

# Evaluation of Deep Learning for predicting rice traits using structural and single-nucleotide genomic variants

Ioanna-Theoni Vourlaki<sup>1</sup>, Sebastián E. Ramos-Onsins<sup>1</sup>, Miguel Pérez-Enciso<sup>1,2,3\*</sup>, Raúl Castanera<sup>1\*</sup>

<sup>1</sup>*Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Bellaterra, Barcelona 08193, Spain.*

<sup>2</sup>*Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain.*

<sup>3</sup>*Universitat Autònoma de Barcelona, Barcelona, 08193, Spain.*

\* For correspondence: [mperezenciso@gmail.com](mailto:mperezenciso@gmail.com), [raul.castanera@cragenomica.es](mailto:raul.castanera@cragenomica.es)

## Abstract

Structural variants (SVs) such as deletions, inversions, duplications, and Transposable Element (TE) Insertion Polymorphisms (TIPs) are prevalent in plant genomes and have played an important role in evolution and domestication, as they constitute a significant source of genomic and phenotypic variability. Nevertheless, most methods in quantitative genetics focusing on crop improvement, such as genomic prediction, consider Single Nucleotide Polymorphisms (SNPs) as the only type of genetic marker. Here, we used rice to investigate whether combining the structural and nucleotide genome-wide variation can improve prediction ability of traits when compared to using only SNPs. Moreover, we also examine the potential advantage of Deep Learning (DL) networks over Bayesian Linear models, which have been widely applied in genomic prediction. Specifically, the performance of BayesC and a Bayesian Reproducible Kernel Hilbert space regressions were compared to two different DL architectures, the Multilayer Perceptron, and the Convolution Neural Network. We further explore their prediction ability by using various marker input strategies and found that exploiting structural and nucleotide variation improves prediction ability on complex traits in rice. Also, DL models outperformed Bayesian models in 75% of the studied cases. Finally, DL systematically improved prediction ability of binary traits against the Bayesian models.

## Introduction

Rice (*Oryza sativa*) constitutes a fundamental staple crop, essential to humans for its nutritional and caloric value. World rice production has reached a plateau ([FAO, 2023](#)), and following the conventional breeding techniques, rice yield will soon not meet the high demand caused by the increasing world population. Therefore, we need to explore new approaches to secure global nutritional requirements by increasing at the same time the quality and quantity of rice yield. Genomic Prediction (GP) can help to achieve these goals, accelerating the breeding progress ([Meuwissen et al. 2001](#)). Various studies in plants have shown the effectiveness of GP in increasing breeding speed ([Jighly et al. 2019](#), [Tessema et al. 2020](#), [Xu et al. 2020](#), [Krishnappa et al. 2021](#)). GP framework has been widely used in rice studies for predicting various quantitative traits, reporting moderate to high predictive performance ([Xu et al. 2021](#)). Complex traits are controlled by numerous loci that are difficult to detect with genetic mapping. GP assumes that quantitative trait loci (QTL) will be in linkage disequilibrium (LD) with molecular markers. Thus, instead of detecting all the QTL associated with a trait, an indirect association between marker and trait is exploited.

Conceptually, since the number of genotyped individuals  $n$ , is typically smaller than the number of molecular markers  $p$ , GP faces statistical challenges such as large sampling variance and increased mean-square error. To overcome this limitation, variables must be selected or restrictions on the solutions must be applied or sometimes both. The main classes of GP methods are the genomic relationship-based method such as Genomic Best Linear Unbiased Prediction (GBLUP, [VanRaden, 2008](#)) and the SNP effect-based methods such as the Bayesian family ([Meuwissen et al. 2001](#), [Habier et al. 2011](#), [Pérez and De Los Campos, 2014](#)) and LASSO ([Tibshirani, 2011](#)). Some Bayesian models do not necessarily assume homogeneity across marker effects. They perform variable selection and shrinkage on the effects simultaneously using priors other than

Gaussian. BayesC is an example of this category assuming as a prior a normal distribution with constant variance while a fraction of marker has no effect ([Habier et al. 2011](#)). On the other hand, methods such as GBLUP involve restriction on the square of solutions (L2 norm), with the effect of the markers assuming to be normally distributed with equal variance.

Deep Learning (DL) networks are a collection of machine learning algorithms that have exhibited excellent performance in some prediction tasks ([Min et al. 2017](#); [Pattanayak 2017](#)). The DL models are trained in such a way to find complex relationships between data. DL networks consist of multiple layers and interconnected nodes. Each layer uses as input the output of the previous layer to optimize the prediction or classification. Numerous DL architectures have been proposed, such as Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN, [LeCun et al. 2015](#)). DL has been around for decades but only recently started to be widely implemented because of the easy implementation framework provided by various online libraries (e.g <https://keras.io/>, <https://pytorch.org/>). The performance of the DL networks depends on the accuracy of hyperparameter choice, which is not an easy task and requires abundant computation resources ([Young et al. 2015](#), [Chan et al. 2018](#)). For a review of DL tools applied to genomic prediction, see ([Pérez-Enciso and Zingaretti, 2019](#)).

Despite their features, various works have shown a performance of DL in genomic prediction comparable to linear models ([González-Recio et al. 2014](#); [Ma et al. 2017](#); [Bellot et al. 2018](#); [Montesinos-López et al. 2018](#)). [Zingaretti et al. \(2020\)](#) did not find a considerable advantage of DL over linear models, except when epistasis component was important. [Ehret et al. \(2015\)](#) found non-relevant differences between a GBLUP and a MLP model. In a wheat study ([Ma et al. 2017](#)), DL performed better than GBLUP when used to predict phenotypes from genotypes. Similarly, [Gianola et al. \(2011\)](#) found that MLP performed better than a Bayesian linear model in wheat. In another study in wheat, [Pérez-Rodríguez et al. \(2012\)](#) extensively compared the prediction performance of Radial Basis Function Neural Networks and Bayesian Regularizes Neural Networks against several linear models and semiparametric models such as Reproducible Kernel Hilbert Space (RKHS). The authors concluded that the non-linear models, such as DL, demonstrated a higher prediction ability than the linear models. Evaluating the potential of DL algorithms, [Sandhu et al. \(2021\)](#) showed higher prediction ability against the linear models in spring wheat. For an extensive review in GP using DL models see [Montesinos-López et al. \(2021\)](#).

While genomic prediction of quantitative traits has been broadly studied, binary and ordinal traits are of great importance in plant breeding as well. Particularly, in rice, culm morphology and stay green related traits are important targets for genetic improvement ([Kamal et al. 2019](#), [Chigira et al. 2023](#), [Lee and Masclaux-Daubresse 2021](#)). However, the prediction ability of different types of traits depends on the statistical models used. Thus, many studies in Genomic Selection (GS) focus on improving existing models or develop new ones as it was pointed out by [Montesinos-López et al. \(2019\)](#).

Most of the studies in GP rely only on SNPs, disregarding other sources of genomic variation such as structural variants (SVs). Among the different types of SVs, TIPs account for an important fraction and play a key role in plant evolution, from domestication to adaptation and breeding ([Dubin et al. 2018](#)). Studies in tomato and in rice found that the use of TIPs can lead to the identification of novel associations not detected by SNPs in genome-wide association studies ([Akakpo et al. 2020](#); [Carpentier et al. 2019](#); [Domínguez et al. 2020](#); [Castanera et al. 2021](#)). In a recent study, we showed that TIPs explain a sizable fraction of the genetic variance in several rice agronomic traits and significantly improve genomic prediction ([Vourlaki et al. \(2022\)](#)). However, TIPs are not the only type of structural variation in the genome. Other SVs such as deletions, inversions, and

duplications account for an important fraction of genetic variation and have been key in the domestication and diversification of plant crops ([Lye and Purugganan, 2019](#)). SVs have a high potential to generate large effect mutations when they affect genes or gene regulatory regions. Over the last few years, several studies have demonstrated the importance of presence-absence variation and structural variation as a source of gene expression and phenotypic variability in plant crops including rice, tomato or soybean ([Qin et al. 2021](#), [Alonge et al. 2020](#), [Liu et al. 2020](#)).

Here we investigate whether merging all the structural and nucleotide genome-wide variation can improve phenotypic prediction compared to using only SNPs in rice. Finally, we further explore the performance of DL in GP by (i) using multiple marker input strategies, (ii) proposing several approaches to accommodate large scale marker information, and (iii) optimizing network architectures. We also provide documented python code based in tensorflow 2 ([Abadi et al. 2016](#)) in the following link <https://github.com/ivourlaki/Deep-Learning-in-rice-for-prediction.git>.

## Materials and Methods

### Rice accessions and traits

In this study we used a subset of 738 accessions from the 3,000-rice genome project ([Li et al. 2014](#)). Accessions were chosen based on the availability of at least 15x of sequencing depth. The 738 accessions are representative of the main rice population groups: Aus/Boro (AUS, N=75), Indica (IND, N=451), Japonica (Jap, N=166), Aromatic (ARO, N=17). An additional group of admixed varieties (ADM, N=29) consisting of accessions that cannot be assigned to a specific rice group was also used. SNP-based group assignment from [Alexandrov et al. \(2015\)](#) and [Sun et al. \(2017\)](#) was used to identify the different subsets of this study. Studied traits are publicly available at IRRI SNP-Seek database (<https://snp-seek.irri.org/>). Among continuous traits, grain weight and time to flowering were used, whereas culm diameter (1<sup>st</sup> internode) and leaf senescence were selected among the binary traits available. Some of the ordinal traits were binned to balance the number of observations per class. Particularly, in leaf senescence values equal to 1 were assigned to class 1 while the rest ones range from 2-9 were recorded as 2 and were assigned to class 2. Finally, time to flowering was log-transformed.

