of 80%-20%. Males were kept for breeding across four generations, while females were kept for two generations only. This selection and mating design were chosen to mimic the breeding program of a dairy cattle population. A base population of independent individuals (i.e. generation zero, or G0) was considered to start the selection process, followed by seven simulated generations under selection referred to as G1, …, G7.

This simulation scheme under selection allowed us to both study the evolution of the nonlinear genetic relationships between traits when the population is under selection, and compare the genetic gain of selection, based on the additive GV obtained with the different models, using their values to construct the SI in equation (5). Because of the way we simulate the data, the PGV obtained by all prediction models could express non-additive effects. Since we are interested in the additive PGV, we computed it as follow: at each generation we simulated the genotype of 20 offspring for each eligible male based on random mating schemes with the females. Then, the PGVs of these simulated offspring were predicted with the different models (GBLUP, DL, and DLGBLUP) trained on a dataset where all individuals until the latest generation were genotyped and phenotyped. The additive PGV of the eligible males were finally computed as the average of the PGVs of their simulated offspring.

The GVs for the reference trait on all subsequent generations were simulated using the same original QTL effects of G0, and the other dependent traits were simulated as previously described in the 'Simulated Data' section. The number of individuals in each of the generations G1 to G7 was maintained as 5,000. With each new generation, the reference population to perform the prediction models was updated to comprise all individuals from G0 to the latest with both genotypes and phenotypic records, and then used to perform the genetic evaluation. For DL and DLGBLUP, the genetic evaluation model with a new generation was initialized using the weights (DL parameters) obtained at the evaluation of the previous generation.

Evaluation metrics

To evaluate the performance of each model in predicting genetic values, we used the Pearson correlation coefficient to calculate the prediction accuracy for each individual trait, as well as the MSE between the TGV and PGV, and performed a visual assessment of the relationship form between traits. To evaluate the performance of each model with respect to genomic selection, we used the genetic gain computed as the difference between the mean of the TGV for individuals in G0 and the mean of the TGV for individuals in the next generation.

Results

Here we present the results of DL-based model trained following the second training approach (sequential training) which gave better results. In addition, for the simplicity, we called the traits based on the form of relationship they have with the reference trait, for example quadratic trait correspond to the trait with f as quadratic.

Genomic Prediction

Prediction Accuracy

Figure 2 presents the boxplots comparing the performance of the GBLUP, pure DL, and DLGBLUP models using only the QTL as input data, with respect to the prediction accuracy and the MSE, over the 20 replicates performed.

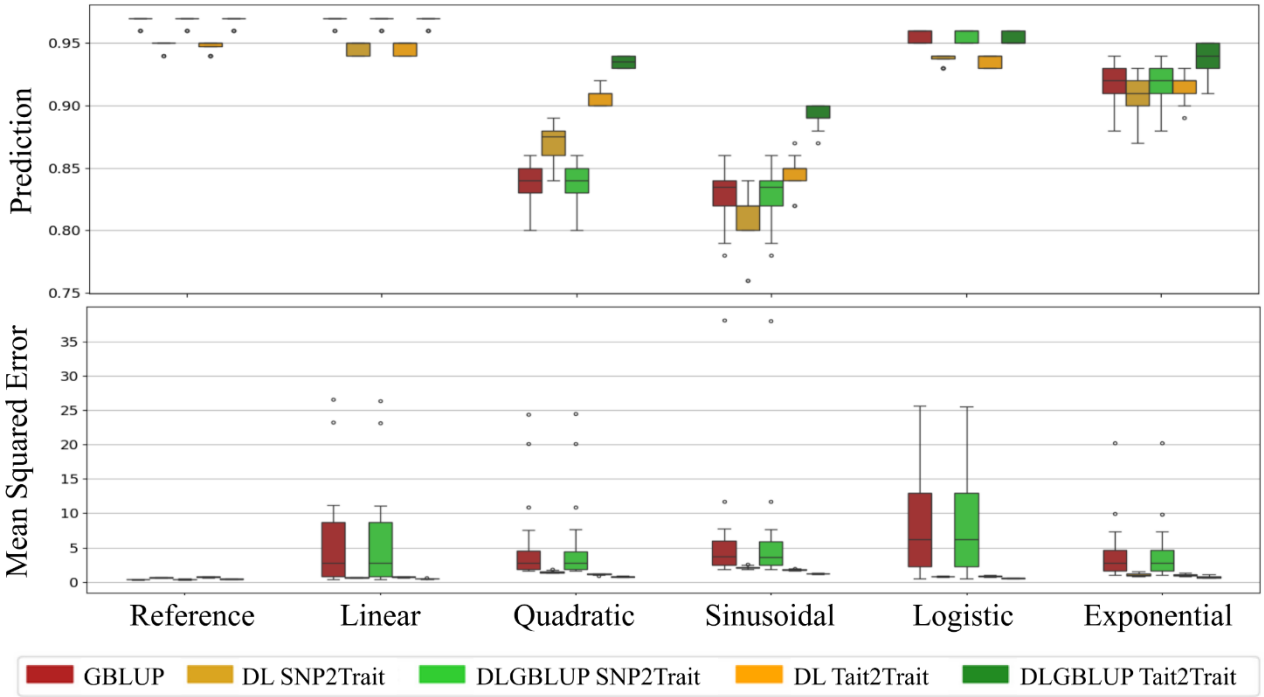


Figure 2: Performance comparison of different prediction models for various traits. GBLUP (red), Deep Learning (DL) (yellow) and DLGBLUP (green); with the two components SNPs2Trait and Trait2Trait. The prediction was made using genomic data with just QTL.