### Markers

We used the filtered SNP dataset in [Yourlaki et al. \(2022\)](#). Specifically, a binary ped file format with the 3K RG CoreSNPs dataset for all chromosomes was downloaded from the SNP-Seek database. The original dataset consisted of 404,399 bi-allelic SNPs from 3,034 rice accessions, including the 738 accessions selected. We filtered out markers with minor allele frequency  $\leq 0.01$  and missing rate  $> 1\%$  using plink2 ([Purcell et al. 2007](#); [Chang et al. 2015](#)). Also, missing genotypes were imputed using Beagle 5.2 with default parameters ([Browning et al. 2018](#)). After filtering the final dataset consisted of 228,871 SNPs.

The TIP dataset described in [Castanera et al. \(2021\)](#) (containing two categories: MITE-DTX and RLX-RIX) was complemented with non-TE deletions (DEL), duplications (DUP) and inversions (INV) downloaded from SNP-seek database (3K RG Large Structural Variants release 1.0). We filtered out SVs events containing multiple overlapping deletions as these complex variants are difficult to genotype with short reads and are thus less reliable. SVs genotypes were recorded as 0/1 (absence/presence). All the markers with minor allele frequency  $\leq 0.01$  were filtered out. Finally, the dataset used in our analysis consists of 52,120 MITE-DTX, 21,517 RLX-RIX, 74,136 DEL, 25,670 DUP and 7,527 INV.

### Genetic variance inference

To estimate the genetic variance components explained by each marker set, we fitted the following linear model using RKHS [\(Gianola et al. 2006\)](#):

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z} \mathbf{u}_1 + \mathbf{Z} \mathbf{u}_2 + \mathbf{e} \quad (1)$$

where  $\mu$  is the general mean,  $y$  is the phenotype vector of size  $n$  (the number of accessions),  $Z$  is an identity incidence matrix,  $u_1$  and  $u_2$  are random effects of each of the marker groups and  $e$  is the residual. Random effects are assumed to be normally distributed  $u_1 \sim N(0, K_1 \sigma_1^2)$ ,  $u_2 \sim N(0, K_2 \sigma_2^2)$ , with constant variance  $K_1 \sigma_1^2$  and  $K_2 \sigma_2^2$ . Where  $K_1, K_2$  are genomic relationship matrices (GRM) obtained from the markers used in the corresponding model. We fitted the model five times using as  $K_1$  the GRM form SNPs while as  $K_2$ , GRM was obtained from MITE-DTX, RLX-RIX, DEL, DUP, INV, separately. The GRM were calculated using AGHMatrix [\(Amadeu et al. 2016\)](#). Model was implemented in BGLR package [\(Pérez and de Los Campos 2014\)](#) using default priors to estimate  $\sigma_1^2, \sigma_2^2$ .

### Genomic Prediction Models

#### Bayesian Regression Models

Two Bayesian methods were employed in this study: Bayesian RKHS and BayesC. RKHS is a kernel-based method that uses a ridge regression L2 regularization technique like GBLUP. BayesC is a variable selection method that estimates the effect of the markers. Both methods were applied to each trait separately. Particularly, for each method, various models were designed and applied comparing the predictive performance of using all the markers together versus using only SNPs. For RKHS, the models are described as follows:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z} \mathbf{u}_1 + \mathbf{Z} \mathbf{u}_2 + \mathbf{Z} \mathbf{u}_3 + \mathbf{Z} \mathbf{u}_4 + \mathbf{Z} \mathbf{u}_5 + \mathbf{Z} \mathbf{u}_6 + \mathbf{e}, \quad (2)$$

Where  $u_1, u_2, u_3, u_4, u_5, u_6$  are random effects  $u \sim N(0, K \sigma^2)$  with  $K$  corresponds to a GRM computed using SNPs, MITE-DTX, RLX-RIX, DEL, DUP, INV, respectively. Apart from model (2) where the different marker types are applied at the same time, other models were studied as well. Particularly, we were interested in investigating whether one matrix represented by different type of markers could result in a higher prediction in comparison to model (2) and to use only SNPs. As a result, the following models were studied:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z} \mathbf{u}_{10k\_variants} + \mathbf{e} \quad (3)$$

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z} \mathbf{u}_{10k\_SNPs} + \mathbf{e} \quad (4)$$

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z} \mathbf{u}_{Linked\_SNPs} + \mathbf{e} \quad (5)$$

Where  $u_{10k\_variants}$ ,  $u_{10k\_SNPs}$ ,  $u_{Linked\_SNPs}$ , are the effects of the 10,000 most associated markers among the six sets, the 10,000 most associated SNP effects and by using the SNPs linked to structural variation. LD between SNPs and SVs was calculated using ngsLD software ([Fox et al. 2019](#)), and we considered that a SNP was linked to an SV when  $r^2 \geq 0.8$ .

For BayesC, the complete models were:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}_{10k\_variants} \boldsymbol{\beta} + \mathbf{e} \quad (6)$$

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}_{10k\_SNPs} \boldsymbol{\beta} + \mathbf{e} \quad (7)$$

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}_{Linked\_SNPs} \boldsymbol{\beta} + \mathbf{e} \quad (8)$$

Where  $X_{10k\_variants}$ ,  $X_{10k\_SNPs}$ ,  $X_{Linked\_SNPs}$  are the standardized genotypes for the 10,000 most associated markers among all marker sets, the 10,000 most associated SNPs and the linked SNPs respectively. The 10,000 most associated were selected based on a linear regression-based genome-wide association analysis (GWAS) analysis. A detailed description is given in “Input Strategies” section. Finally, where  $\boldsymbol{\beta}$ , is the vector of effects for the corresponding matrix.

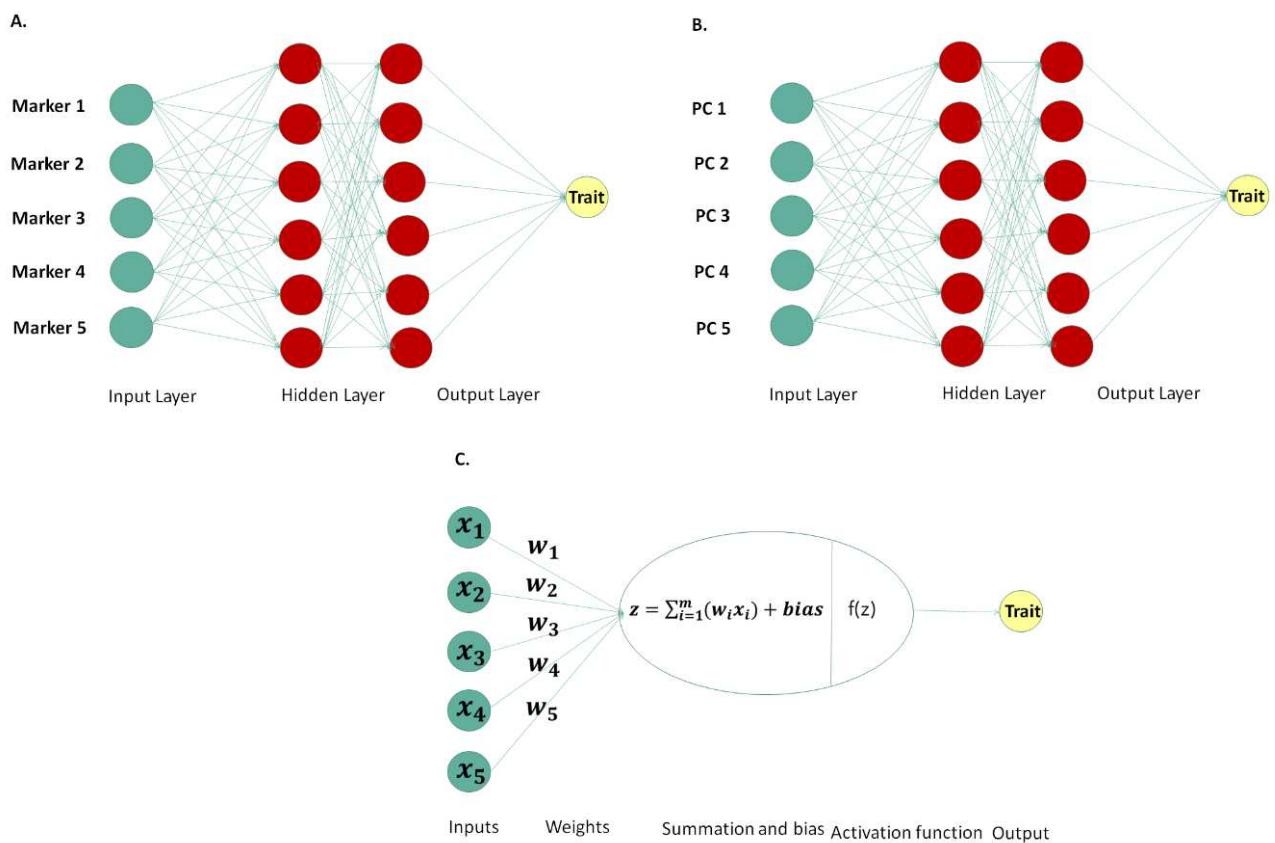
Using either RKHS or BayesC, phenotypes to be predicted were removed from the dataset and the model fitted using the remaining phenotypes. Prediction ability was assessed by computing two different metrics related to the type of trait. We computed the mean squared error (MSE) between predicted and observed phenotypes for the quantitative traits, whereas binary cross-entropy was employed for the binary traits. Both models were implemented using BGLR package. BayesC assumes that a proportion of markers will have zero effect with probability sampled from a beta distribution,  $\pi \sim Beta(p_0, \pi_0)$ . Here we chose  $p_0 = 5$  and  $\pi_0 = 0.01$ . For the case of binary traits option “response\_type=ordinal” was applied in both methods (RKHS, BayesC). Finally, BGLR was run for 100,000 iterations using default priors for RKHS.