At the additive effect level, GBLUP demonstrated superior prediction accuracy compared to the pure DL SNP2Trait models for almost all trait except the quadratic one; nonetheless, the later presented a consistently lower MSE. Once trained to predict the output of GBLUP, the results of DLGBLUP SNP2Trait were very similar to those of GBLUP.

Both the GBLUP and DLGBLUP SNP2Trait models excelled in predicting the GV for the linear and logistic traits, achieving remarkable accuracy, while their performance was diminished for the other nonlinear traits, but always superior to an accuracy of 0.8 for every trait.

The prediction accuracy of DLGBLUP for the quadratic, sinusoidal, and exponential traits was improved after passing the PGV through the Trait2Trait step of the model that predicted the GV from

its relationship with other traits. This improvement varied from 0.01 to 0.1, depending on the type of nonlinear relationship between traits, and on the previous GV predicted from the additive effect. After passing through the Trait2Trait step, the DLGBLUP model did not provide any further improvement for the prediction accuracy of the linear traits (reference included), which were already well predicted by GBLUP.

The DLGBLUP model outperformed the pure DL model by a margin of 0.01 to 0.04 in terms of prediction accuracy. Of all the evaluated models, the DLGBLUP Trait2Trait model consistently yielded the lowest MSE across all dependent traits, with more reliable and precise estimation, with a prediction accuracy either higher than all the other models, or just as high as GBLUP.

Prediction of nonlinear relationship function

Both DLGBLUP and pure DL models, with the MeanTrait2Trat step, were capable of identifying the true form of the relationship between traits, whether it was linear or nonlinear. Moreover, the shape of the PGV was adjusted according to the trait's relationship through the Trait2Trait part. A more accurate input, i.e. the GBLUP predictions, to the MeanTrait2Trait and Trait2Trait steps, enabled the prediction of a clearer and more precise relationship between traits, as illustrated in Figure 3.

Additionally, MeanTrait2Trait was able to detect the genetic relationships between traits, while GBLUP was able to do so only for the linear trait. For the other nonlinear traits, we observed that GBLUP transformed their relationship to be linear, sometimes identifying a level of linear relationship (logistic and exponential traits), or completely missing any relationship between traits (quadratic and sinusoidal traits).

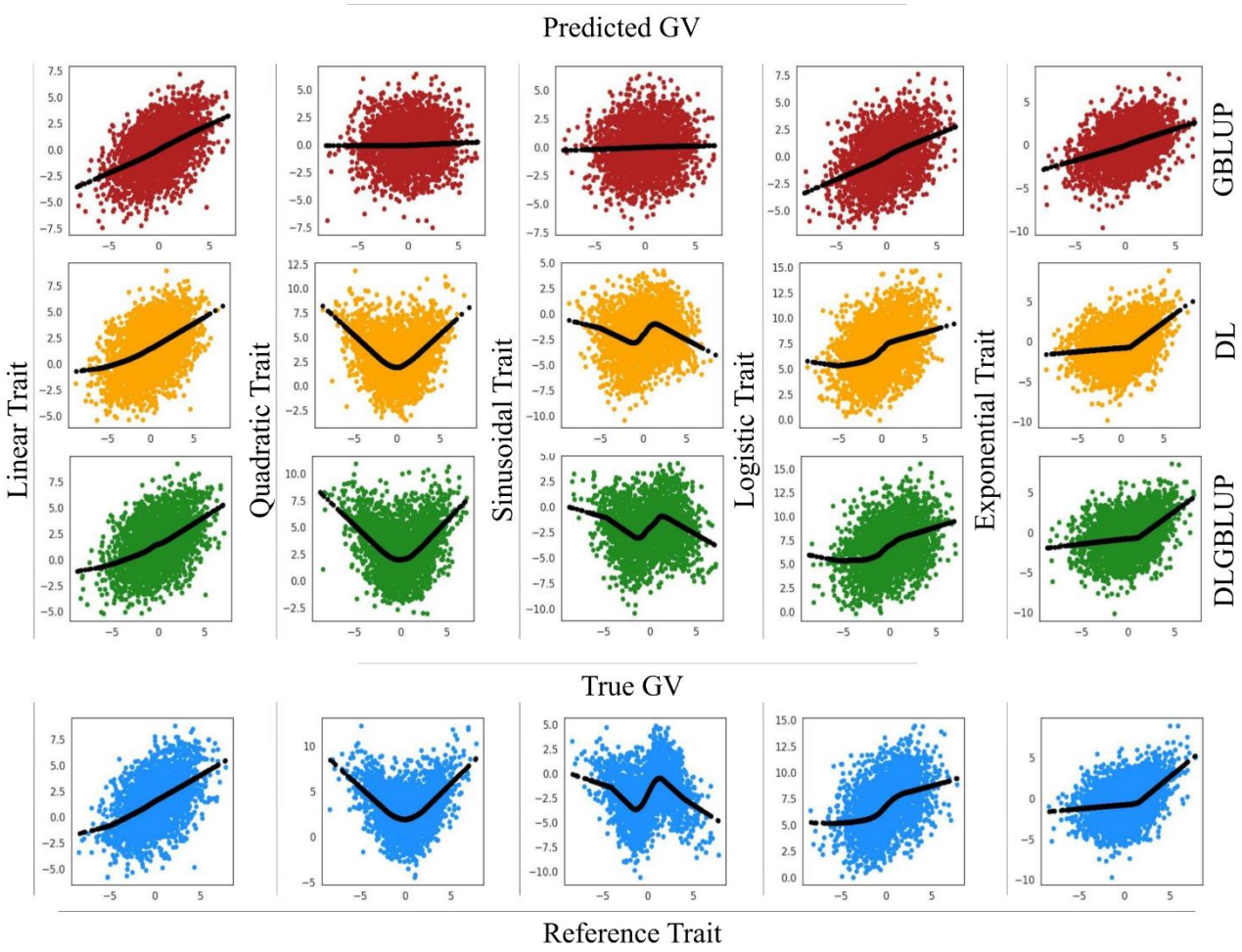


Figure 3: Comparison plots of predicted genetic relationship between traits with different prediction model. The reference trait is represented on the x-axis and all dependent traits on the y-axes. The colored points represent the PGV of GBLUP (red), DL (yellow) and DLGBLUP (green) and the TGV (blue) of a population. In black the mean relationship between traits as obtained from MeanTrait2Trait module. The prediction was made using genomic data with just QTL, h^2 equal 0.3 and 50 specific QTLs.

Effect of Heritability, number of specific QTL, and level of Genetic Relationship

Heritability directly influences the quality of GV predictions from additive effects. The higher the heritability, the higher the prediction accuracy of GBLUP, as well as the prediction accuracy made

by the Trait2Trait part and its improvement in comparison to GBLUP, for the nonlinear traits, reaching 0.17 for the quadratic traits with heritability of 0.6. Moreover, there was an effect of the number of specific QTL that control the genetic relationship (Figure 4).