### Multilayer Perceptron

One of the most popular DL architectures is the Multilayer Perceptron (MLP). MLP is a fully connected feedforward artificial neural network which transforms any input dimension to the desired dimension. All the neurons are connected to every neuron in the previous layer and then connected to every neuron in the next layer. Each neuron receives the initial inputs multiplied by a corresponding weight coefficient. Then the sum of all inputs multiplied by weight plus a bias, is passed to an activation function which introduces the non-linearity to the network transforming the inputs accordingly. We can represent the output of each hidden layer as (note the transposes):

$$\mathbf{Z}_l = f(\mathbf{X} \mathbf{W}^{(l-1)T} + \mathbf{b}^{(l-1)T}) \quad (9)$$

Where  $Z_l$  is the output of the  $l$  layer,  $b^{(l-1)T}$  is the bias vector of the first layer,  $X$  is a single matrix of all training examples so that we could compute all the prediction using a single matrix multiplication,  $W^{(l-1)T}$  is the weight matrix and  $f$  is a nonlinear activation function. The model is trained successively, that is, the output of neurons from the previous layer will be the input for the next layer. Figure 1 shows the basic workflow of MLP network.



**Figure 1:** Multilayer perceptron (MLP) representation with markers (A) and PCs (B) as input layers; bottom center shows the basic workflow of a perceptron or else neuron (C).

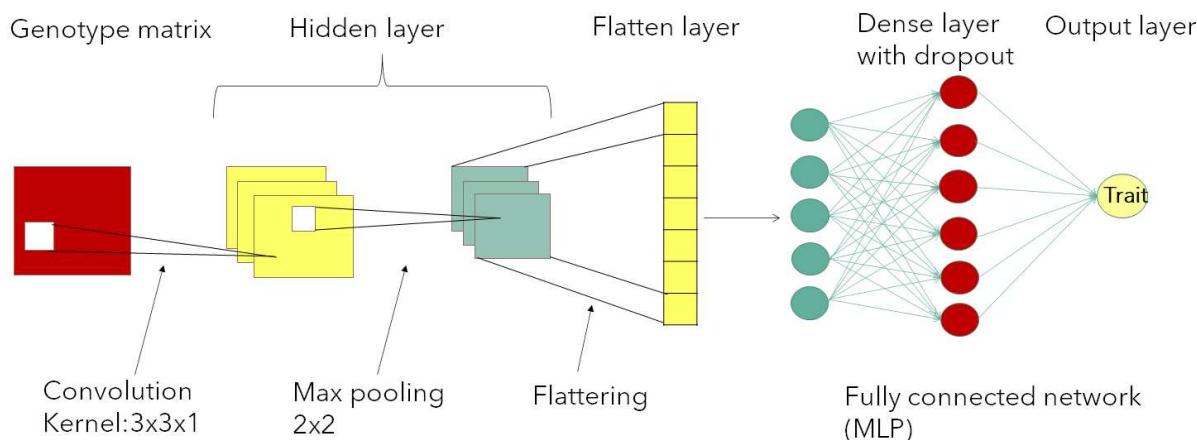
### Convolutional Neural Networks

Convolutional neural networks (CNNs) can utilize spatial relationships between nearby variables (e.g., pixels) of the input matrix. This architecture can accommodate situation where input variables are distributed along a space and are associated with each other such as linkage disequilibrium between nearby markers ([Pérez-Enciso and Zingaretti, 2019](#)). A CNN has hidden layers which typically consist of convolutional layers, pooling layers, flatten layers and fully connected dense layers. In each convolutional layer, CNN automatically performs a convolution that is a linear operation performed along the input of predefined width and strides by applying kernels or filters. The weights used are the same for all marker windows. The filter moves along windows of same sizes consist of markers performing a multiplication operation (dot product) until the entire matrix is traversed. The output of the convolutional function can be described as an integral transformation ([Widder, 1954](#)), as follows:

$$s(t) = (f * k)(t) = \sum_x k(t-x)f(x) \quad (10)$$

where  $k$  represents the kernel, convolution is the transformation of  $f$  into  $s(t)$ . The operation is performed over an infinite number of copies  $f$  resulting in the weighted sum shifting over the kernel. An activation function is applied after each convolution to produce the output layer. After nonlinearity has been applied to the feature map produced by the first layer, a pooling layer usually follows, aiming to reduce the dimensionality and smoothen the representation (Figure 2). The benefit of using CNN is their ability to develop an internal representation of a two-dimensional

matrix extracting the most important features. CNN leverages the fact that nearby input variables are more strongly related than the distant ones.

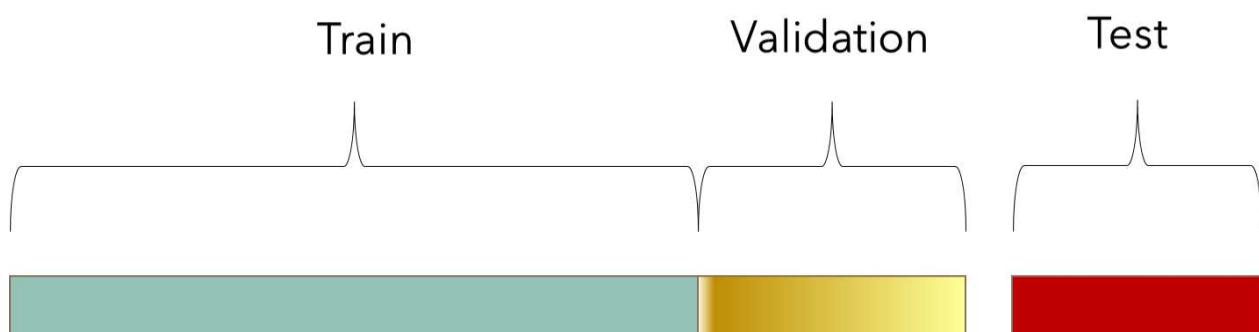


**Figure 2:** Convolutional Neural Network (CNN) representation used in study.

### Cross-validation and Independent Prediction

Here we evaluate the prediction accuracy by following two challenging validation scenarios relevant for breeding programs: prediction of individuals from two different population groups and prediction of randomly selected individuals from the rest ones. For the first strategy, we predicted performance of the admixed (ADM, N=29) and aromatic (ARO, N=17) groups using the rest accessions (IND, JAP and AUS/Boro). Since accessions to be predicted are not phylogenetically close to the accessions in the training set, it would be expected a low prediction ability from the models for this scenario.

In the second strategy, prediction accuracy was evaluated by implementing a 10-fold cross-validation (CV) where training population consisted of 90% of the data and testing set included 10% of the remaining data. Analysis was performed in each of the ten training sets separately assuming ten different validation scenarios. Since accessions are randomly selected and not based on their population group, samples in the training set might be related to the predicted ones. Note that, in the case of DL application, training population was further split in a validation dataset which included 20% of the training data set (Figure 3).



**Figure 3:** A visualization of how the three data sets, training, validation, and test are divided.

Validation data set is used during the training process of our network to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. It is important to mention that the model “sees” the data and use it for an evaluation of the process but never “learn from these”. After the model was trained, we retrieved the best

hyperparameters and performed prediction using the test dataset. The test dataset provides the gold standard used to make an unbiased final evaluation of the model. It is used only once a model is completely trained using the train and validation sets.

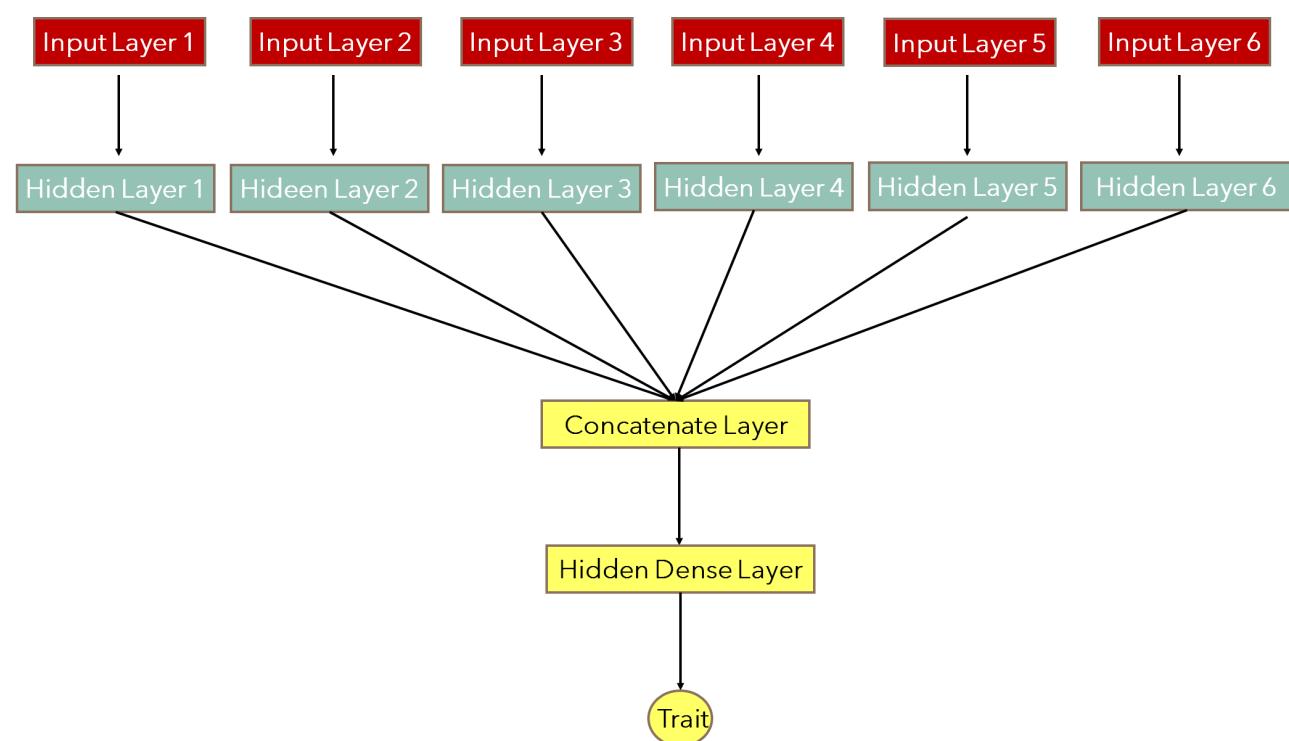
## Marker Input strategies

We used different marker input strategies aiming to enhance network flexibility and thus improve prediction ability. Four strategies were designed as follows:

1. **Most associated markers:** In the first input strategy, we merged the structural and nucleotide genomic-wide variation to test whether prediction accuracy can be improved. However, using the whole six genotype matrices (SNPs, MITE-DTX, RLX-RIX, DEL, DUP, INV) would add a high complexity in our network that might cause an overfitting. Studies have shown that using a subset of markers can result in equivalent prediction ability to using all the data sets ([Müller et al. 2017](#), [Bellot et al. 2018](#), [Vourlaki et al. 2022](#)). Thus, from the total of 409,892 molecular markers we selected the 10,000 most associated to the traits of interest. Specifically, we performed a genome wide association study (GWAS) fitting a linear model (single-marker regression analyses) to find associations between each of the six-marker set and each of the four traits (4x6). For each fitted model, a p-value corresponding to each marker was collected. From the collection of the p-values the 10,000 most associated was selected. GWAS was performed only to the training sets. Since we followed two different cross-validation strategies the process was repeated for each of those, that is for the across population training set, for the ten partitions training sets and for each trait. This strategy (hereafter referred to as “COMBINED\_VARIANTS”) was applied to DL and Bayesian linear models. Additionally, we selected the 10,000 most associated SNPs (hereafter referred as “SNPs”) to perform the same analysis and compare directly to the COMBINED\_VARIANTS strategy.
2. **Linked SNPs:** Using causal variants associated to the specific traits can result to prediction accuracy almost 1 ([Pérez-Enciso et al. 2015](#)), and structural variants are often causative of trait variability. Nevertheless, SNPs are easier to genotype in populations than SVs (ie, by genotyping chips). We reasoned that SNPs in LD with SVs could be used as SV replacement and would be easier to use in further experimental or breeding programs. Hence, we used ngdLD to detect SNPs in LD with SVs ( $r^2 \geq 0.8$ ) and used them for prediction. The software was applied to each trait and cross-validation scenario using only the training set. Then, for each case the unique pairs of SNPs-SVs with LD  $\geq 0.8$  were selected. The type, position and chromosome of the variants meeting the criteria were collected in a list and is provided in the github repository. The linked SNPs sets were used as marker sets across all the analysis. The selected abbreviation for this input strategy was “LINKED\_SNPs”. The average number of linked SNPs across the analyses was  $\sim 12,000$ .
3. **PCs single matrix:** In the third strategy, we exploited the advantages of principal components analysis (PCA) by incorporating it to neural networks. Studies have shown that using principal components in DL framework can be particularly advantageous ([Seuret et al. 2017](#)). In our study, PCs were computed based on eigenvectors for each of the obtained GRM. We run the analysis by feeding the network with a single matrix merging the PCs computed by the six GRMs introducing them as a single layer (Figure 1 (B)). Also, a separate analysis was performed for SNPs and linked SNPs testing whether this strategy will enhance the performance of the

model. The selected abbreviation for this input strategy was “COMBINED\_VARIANTS” with method applied MLP\_PCs.

4. **Multiple Inputs:** Here, we tested whether multiple input strategy could improve the prediction of traits. Particularly, we used the six PCs sets (computed by the GRM of each marker set) as six inputs feeding to the network in different input layers simultaneously. Other works have shown that a multiple input strategy can reduce overfitting and computational cost while at the same time exploits mixed data improving prediction ([Livieris et al. 2020](#), [Xiong et al. 2021](#)). Thus, the network accepted six different input layers which independently forwards in six different hidden dense layers. Next the six layers are merged by a concatenate layer (Figure 4). In order to compare directly this DL strategy against Bayesian Linear models, RKHS was performed with model eq. 2 using all the GRM as inputs at the same time. The selected abbreviation for this input strategy was “MULTIPLE INPUTS”.



**Figure 4:** Representation of Multiple inputs strategy employed in the present study.

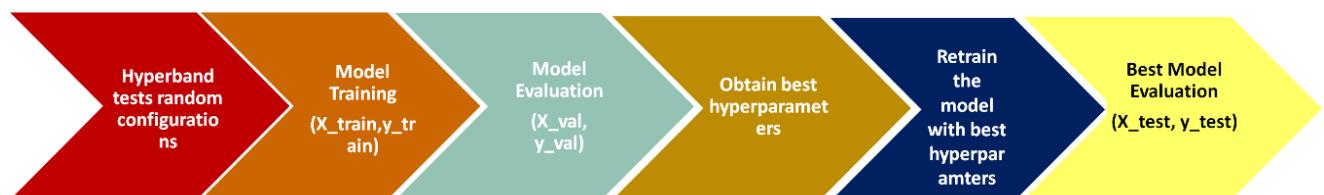
### Optimization of Hyperparameters

CNNs, MLP, BayesC and RKHS were implemented using the COMBINED\_VARIANTS (SNPs+SVs), as well as SNPs and LINKED\_SNPs as inputs separately. Additionally, MLP network was employed for using the six PCs as a single input matrix, as six different input layers and for PCs produced by GRM of SNPs. All the models were applied separately to each trait and to eleven cross-validation scenarios (10-fold and ARO/ADM across-population prediction). Table 1 shows the different models implemented in our analysis.

**Table 1:** Summary of the analyses performed.

	MLP, CNN, BayesC, RKHS, MLP PCs			RKHS, MLP PCs	
Trait	Combined variants	SNPs	Linked SNPs	Multiple Inputs	
<b>Culm diameter</b>	x	x	x	x	x
<b>Leaf senescence</b>	x	x	x	x	x
<b>Grain weight</b>	x	x	x	x	x
<b>Time to flowering</b>	x	x	x	x	x

For each of the different runs, hyperparameter tuning was performed obtaining the best hyperparameters and then retrained the model with the hyperparameters obtained by the search. Here Keras Tuner ([O'Malley et al. 2019](#)) library was used to pick the optimal set. Hyperparameters are the variables that control the training process and the topology of our model. When the model is built for hyperparameter tuning, the search space is also defined in addition to the model architecture. Then a tuner must be selected to determine which hyperparameter combinations should be tested. In our analysis we used the Hyperband tuner. The Hyperband tuning algorithm uses adaptive resource allocation and early stopping to quickly converge on a high-performing model. The algorithm trains a large number of configurations for a few epochs and carries forward only the top-performing half of models to the next round ([Li et al. 2018](#)) evaluating the performance by computing the MSE (for quantitative traits) and binary cross-entropy (for binary traits) were computed on a held-out validation set. The best model is the one that minimizes errors. After the hyperparameter search was finished, we evaluated the model on the test data and performed prediction computing the pre-mentioned evaluation metrics of interest on the test data set. Figure 5 displays the suggested scheme.



**Figure 5:** Figure depicts the basic scheme performing from hyperband tuner to determine the best configuration towards to the final evaluation of the model.

DL performance is controlled by various parameters and thus the optimization of the hyperparameters is not a trivial step. Here, we designed the tuner search space based on the available literature ([Sandhu et al. 2021](#), [Zingaretti et al. 2020](#)). The hyperparameters chosen to be optimized were: activation function (Relu, Tanh, Linear), number of neurons for the first layer in MLP and number of filters in CNN (16,38,64,128), number of hidden dense layers for MLP (0,1,2,3) while for CNN (1,2,3), number of neurons of hidden layers (2,4,8,16), numbers of optimizers

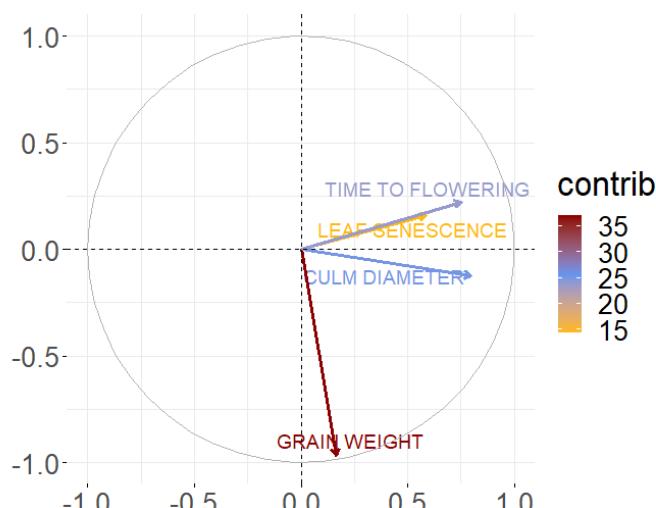
(Adam, RMSprop, SGD), dropout rate (0,0.05,0.1,0.15,0.2,0.25,0.3), L1 and L2 regularizers with optimized weight decay parameter (0.001, 0.01, 0.1). For the output layer we used one unit with linear as activation function for the quantitative traits whereas sigmoid was used for the binary traits as it is suggested ([Montesions-López et al. 2022](#)) For the hyperparameter optimization , 80% of training set was used and the remaining 20% validation data set was applied for inner testing. Training a DL network that can generalize well new data set is a challenging issue. A model with too little capacity cannot learn from the data, a problem known as underfitting, whereas a model with a large capacity can learn and fit too well to the training dataset results in overfitting. For avoiding and reducing the effects of these two phenomena there are techniques that can be adjusted to a DL network.

Here we used two regularization techniques such as L1 and L2 with a weight decay parameter. These techniques penalize the weight values of the network making values tend to zero and negative equal to 0 and avoid a parsimonious model. L1 adds “squared absolute value of magnitude” of coefficient as penalty term to the loss function while L2 adds “squared magnitude” of coefficient as penalty term to the loss function. We added L1 and L2 regularizers in the hidden layers. Additional to the regularization, dropout and early stopping were applied to reduce the effect of overfitting and underfitting on our models. A dropout layer was applied before the output layer. Our analysis was implemented using Tensor Flow 2.8.0 library with Keras 2.8.0 interface and Keras Tuner 1.1.2.