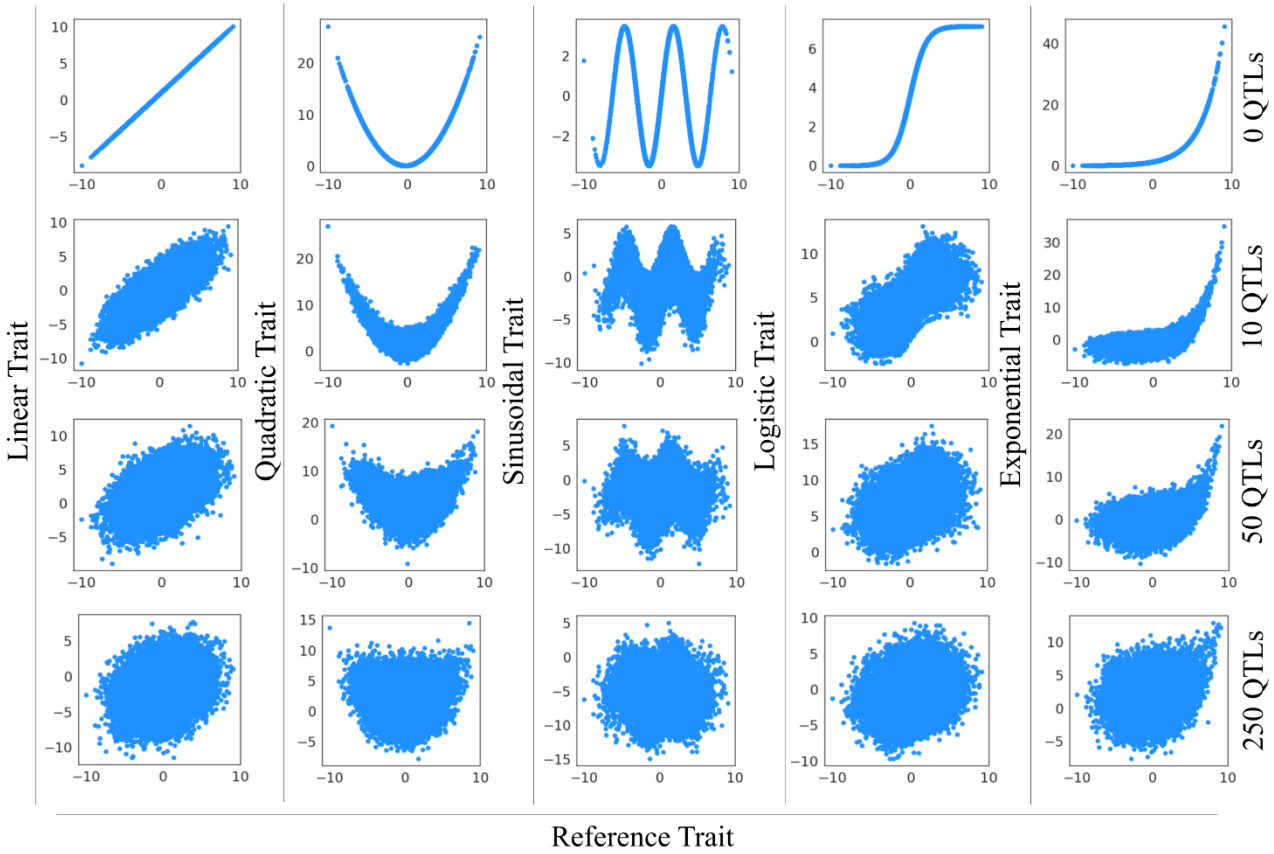


Figure 4: Comparison plots of true genetic relationship between traits with different number of specific QTLs. The reference trait is represented in x-axis and all dependent traits in y-axes. The blue points represent the True GV of a population.

The higher the number of specific QTL the less clear the relationship between traits. A higher genetic relationship -a lower number of specific QTL in our case- intensifies the nonlinear effects, reducing GBLUP's prediction accuracy while simultaneously increasing the potential for improvement via the Trait2Trait component, as shown in Table 1.

Table 1: Prediction accuracy comparison of different prediction models with different heritability and number of specific QTL.

			Traits					
Heritability	**QTLs**	**Model**	**Reference**	**Linear**	**Quadratic**	**Sinusoidal**	**Logistic**	**Exponential**
0.05	50	GBLUP	0.83	0.8	0.8	0.82	0.84	0.83
		DL	0.72	0.67	0.7	0.72	0.74	0.76
		DLGBLUP	0.82	0.79	0.8	0.81	0.83	0.83
0.15		GBLUP	0.93	0.94	0.82	0.87	0.93	0.91
		DL	0.88	0.89	0.81	0.83	0.89	0.87
		DLGBLUP	0.93	0.93	0.87	0.87	0.92	0.92
0.3		GBLUP	0.97	0.97	0.84	0.83	0.96	0.92
		DL	0.95	0.95	0.91	0.85	0.94	0.92
		DLGBLUP	0.97	0.97	0.94	0.89	0.95	0.94
	0	GBLUP	0.99	0.99	0.13	0.15	0.94	0.79
		DL	0.97	0.97	0.85	0.28	0.93	0.88
		DLGBLUP	0.99	0.99	0.95	0.89	0.97	0.93
	10	GBLUP	0.98	0.98	0.42	0.6	0.95	0.86
		DL	0.96	0.96	0.82	0.6	0.94	0.9
		DLGBLUP	0.98	0.97	0.9	0.84	0.96	0.94
	250	GBLUP	0.92	0.93	0.9	0.9	0.93	0.92
		DL	0.89	0.89	0.86	0.86	0.89	0.88
		DLGBLUP	0.92	0.92	0.91	0.9	0.93	0.92
0.6	50	GBLUP	0.99	0.99	0.79	0.81	0.96	0.93
		DL	0.98	0.98	0.96	0.86	0.97	0.96
		DLGBLUP	0.99	0.99	0.96	0.88	0.98	0.97

Effect of Non-QTL SNPs

Table 2 presents the effect on the model's performances of incorporating SNPs that were not QTL in the genomic data used. The most accurate predictions were achieved when the input data comprised exclusively all the QTLs implicated in the TGV of a trait, as expected, with DLGBLUP achieving the best predictions. When excluding half of the QTLs, or when adding non-causative SNPs to the input genotypes the prediction accuracy decreased. The decrease in prediction accuracy when non-causative SNPs were introduced was progressive as more and more SNPs were included to the genomic data, which results in a decrease in the advantage of DLGBLUP. Excluding all QTL, and

keeping only non-causative SNPs in LD with the QTL, resulted in PGV with accuracies ranging from 0.55 to 0.7, and an improvement made by DLGBLUP ranging from 0.01 to 0.03.

Table 2: Prediction accuracy comparison of different prediction models using genomic data with different numbers of SNPs, a $h^2$of 0.3 and 50 specific QTLs.