## RESULTS

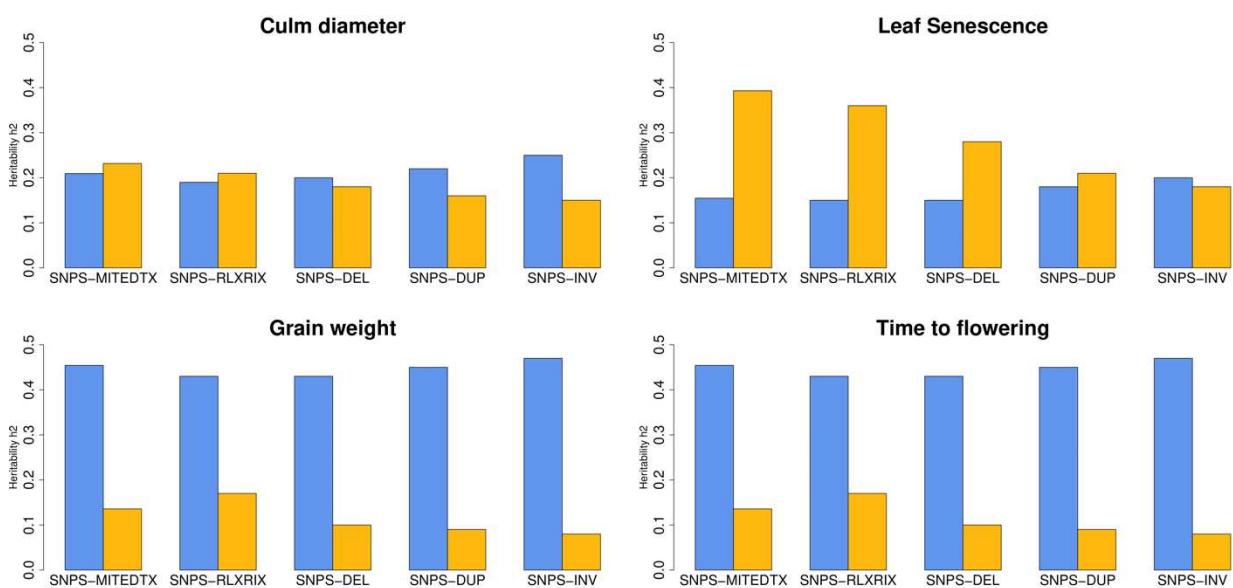
### Phenotypic Structure and Genetic Inference

We used PCA to determine the underlying structure of our data and the direction of the maximum variation when projected in a lower dimension space. Figure 6 shows the projections of variables of each trait onto the principal components. The length of the arrow is proportional to trait contribution, whereas the angle between arrows shows whether traits are correlated (pointed out in the same direction) or not. An analysis in two principal components showed that the first component depends on grain weight, which contributes the most to the total phenotypic variation. The main contributors to the second component in descending sequence are Time to flowering, Culm Diameter and Leaf Senescence.



**Figure 6:** PCA loadings of each trait for the two first standardized principal components. The different colors indicate the percentage of contribution to the PCAs displayed as “contrib”.

Genetic variance estimates were obtained for each trait (Figure 7). Particularly, we estimated the genetic variance explained by each SV marker set in comparison to SNPs in order to understand the relative importance of each set to determine the observed phenotype. Figure 7 shows that structural variants explain a significant fraction of genetic variance, larger than that explained by SNPs in the two binary traits, Culm diameter and Leaf Senescence. Among the different types of SVs, TE-related variants (MITE-DTX and RLX-RIX) explained more genetic variance than non-TE variants. Among the latter, deletions explained more variance than duplications and inversions.

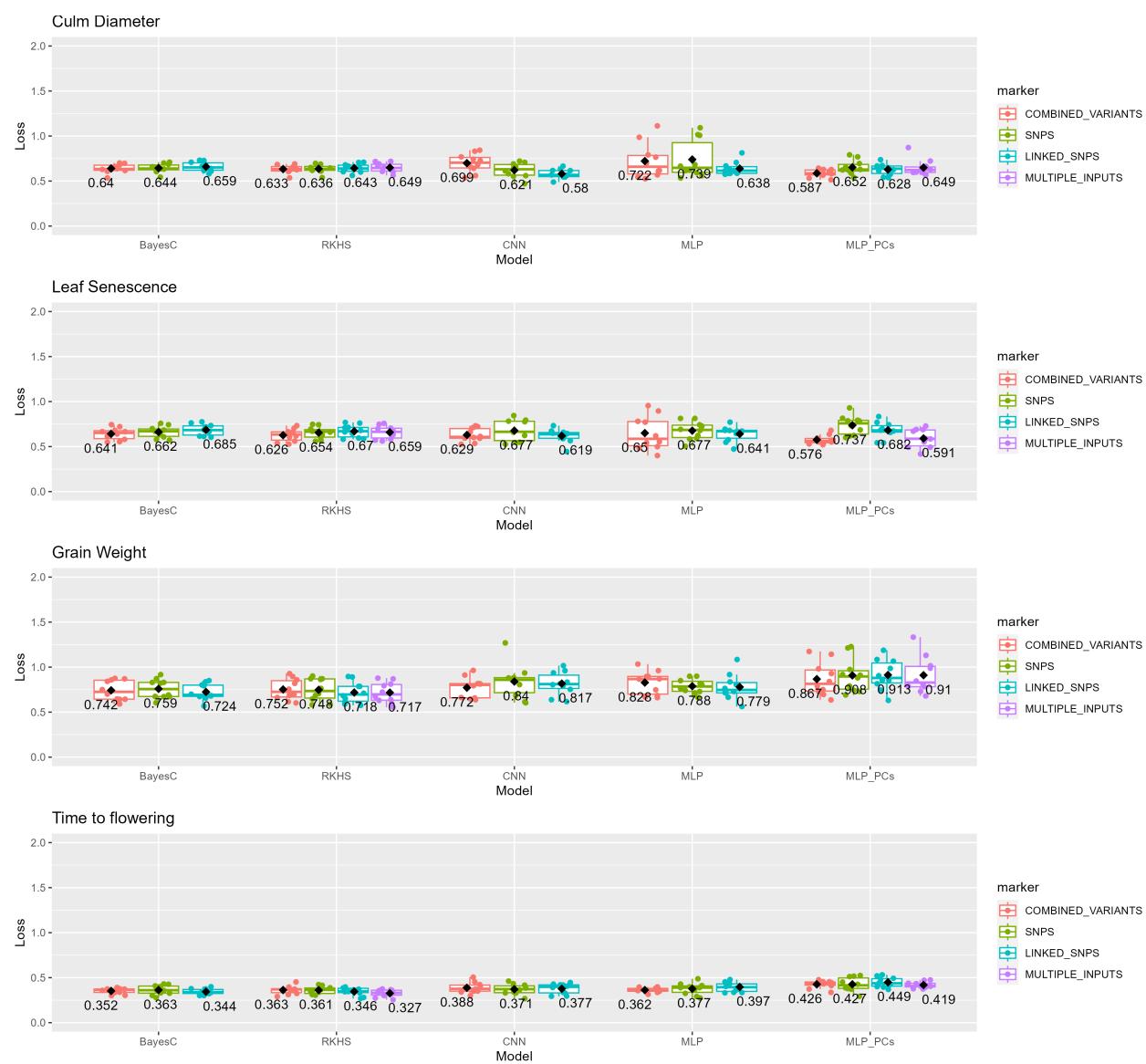


**Figure 7:** Means of posterior distributions of genetic variances explained by each marker set. Blue bars represent SNPs and yellow bars are the different kinds of SVs.