		Traits					
	Method	**Reference**	**Linear**	**Quadratic**	**Logistic**	**Sinusoidal**	**Exponential**
Half QTL	DL	0.82	0.75	0.72	0.77	0.63	0.65
	GBLUP	0.82	0.76	0.66	0.78	0.61	0.64
	DLGBLUP	0.81	0.74	0.72	0.77	0.63	0.65
Just QTL	DL	0.95	0.95	0.91	0.94	0.85	0.92
	GBLUP	0.97	0.97	0.84	0.96	0.83	0.92
	DLGBLUP	0.97	0.97	0.94	0.96	0.89	0.94
10% SNPs	DL	0.94	0.94	0.9	0.94	0.86	0.92
	GBLUP	0.95	0.95	0.8	0.95	0.81	0.91
	DLGBLUP	0.94	0.94	0.91	0.94	0.87	0.93
25% SNPs	DL	0.87	0.89	0.82	0.88	0.76	0.86
	GBLUP	0.91	0.87	0.75	0.9	0.76	0.87
	DLGBLUP	0.9	0.9	0.84	0.89	0.8	0.88
50% SNPs	DL	0.82	0.83	0.74	0.82	0.68	0.8
	GBLUP	0.86	0.86	0.7	0.86	0.7	0.82
	DLGBLUP	0.85	0.85	0.78	0.85	0.73	0.83
75% SNPs	DL	0.78	0.79	0.71	0.79	0.66	0.77
	GBLUP	0.84	0.84	0.68	0.84	0.69	0.79
	DLGBLUP	0.82	0.83	0.75	0.82	0.7	0.8
All SNPs	DL	0.76	0.77	0.69	0.78	0.65	0.75
	GBLUP	0.82	0.82	0.67	0.82	0.68	0.78
	DLGBLUP	0.8	0.81	0.73	0.81	0.69	0.78
NO QTL	DL	0.61	0.64	0.57	0.67	0.55	0.64
	GBLUP	0.63	0.67	0.56	0.7	0.57	0.65
	DLGBLUP	0.63	0.67	0.59	0.69	0.58	0.66

Genomic Selection

Genetic relationship across generations

Figure 5a shows the evolution of the genetic relationships between the TGV of the traits over the seven generations, under a multi-trait selection using the linear SI based on the additive GV obtained with pure DL. As expected, the relationship between the reference trait and the linear trait remained

linear over all the generations under selection. In contrast, the relationship between the reference trait and the quadratic trait evolved from nonlinear to linear very quickly, with its nonlinearity being almost imperceptible from G1. The remaining three traits that were simulated with a nonlinear relationship with the reference trait (sinusoidal, logistic, and exponential) preserved their nonlinear characteristics. However, while the shape of the relationship of the sinusoidal and exponential traits remained the same, the latter trait had its relationship attenuated, and the shape of the relationship of the logistic trait changed completely with respect to the original one in G0. These trends of transformations (or not) of the shape of the relationships persisted whether selection was based on any of the predicted additive GV [See Additional file 1, Figure S1].