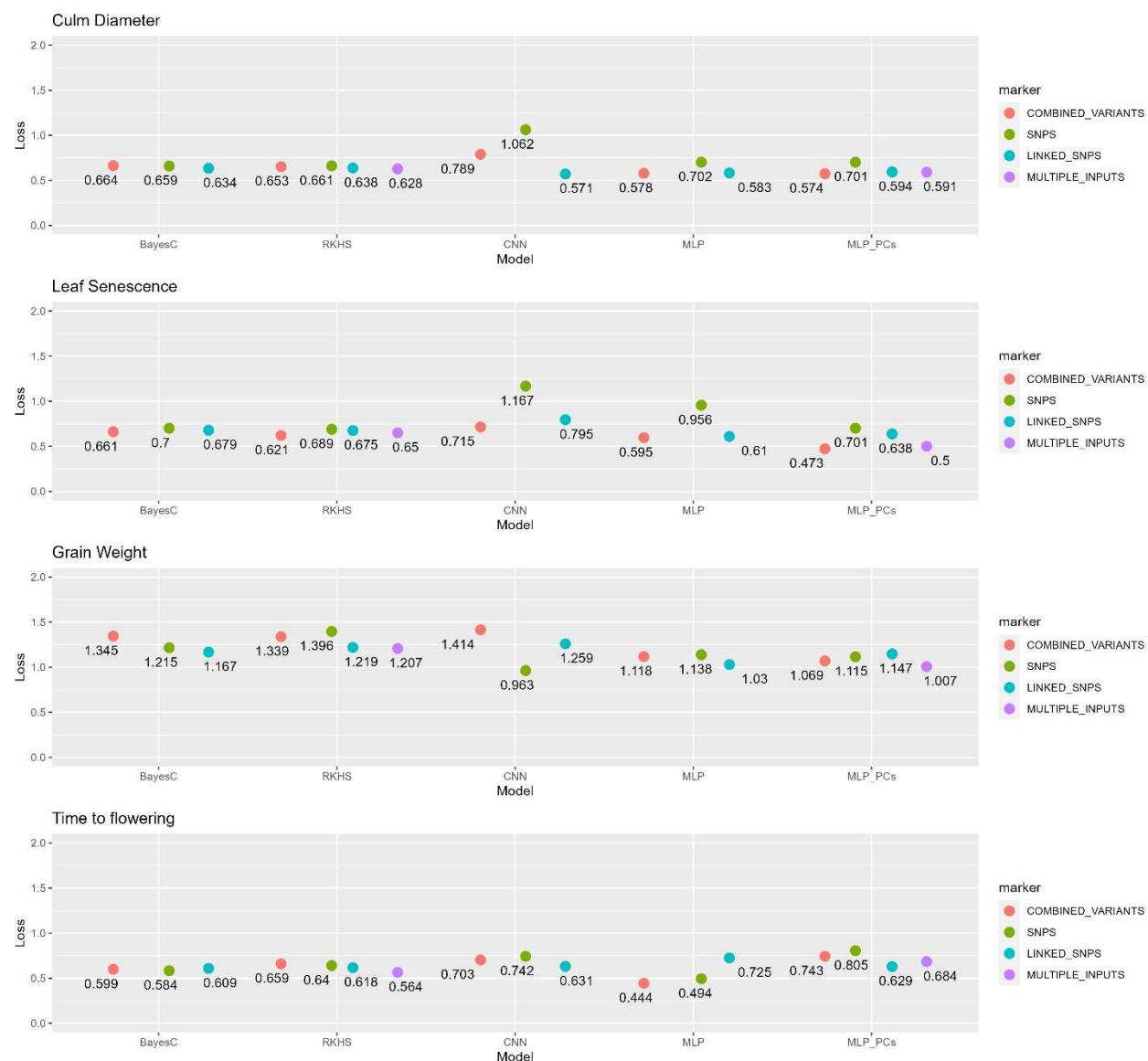
## Comparison of Model performances

The prediction ability of DL implementations was compared to those of Bayesian regression using RKHS and BayesC for each trait and under eleven cross validation sets. Particularly, we assessed prediction by following two different validation strategies, prediction using ten randomly selected training sets produced by a 10-fold cross validation strategy and prediction across populations. All the models were applied separately to each of the eleven validation scenarios (see Materials and Methods). Figure 8 shows the performance of each of the models under the 10-fold cross validation. We used binary-cross entropy as evaluation metric for binary traits and MSE for quantitative traits. Other metrics can also be applied (e.g Pearson’s correlation for quantitative traits, accuracy for binary traits) however here we measure the prediction ability in terms of loss of the model as previously reported ([Montesinos-López et al. 2023](#)). The highest prediction ability (minimum loss) for culm diameter was obtained using CNN and linked SNPs (loss = 0.58). For leaf senescence the best prediction value was reported using MLP with PCs computed by the

combined variants (Loss = 0.576). For the case of quantitative traits, Bayesian Regression models showed higher prediction accuracy values than those of DL models. Particularly, grain weight and time to flowering was better predicted under RKHS model using multiple inputs (Loss = 0.72 and 0.33, respectively).



**Figure 8:** Performance of each of the model-input combinations under the 10-fold cross validation strategy and four predicted phenotypes. Each model was applied separately to each of the ten partitions. Points represent the evaluation metric for each partition whereas the boxplot shows the distribution of the numerical values displaying the data quartiles. The value that appears in bold is the average value of each model. The y-axis shows the loss metric values which for binary traits is the binary-cross entropy and for quantitative traits the MSE.



**Figure 9:** Performance of each of the model-input combinations under the across population strategy. Points represent the evaluation metric. The y-axis shows the loss metric values which for binary traits is the binary-cross entropy and for quantitative traits the MSE.

In the independent prediction, phenotypes of all ADM and ARO accessions were predicted given the rest of the accessions. Figure 9 shows the prediction ability for across population strategy under eleven different models. Here, DL models outperformed Bayesian ones in all the traits. Particularly, culm diameter and leaf senescence were better predicted under the MLP\_PCs model and “Combined\_VARIANTS” marker strategy as in 10-partitions. However, grain weight and time to flowering showed the lowest loss values under CNN with SNPs (Loss = 0.96) and MLP with combined variants (Loss = 0.44) respectively. In general, time to flowering was better predicted compared to the rest traits in both cross-validation strategies. On average, prediction across populations was less accurate for the quantitative traits than in 10-fold scenarios as it was expected because of the more distantly related training and test data sets.

### Impact of marker selection on genomic prediction

Our results showed that phenotypic traits such as leaf senescence and time to flowering were better predicted using combined variants or multiple inputs. Also, using SNPs linked to SVs

exhibited an efficient prediction ability especially for culm diameter under both validation strategies. We observed that incorporating structural variation in a genomic prediction framework either combining with SNPs or generating the linked SNPs to these variants resulted in an improved prediction performance in near 90% of the studied cases against using only SNPs (Table 6).

**Table 6:** Minimum prediction loss and corresponding model with input strategy.

Traits	Validation Strategy	
	10-folds partitions	ARO/ADM accessions
<b>Culm diameter</b>	0.580 (CNN, Linked SNPs)	0.571 (CNN, Linked SNPs)
<b>Leaf senescence</b>	0.576 (MLP_PCs, Combined variants)	0.473 (MLP_PCs, Combined variants)
<b>Grain weight</b>	0.717 (RKHS, Multiple inputs)	0.963 (CNN SNPs)
<b>Time to flowering</b>	0.327 (RKHS, Multiple Inputs)	0.444 (MLP, Combined variants)

## Discussion

In this study we investigated whether combining structural and nucleotide genome-wide variation for genomic prediction can improve prediction ability for important agronomic traits in rice. Previous studies on plants have shown the association between structural variants and phenotypic traits ([Żmierko et al. 2014](#), [Shang et al. 2022](#), [Zhou et al. 2022](#)), and some have been demonstrated to be the causal variants for a diversity of phenotypes across major traits in plants ([Sutton et al. 2007](#), [Cook et al. 2012](#)). In example, late or early flowering on wheat depends on the increased copy number of Vrn-A1 and Ppd-B1 genes respectively ([Würschum et al. 2015](#)). In addition, plant height in wheat is associated with a specific tandem duplication ([Li et al. 2012](#)). The strong regulatory potential of SVs could be an explanation for the high performance of SVs in the prediction of phenotypic traits. However, the incorporation of SVs in breeding programs demands their genotyping to be automatized. This can be a complex task, as SVs are highly diverse and commercial SV genotyping assays do not exist. Our results show that linked SNPs can be effectively used to indirectly incorporate SVs in genomic prediction. We propose that SNPs linked to known SV variants (ie, resulting from recent pangenome studies) constitute a promising marker resource to be used for future genomic prediction analyses.

A second objective addressed is the evaluation of the performance of DL and Bayesian methods to predict agronomically important traits in rice. Culm diameter, leaf senescence and time to flowering are correlated (Figure 6), whereas grain weight is uncorrelated to them. Traits such as time to flowering and grain weight are polygenic, controlled by many QTLs of large effects ([Begum et al. 2015](#), [Xu et al. 2015](#), [Chen et al. 2021](#)). Studies in culm diameter have shown that it is controlled by at least twelve QTLs associated with lodging resistance in dry direct-seeded rice ([Yadav et al. 2017](#)). In addition, delayed leaf senescence or stay-green is associated to forty-six QTLs that made up the genetic basis of this important trait in rice ([Jiang et al. 2004](#)). Genomic prediction of traits such as time to flowering was quite accurate with the loss metric reported being the lowest values across all the study (average MSE value equal to 0.33, Table 6). For leaf senescence the GP ability was lower than that in time to flowering yet accurate. It is worth mentioning that, in the ARO/ADM validation strategy the prediction ability of leaf senescence and time to flowering was improved by DL against the best values of Bayesian models by 24% and 21% respectively (Figure 9). Since the genetic relatedness of the accessions used for training increases

prediction accuracy, it is interesting that DL models outperform the Bayesian ones in both binary traits for genetically distant-lines. (Table 6).

Increasing the prediction accuracies of traits in rice breeding is challenging but at the same time of high importance, taking into consideration the increasing environmental constraints that limit world production. New methods attempt to improve prediction of agronomic traits promising lower computational cost and better results. DL is a state-of-the-art method applied in many different fields, and many recent studies have started to compare DL with standard linear models for genomic prediction ([González-Recio et al. 2014](#); [Ma et al. 2017](#); [Bellot et al. 2018](#); [Montesinos-López et al. 2018](#), [Zingaretti et al. 2020](#), [Sandhu et al. 2021](#), [Montesinos-López et al. 2019](#), [Montesinos-López et al. 2023](#)). Here, we studied the performance of DL models for predicting complex traits in rice comparing them to Bayesian regression methods under different input strategies and scenarios. Overall, our results showed that DL can increase prediction accuracy compared to Bayesian methods in 75% of the implementations. Across DL architectures, MLP and CNN were the optimal choices in the same number of cases depending on the trait and training population. This observation shows that there is not a clear winner, as evidenced by contrasting findings in the literature, where MLP outperforms CNN according to ([Sandhu et al. 2021](#)), whereas ([Bellot et al. 2018](#), [Zingaretti et al. 2020](#)) report the opposite trend. For the case of Bayesian regression models, RKHS clearly outperformed BayesC.

Another critical and challenging issue in DL models is the optimization of hyperparameters, mainly due to the high computational cost. The tuning of the hyperparameters for each trait depends on the genetic basis and architecture of the trait. As we show in Supplementary Tables 1-4, different combinations of hyperparameters were selected for the various traits as the prediction ability is highly associated with the interaction of these factors ([Bellot et al. 2018](#), [Montesinos-López et al. 2018](#)). We observed that Tanh was the most useful activation function in quantitative traits being selecting in 75% of the cases (6/8) whereas in binary traits, Relu function was the optimal choice in 63% of the cases (5/8). Moreover, Adam optimizer was the most frequently chosen in binary traits during the hypertuning with 63%. Nevertheless, RMSprop was the optimal option with percentage of 50% in quantitative traits. DL models can capture interactions of large orders because of the presence of hidden layers ([Goodfellow et al. 2016](#), [LeCun et al. 2015](#)). However, RKHS models are also able to capture complex interaction patterns. This ability of both methods can be reflected in our results demonstrating that both can capture complex interactions.