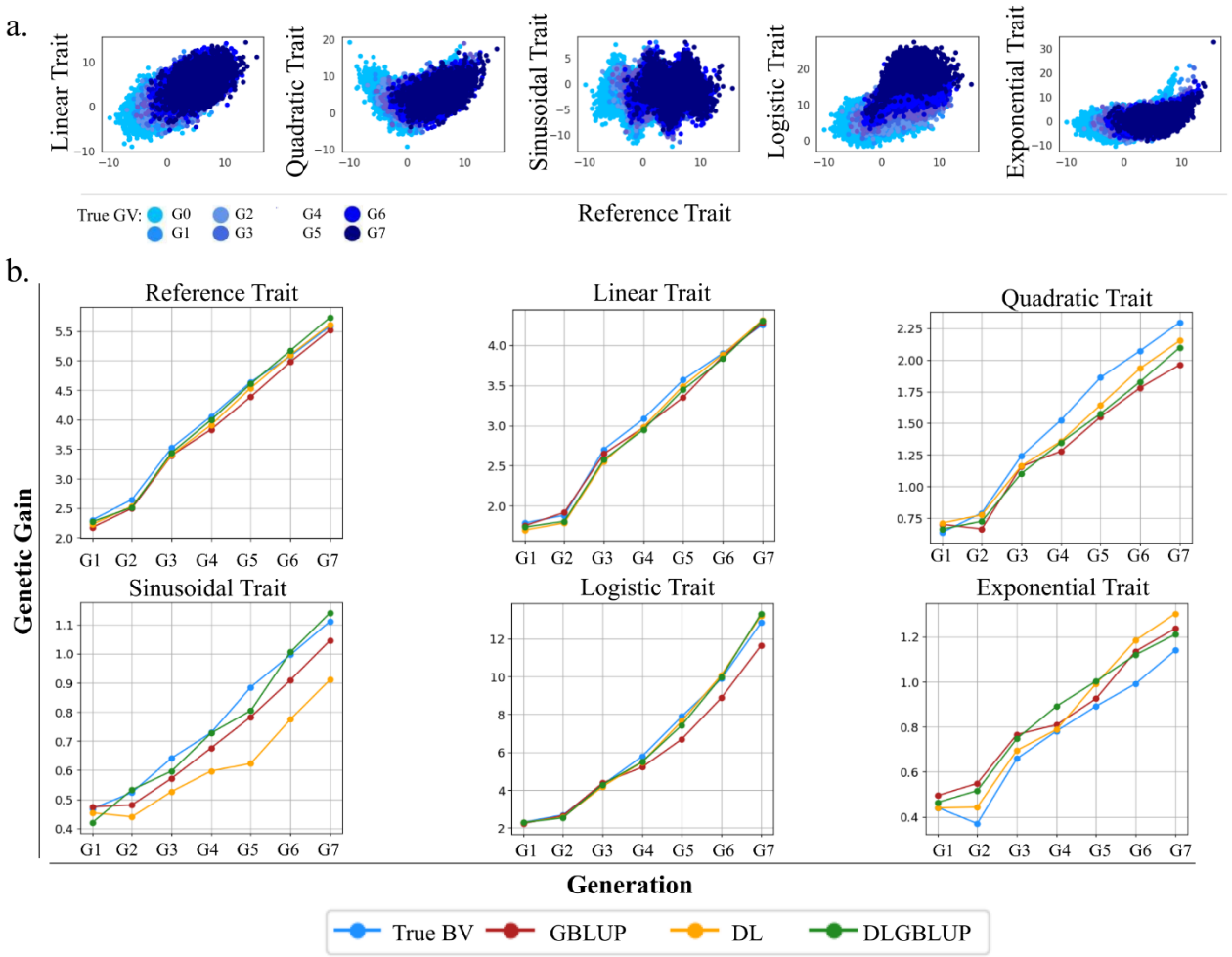


Figure 5: Comparative analysis of traits progression across generations. (a) Plots of relationship between the TGV of reference trait (x-axis) and dependent traits (y-axis) over 8 generations based on DL predictions. (b) Comparison between the Genetic Gain of 7 generationswith 15% male selection using the true and the predicted additive GV with GBLUP, DL, and DLGBLUP for all traits.

Genetic Gain

Genetic gain was achieved for all the traits, following the selection of the top 10% males (Figure 5b), based on the SI combined with any of the additive GV. The magnitude of this gain varied depending on the form of relationship, with the most substantial gains observed for the logistic trait. For both reference and linear traits, the progress based on the SI using the additive GV from all models were equivalent. Conversely, for each of the nonlinear traits, the genetic progress achieved based on the SI using the additive GV from either DL or DLGBLUP, or both for some traits, was greater than the genetic progress achieved using the additive GV from GBLUP.

Discussion

In this study, we proposed a novel hybrid model, DLGBLUP, which integrates DL and statistical methods to perform a multi-trait prediction of nonlinearly related genetic values. Such nonlinear genetic relationships between traits have not been greatly explored in GP, due to the difficulty in addressing such nonlinearity using the classical statistical methods. Although DL methods have been widely explored in the recent years as an alternative for the statistical methods in GP, the results obtained with DL in real data sets have not yet significantly overcome those from GBLUP [6]. Moreover, in spite of DL's capacity to model nonlinear patterns, these methods have not been greatly explored to model between-traits relationships.

One possible reason why DL did not achieve higher performance in GP compared to GBLUP, may be the fact that in GP, the objective falls in predicting the individuals GV, which are not

observed. While most studies implemented models that have been previously applied in different domains, in GP these models have to operate on the genotype-phenotype relationships - built-up in the hidden layers of DL rather than present in the target- to predict the GV. Some studies incorporated a biological interpretation into their models, which performed better than those DL models that maintained the same assumptions as GBLUP. GenNet [25] used layers embedded with prior biological knowledge, such as gene and chromosome annotation. DeepGBLUP [15], on the other hand, considered that a phenotype is the sum of additive, dominance, and epistasis effects (computed using GBLUP) in addition to an initial GV predicted from adjacent markers' effects extracted using a locally connected layer. In this study, we proposed DLGBLUP model that can predict a GV by learning GBLUP's predictions and then adjusting the predictions according to the relationships between traits.

Genomic selection in livestock or plant breeding programs consists of integrating the information of various traits into a SI, to select individuals that ensure genetic progress for traits of commercial interest, while not adversely impacting other traits that may be correlated to the targeted ones. To achieve this progress, two conditions are required: 1) A precise prediction of the individuals' GV and 2) A good understanding of the relationships between traits. If one of these conditions is unavailable, accurately selecting animals for multiple traits may be challenging. Until now, statistical methods – and particularly GBLUP – have been the leading choice of breeders for performing genetic evaluations. On the prediction level, these statistical methods showed a competitive performance regarding their accuracy of prediction and current computational resources required. These methods are, however, restricted to the assumption of linearity for the relationships between traits. While this assumption may not affect the ranking of individuals on an intra-trait perspective, in the case of nonlinearity in the relationships between traits, the assumption of linearity may have an effect on the inter-trait level, when performing a genetic evaluation.

Indeed, we have shown in our study that the pure DL and DLGBLUP models, which are flexible to the possibility of nonlinear relationships between traits, outperformed GBLUP on the simulated nonlinear traits, successfully capturing the shape of the relationship, especially if the provided input is more accurate. In addition, we demonstrated the critical role of an accurate SNPs selection step in minimizing input noise without eliminating important information, to enhance prediction accuracy. A significant advantage of DLGBLUP is that it can be complementary to any other single or multi-trait statistical methods, enabling it to achieve superior results, when compared to using DL alone.

While research on the applications of DL for genomic prediction has become more and more widespread in the recent years, works that apply DL for multitrait prediction models was less popular. Nevertheless, a previous study considered that relationships between the elements of the output layer could be learned and captured automatically by a neural network with shared neurons and weights [18]. However, this study did not explore whether these trait relationships are linear or nonlinear. In contrast, our study reveals that the SNP2Trait module for the DL model was unable to detect nonlinear relationships. Instead, the DL model required a proper and exclusive network to map one trait (input) to another (output), in order to capture nonlinearity.