The third aim of this work is to study the impact of various input strategies on the prediction results. It is commonly believed that GP requires a large marker set to be used for an efficient prediction. However, our current results and some of related works ([Vourlaki et al. 2022](#), [Bellot et al. 2018](#)) support that GP models can be effective even with a smaller dataset of markers. However, the optimal marker size can be related to the studied trait ([Sandhu et al. 2021](#)). We also observed that the best input strategy is affected by the chosen phenotypic trait and the training set in some cases (Table 6). Note that MLP models using PCs as input strategy proved beneficial in 66.7% of the cases with MLP as best model. In any case, the different input strategies that we followed indicated that the accommodation of subsets of the markers in GP framework can be equal or even more informative than using the whole marker sets (compared to our results in [Vourlaki et al. 2022](#)).

Finally, we would like to mention the challenges and limitations of DL models. Firstly, DL models do not provide clear insights into the genetic architecture of the traits, nor do they give information about the effects of specific markers in the studied traits. Different hyperparameters act on different parts of the data, making it hard to interpret the biological significance and

importance of each marker in the model ([Bellot et al. 2018](#), [Cuevas et al. 2019](#)). Also, the high computational cost of training models is a significant drawback, especially when multiple hyperparameters must be optimized for each trait separately ([Gulli and Pal et al. 2017](#)). The outperformance of DL over linear models is not always the case. The prediction ability depends on the studied traits and can be influenced by many factors. There is not a single algorithm that performs better in all species and traits ([Perez-Enciso and Zingaretti, 2019](#)) since its performance depends on various factors. Nevertheless, even though the advantage of DL networks against linear methods has not been established yet, their incorporation into plant breeding can be important to improve genetic merit for complex traits.

## Acknowledgements

We acknowledge Josep M. Casacuberta for his helpful advice on the manuscript and Alejandro Navarro Martínez for his contribution to the initial work.

## Funding

Ministerio de Ciencia e Innovación grant PID2019-108829RB-I00 to MPE and RC.

Ministerio de Ciencia e Innovación Postdoctoral Fellowship IJC2020-045949-I to RC.

Programa Severo Ochoa para Centros de Excelencia en I+D" 2020-2023 (CEX2019-000902-S)

## References

Abadi, Mart & others. (2016). Tensorflow: A system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16) (pp. 265–283).

Akakpo R., Carpentier M.C., le Hsing Y., Panaud O. (2020) The impact of transposable elements on the structure, evolution, and function of the rice genome. *New Phytologist* 226:44–49. <https://doi.org/10.1111/nph.16356>.

Alexandrov N, Tai S, Wang W, Mansueto L, Palis K, Fuentes RR, Ulat VJ, Chebotarov D, Zhang G, Li Z, Mauleon R, Hamilton RS, McNally KL. 2015 SNP-Seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Res.* 43(Database issue):D1023-7. doi: 10.1093/nar/gku1039.

Alonge M., Wang X., Benoit M., Soyk S., Pereira L., Zhang L., Suresh H., Ramakrishnan S., Maumus F., Ciren D., Levy Y., Harel T.H., Shalev-Schlosser G., Amsellem Z., Razifard H., Caicedo A.L., Tieman D.M., Klee H., Kirsche M., Aganezov S., Ranallo-Benavidez T.R., Lemmon Z.H., Kim J., Robitaille G., Kramer M., Goodwin S., McCombie W.R., Hutton S., Van Eck J., Gillis J., Eshed Y., Sedlazeck F.J., van der Knaap E., Schatz M.C., Lippman Z.B. (2020) Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell.* 182(1):145-161.e23. doi: 10.1016/j.cell.2020.05.021.

Amadeu R.R., Cellon C., Olmstead J.W., et al. (2016) AGHmatrix: R Package to Construct Relationship Matrices for Autotetraploid and Diploid Species: A Blueberry Example. *The Plant Genome* 9: <https://doi.org/10.3835/plantgenome2016.01.0009>.

Begum H., Spindel J.E., Lalusin A., et al (2015) Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (*Oryza sativa*). *PLoS One.* 2015 Mar 18;10(3):e0119873. doi: 10.1371/journal.pone.0119873.

Bellot P., de Los Campos G., Pérez-Enciso M. (2018). Can deep learning improve genomic prediction of complex human traits? *Genetics* 210, 809–819. doi: 10.1534/genetics.118.301298.

Browning BL, Zhou Y, Browning SR (2018) A one-penny imputed genome from next-generation reference panels. *Am J Hum Genet* 103:338–348. <https://doi.org/10.1016/j.ajhg.2018.07.015>.

Carpentier M.C., Manfroi E., Wei F.J., et al (2019) Retrotranspositional landscape of Asian rice revealed by 3000 genomes. *Nature Communications* 10(1):24. doi: 10.1038/s41467-018-07974-5.

Castanera R., Vendrell-Mir P., Bardil A., et al (2021) Amplification dynamics of miniature inverted-repeat transposable elements and their impact on rice trait variability. *Plant Journal* 107:118–135. <https://doi.org/10.1111/tpj.15277>.

Chan M., Scarafoni D., Duarte R., Thornton J., Skelly L. (2018). “Learning network architectures of deep CNNs under resource constraints,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. (IEEE), 1784–1791. doi: 10.1109/CVPRW.2018.00222.

Chang CC, Chow CC, Tellier LCAM et al (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*. <https://doi.org/10.1186/s13742-015-0047-8>.

Chen K., Łyskowski A., Jaremko Ł., Jaremko M. (2021) Genetic and Molecular Factors Determining Grain Weight in Rice. *Frontiers in Plant Science* 12. doi: 10.3389/fpls.2021.605799.

Chigira K., Yamasaki M., Adachi S., Nagano A.J., Ookawa T. (2023) Identification of Novel Quantitative Trait Loci for Culm Thickness of Rice Derived from Strong-Culm Landrace in Japan, Omachi. *Rice (N Y)*. 16(1):4. doi: 10.1186/s12284-023-00621-8.

Cook D. E., Lee T. G., Guo X., Melito S., Wang K., Bayless A. M., et al. (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science* 338, 1206–1209. doi: 10.1126/science.1228746.

Cuevas J., Montesinos-López O., Juliana P., Guzmán C., Pérez-Rodríguez P., González-Bucio J., Burgueño J., Montesinos-López A., Crossa J. (2019) Deep Kernel for Genomic and Near Infrared Predictions in Multi-environment Breeding Trials. *G3 (Bethesda)*. 9(9):2913–2924. doi: 10.1534/g3.119.400493.

Domínguez M., Dugas E., Benchouaia M., et al (2020) The impact of transposable elements on tomato diversity. *Nature Communications* 11: <https://doi.org/10.1038/s41467-020-17874-2>.

Dubin MJ, Mittelsten Scheid O, Becker C. (2018) Transposons: a blessing curse. *Curr Opin Plant Biol.* 42:23–29. doi: 10.1016/j.pbi.2018.01.003.

Ehret A., Hochstuhl D., Krattenmacher N., Tetens J., Klein M.S., Gronwald W., Thaller G. (2015) Short communication: Use of genomic and metabolic information as well as milk performance records for prediction of subclinical ketosis risk via artificial neural networks. *J Dairy Sci.* 98(1):322–9. doi: 10.3168/jds.2014-8602.

FAO (2023) Crop Prospects and Food Situation – Quarterly Global Report No. 2, July 2023. Rome. <https://doi.org/10.4060/cc6806en>

Fox E. A., Wright A. E., Fumagalli M., Vieira F. G. (2019) *ngsLD*: evaluating linkage disequilibrium using genotype likelihoods, *Bioinformatics*, Volume 35, Issue 19, Pages 3855–3856, <https://doi.org/10.1093/bioinformatics/btz200>

Gianola D., Fernando R.L., Stella A. (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*. 173(3):1761–76. doi: 10.1534/genetics.105.049510.

Gianola D., Okut H., Weigel K.A., Rosa G.J. (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet.* 12:87. doi: 10.1186/1471-2156-12-87.

González-Recio O., Rosa G. J. M., Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livest. Sci.* 166, 217–231. doi: 10.1016/j.livsci.2014.05.036.

Goodfellow I. J., Bengio Y., and Courville, A. C. (2016). Deep learning. Adaptive computation and machine learning. Cambridge: MIT Press.

Gulli A., and Pal S. (2017). Deep learning with Keras. Birmingham: Packt Publishing Ltd.

Habier D., Fernando R.L., Kizilkaya K., Garrick D.J. (2011) Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*. 12:186. doi: 10.1186/1471-2105-12-186.

Jiang G.H., He Y.Q., Xu C.G., Li X.H., Zhang Q. (2004) The genetic basis of stay-green in rice analyzed in a population of doubled haploid lines derived from an indica by japonica cross. *Theor. Appl. Genet.*, 108, 688–698.

Jighly A., Lin Z., Pembleton L.W., et al. (2019) Boosting Genetic Gain in Allogamous Crops via Speed Breeding and Genomic Selection. *Frontiers in Plant Science* 10: <https://doi.org/10.3389/fpls.2019.01364>.

Kamal N.M., Alnor Gorafi Y.S., Abdelrahman M., Abdellatef E., Tsujimoto H. (2019) Stay-Green Trait: A Prospective Approach for Yield Potential, and Drought and Heat Stress Adaptation in Globally Important Cereals. *Int J Mol Sci.* 20(23):5837. doi: 10.3390/ijms20235837.

Krishnappa G., Savadi S., Tyagi B.S., et al (2021) Integrated genomic selection for rapid improvement of crops. *Genomics* 113:1070–1086. doi: 10.1016/j.ygeno.2021.02.007.

LeCun Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539.

Lee S., Masclaux-Daubresse C. (2021) Current Understanding of Leaf Senescence in Rice. *Int J Mol Sci.* 22(9):4515. doi: 10.3390/ijms22094515.

Li L., Jamieson K., DeSalvo G., Rostamizadeh A. and Talwalkar A. (2018) Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research* 18 1-52. <http://jmlr.org/papers/v18/16-558.html>.