Our results were obtained on simulated data, for which the dependent traits were partially conditioned on the GV of a reference trait directly simulated from the genomic data. Although the application to real data of the DL models proposed by us would be a natural extension of this work, this practical approach is a complete study in itself, which would derive from the main scope of this work: introducing DL methods capable to identify nonlinear relationships between traits. Beyond the DL model developed and presented here, a number of extra factors must be considered for a real data application: First, the types of phenotypic records are diverse, for example often binary for reproductive traits and sometimes continuous traits deviate from the assumption of normality, such as some of the health traits. Second, some traits have more than one record measure per animal,

requiring a permanent effect included into the model, a lengthy extension to the already complex DL models by us presented. Third, real genomic data usually contain more non-causative SNPs than the true QTL, a scenario in which we demonstrated that the models are more sensitive when it comes to predicting nonlinear relationships between traits. Last, but not least, genetic evaluation models must deal with all effects, genetic and non-genetic (e.g. the permanent environment, contemporary group, potential heterogenous variances). As previously mentioned, contemplating all these effects with the DL models would be a lengthy extension of the complex novel models developed and here presented, meriting a study in itself. Moreover, a preliminary attempt to operate the DL models on yield deviations (which synthetize only the genetic and independent residual effects for each individual) obtained from a previous evaluation accounting for all the genetic and non-genetic effects showed to be an unsuccessful approach. This approach failed to capture nonlinear relationships between traits, not due a failure from the DL models, but because the BLUP pre-processing to obtain the yield deviations forced a linearized output due to the model's nature, thus erasing completely any form of nonlinearity. Therefore, further work is required to extend our approach to real data.

Better understanding how multiple traits involved in a breeding program are related is sure to improve the genetic progress obtained by artificial selection. Here, we maintained the use of a linear SI to select individuals, however considered the additive GV obtained with the different models (GBLUP, pure DL, and DLGBLUP) for comparison, and showed that for all traits after seven generations, genetic gains from SI based on the additive GV from either pure DL or DLGBLUP were equal or superior to the gains from SI based on the additive GV from GBLUP.

Conclusions

In this study, we proposed a pure DL model and a hybrid variation, DLGBLUP, two framework that account for nonlinear genetic relationships between traits to predict their GV. From input SNP data and initial GBLUP predictions, that excel on predicting additive effects, the hybrid DL model can

learn to predict GBLUP outputs and then readjusts the prediction for potential nonlinear relationships between traits, when pertinent. Applied to simulated data, DLGBLUP was successful in improving the accuracy of the predicted GV in scenarios where nonlinear relationships between traits was present, in comparison to GBLUP. This greater prediction accuracy of the nonlinearly related traits was due to the ability of DLGBLUP in correctly identifying the mean patterns of such relationships. Moreover, we showed that selection using a SI built based on the additive PGV from either pure DL or DLGBLUP achieved greater genetic gain than selection using a SI built based on the additive PGV from the traditional GBLUP.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests

Funding

This research was funded by a CIFRE PhD grant from Eliance, with financial support from the Association Nationale de la Recherche et de la Technologie (ANRT-Cifre) and APIS-GENE (Paris, France)

Authors' contributions

FS designed the deep learning models and the data simulation method, performed all the analysis, and did the main writing of the manuscript. PC participated in the conceptualization of the study, in

the discussion and interpretation of the results, and in revising the manuscript. HG participated in the design of the deep learning models, and in revising the manuscript. RS participated in the discussion and interpretation of the results. TT contributed to the interpretation of preliminary results, leading to relevant decisions in the conception of the study. TMH participated in designing the data simulation method, in the design of the deep learning models, and in revising the manuscript. BCDC participated in the conceptualization of the study, in designing the data simulation method, in the discussion and interpretation of the results, and contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by INRAE Metaprogramme DIGIT-BIO (Digital biology to explore and predict living organisms in their environment). We would like to think GenIALearn team for their valuable discussion and contribution, and Jocelyn de Goer De Herve from (INRAE /UMR EPIA) for managing the server on which DL work was done.

References

1. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819-29.
2. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414-23.
3. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95:4114-29.

4. Misztal I, Legarra A, Aguilar I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J Dairy Sci. 2009;92:4648-55.

5. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436-44.

6. Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P et al. A review of deep learning applications for genomic selection. BMC Genomics. 2021;22:19.

7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in neural information processing systems. 2017. p. 5998-6008.

8. Alzubaidi L, Zhang J, Humaidi AJ et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data. 2021;8:53.

9. Shinde PP, Shah S. A review of machine learning and deep learning applications. In: Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). Pune, India. 2018. p. 1-6.

10. Zafar I, Anwar S, Kanwal F, Yousaf W, Nisa FU, Kausar T, Ain Q ul, Unar A, Kamal MA, Rashid S, Khan KA, Sharma R. Reviewing methods of deep learning for intelligent healthcare systems in genomics and biomedicine. Biomed Signal Process Control. 2023;86(Part B):105263.

11. Choudhary K, DeCost B, Chen C et al. Recent advances and applications of deep learning methods in materials science. npj Comput Mater. 2022;8:59.

12. Jubair S, Tucker JR, Henderson N et al. GPTransformer: a transformer-based deep learning method for predicting fusarium related traits in barley. Front Plant Sci. 2021;12:761402.

13. Wang K, Abid MA, Rasheed A, Crossa J, Hearne S, Li H. DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. Mol Plant. 2023;16:279-93.

14. Bellot P, de Los Campos G, Pérez-Enciso M. Can deep learning improve genomic prediction of complex human traits? Genetics. 2018 Nov;210:809-19.

15. Lee HJ, Lee JH, Gondro C et al. deepGBLUP: joint deep learning networks and GBLUP framework for accurate genomic prediction of complex traits in Korean native cattle. Genet Sel Evol. 2023;55:56.

16. Van Rheenen W, Peyrot WJ, Schork AJ et al. Genetic correlations of polygenic disease traits: from theory to practice. Nat Rev Genet. 2019;20:567-81.

17. Hill WG. Genetic correlation. In: Brenner's Encyclopedia of Genetics. San Diego, US: Elsevier. 2013. p. 237-39.

18. Montesinos-López OA, Montesinos-López A, Crossa J, Gianola D, Hernández-Suárez CM, Martín-Vallejo J. Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. G3 (Bethesda). 2018;8:3829-40.

19. Fuerst-Waltl B, Sölkner J, Essl A, Hoeschele I, Fuerst C. Non-linearity in the genetic relationship between milk yield and type traits in Holstein cattle. Livest Prod Sci. 1998;57:41-47.

20. Castañeda-Bustos VJ, Montaldo HH, Valencia-Posadas M, Shepard L, Pérez-Elizalde S, Hernández-Mendo O, Torres-Hernández G. Linear and nonlinear genetic relationships between type traits and productive life in US dairy goats. J Dairy Sci. 2017;100:1232-45.

21. Cuyabano B. GenEval. 2021. Available from: https://github.com/bcuyabano/GenEval. Accessed 4 Sep 2023.

22. Pérez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014;198:483-95.

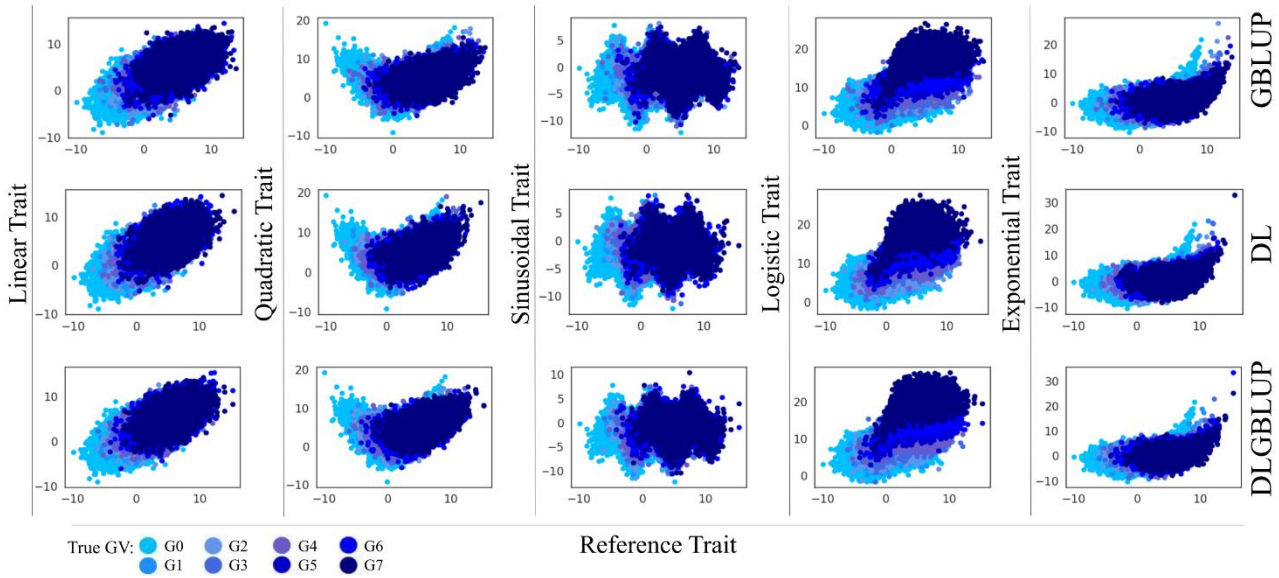
23. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.

24. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY, USA: Springer New York Inc.; 2001.

25. Van Hilten A, Kushner SA, Kayser M et al. GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. Commun Biol. 2021;4:1094.

Additional files

Additional file 1 Figure S1



Title: **Comparative plots of traits progression across 8 generations under selection.**

Description: Plots of relationship between the TGV of reference trait (x-axis) and dependent traits (y-axis) over 8 generations based on additive PGV of GBLUP, DL and DLGBLUP.