Li Y., Xiao J., Wu J., Duan J., Liu Y., Ye X., Zhang X., Guo X., Gu Y., Zhang L., Jia J., Kong X. (2012) A tandem segmental duplication (TSD) in green revolution gene Rht-D1b region underlies plant height variation. *New Phytol.* 196(1):282-291. doi: 10.1111/j.1469-8137.2012.04243.x.

Li Z., Fu B.Y., Gao Y.M., et al. (2014) The 3,000 rice genomes project. *Gigascience* 3:8. doi: 10.1186/2047-217X-3-8.

Liu Y., Du H., Li P., Shen Y., Peng H., Liu S., Zhou GA., Zhang H., Liu Z., Shi M., Huang X., Li Y., Zhang M., Wang Z., Zhu B., Han B., Liang C., Tian Z. (2020) Pan-Genome of Wild and Cultivated Soybeans. *Cell.* 182(1):162-176.e13. doi: 10.1016/j.cell.2020.05.023.

Livieris I.E., Dafnis S.D., Papadopoulos G.K., Kalivas D.P. (2020) A Multiple-Input Neural Network Model for Predicting Cotton Production Quantity: A case study, *Algorithms*, 13(11), 273. <https://doi.org/10.3390/a13110273>.

Lye Z.N., Purugganan MD. (2019) Copy Number Variation in Domestication. *Trends Plant Sci.* 24(4):352-365. doi: 10.1016/j.tplants.2019.01.003.

Ma W., Qiu Z., Song J., Cheng Q., Ma C. (2017) DeepGS: predicting phenotypes from genotypes using deep learning. *bioRxiv* 241414. doi: 10.1101/241414.

Meuwissen T. H. E., Hayes B. J., and Goddard, M. E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157, 1819–1829. <https://doi.org/10.1093/genetics/157.4.1819>.

Min S., Lee B., Yoon S. (2017). Deep learning in bioinformatics. *Brief. Bioinform.* 18, 851–869. doi: 10.1093/bib/bbw068.

Montesinos-López, O.A., Montesinos López, A., Crossa, J. (2022). Artificial Neural Networks and Deep Learning for Genomic Prediction of Binary, Ordinal, and Mixed Outcomes. In: *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer, Cham. [https://doi.org/10.1007/978-3-030-89010-0\\_12](https://doi.org/10.1007/978-3-030-89010-0_12).

Montesinos-López A., Montesinos-López O. A., Hernández-Suárez C. M., Gianola D., Crossa J. (2018). Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3. Genes Genomes Genet.* 8 (12), 3813–3828. doi: 10.1534/g3.118.200740.

Montesinos-López O. A., Martín-Vallejo J., Crossa J., Gianola D., Hernández-Suárez C.M., Montesinos-López A., Juliana P., Singh R. A. (2019) Benchmarking Between Deep Learning, Support Vector Machine and Bayesian Threshold Best Linear Unbiased Prediction for Predicting Ordinal Traits in Plant Breeding. *G3 (Bethesda)*. 9(2):601-618. doi: 10.1534/g3.118.200998.

Montesinos-López O.A., Montesinos-López A., Mosqueda-González. BA., Bentley AR., Lillemo M., Varshney R.K., Crossa J. (2023) A New Deep Learning Calibration Method Enhances Genome-Based Prediction of Continuous Crop Traits. *Front Genet.* 12:798840. doi: 10.3389/fgene.2021.798840.

Montesinos-López O.A., Montesinos-López A., Pérez-Rodríguez P., Barrón-López J.A., Martini J.W.R., Fajardo-Flores S.B., Gaytan-Lugo L.S., Santana-Mancilla P.C., Crossa J. (2021) A review of deep learning applications for genomic selection. *BMC Genomics.* 22(1):19. doi: 10.1186/s12864-020-07319-x.

Müller B.S.F., Neves L.G., de Almeida Filho J.E., Resende M.F.R. Jr, Muñoz P.R., Dos Santos P.E.T., Filho E.P., Kirst M., Grattapaglia D. (2017) Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of Eucalyptus. *BMC Genomics.* 18(1):524. doi: 10.1186/s12864-017-3920-2.

O’Malley, Tom, et al. (2019) Keras Tuner. Available online: [github.com/keras-team/kerastuner](https://github.com/keras-team/kerastuner).

Pattanayak S. (2017). Unsupervised Learning with Restricted Boltzmann Machines and Auto-encoders. *Pro Deep Learning with TensorFlow* (Berkeley, CA: Apress), 279–343. doi: 10.1007/978-1-4842-3096-1\_5.

Pérez-Enciso M., Rincón J.C., Legarra A. (2015) Sequence- vs. chip-assisted genomic selection: Accurate biological information is advised. *Genetics Selection Evolution* 47: <https://doi.org/10.1186/s12711-015-0117-5>.

Pérez-Enciso M., Zingaretti L. (2019). A guide for using deep learning for complex trait genomic prediction. *Genes* (Basel). 10, 553. doi: 10.3390/genes10070553.

Pérez P., de los Campos G. Genome-wide regression and prediction with the BGLR statistical package (2014). *Genetics*. 198(2):483–95. doi: 10.1534/genetics.114.164442.

Pérez-Rodríguez P., Gianola D., González-Camacho J.M., Crossa J., Manès Y., Dreisigacker S. (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3 (Bethesda)* 2(12):1595–605. doi: 10.1534/g3.112.003665.

Purcell S., Neale B., Todd-Brown K. et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575. <https://doi.org/10.1086/519795>.

Qin P., Lu H., Du H., Wang H., Chen W., Chen Z., He Q., Ou S., Zhang H., Li X., Li X., Li Y., Liao Y., Gao Q., Tu B., Yuan H., Ma B., Wang Y., Qian Y., Fan S., Li W., Wang J., He M., Yin J., Li T., Jiang N., Chen X., Liang C., Li S. (2021) Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*.184(13):3542-3558.e16. doi: 10.1016/j.cell.2021.04.046.

Sandhu K.S., Lozada D.N., Zhang Z., Pumphrey M.O. and Carter A.H. (2021) Deep Learning for Predicting Complex Traits in Spring Wheat Breeding Program. *Front. PlantSci.* 11: 613325. doi: 10.3389/fpls.2020.613325.

Seuret M., Alberti M., Liwicki M., Ingold R. (2017) PCA-Initialized Deep Neural Networks Applied to Document Image Analysis. *IAPR International Conference on Document Analysis and Recognition (ICDAR)* 01: 877–882. <https://doi.ieeecomputersociety.org/10.1109/ICDAR.2017.148>.

Shang L., Li X., He H., Yuan Q., Song Y., Wei Z., Lin H., Hu M., Zhao F., Zhang C., Li Y., Gao H., Wang T., Liu X., Zhang H., Zhang Y., Cao S., Yu X., Zhang B., Zhang Y., Tan Y., Qin M., Ai C., Yang Y., Zhang B., Hu Z., Wang H., Lv Y., Wang Y., Ma J., Wang Q., Lu H., Wu Z., Liu S., Sun Z., Zhang H., Guo L., Li Z., Zhou Y., Li J., Zhu Z., Xiong G., Ruan J., Qian Q. (2022) A super pan-genomic landscape of rice. *Cell Res.* 878-896. doi: 10.1038/s41422-022-00685-z.

Sun C., Hu Z., Zheng T., et al. (2017) RPAN: Rice pan-genome browser for ~3000 rice genomes. *Nucleic Acids Research* 45: <https://doi.org/10.1093/nar/gkw958>.

Sutton T., Baumann U., Hayes J., Collins N. C., Shi B.J., Schnurbusch T., et al. (2007). Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science* 318, 1446–1449. doi: 10.1126/science.1146853.

Tessema B.B., Liu H., Sørensen A.C., et al (2020) Strategies Using Genomic Selection to Increase Genetic Gain in Breeding Programs for Wheat. *Frontiers in Genetics* 11: <https://doi.org/10.3389/fgene.2020.578123>.

Tibshirani R. (2011) Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 73:273–282. <https://doi.org/10.1111/j.1467-9868.2011.00771.x>.

VanRaden P.M. (2008) Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91:4414–4423. <https://doi.org/10.3168/jds.2007-0980>.

Vourlaki I.T., Castanera R., Ramos-Onsins S.E., Casacuberta J.M., Pérez-Enciso M. (2022) Transposable element polymorphisms improve prediction of complex agronomic traits in rice. *Theor Appl Genet.* Sep;135(9):3211-3222. doi: 10.1007/s00122-022-04180-2.

Widder D. V. (1954). The convolution transforms. *Bull. Am. Math. Soc.* 60, 444–456. doi: 10.1090/S0002-9904-1954-09828-2.

Würschum T., Boeven P.H., Langer S.M., Longin C.F., Leiser W.L. (2015) Multiply to conquer copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. *BMC Genet* 16: 96. doi:10.1186/s12863-015-0258-0.

Xiong C., Zheng J., Xu L., Cen C., Zheng R., Li Y. (2021) Multiple-Input Convolutional Neural Network Model for Large-Scale Seismic Damage Assessment of Reinforced Concrete Frame Buildings. *Appl. Sci.*, 11, 8258. <https://doi.org/10.3390/app11178258>.

Xu F., Sun X., Chen Y., et al. (2015) Rapid identification of major QTLs associated with rice grain weight and their utilization. *PLoS ONE*. 10(3):e0122206. <https://doi.org/10.1371/journal.pone.0122206>.

Xu Y., Liu X., Fu J., Wang H., Wang J., Huang C., Prasanna B.M., Olsen M.S., Wang G., Zhang A. (2020) Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. *Plant Commun.* 1(1):100005. doi: 10.1016/j.xplc.2019.100005.

Xu Y., Ma K., Zhao Y. et al. (2021) Genomic selection: a breakthrough technology in rice breeding. *Crop Journal* 9:669–677. <https://doi.org/10.1016/j.cj.2021.03.008>.

Yadav S., Singh U.M., Naik S.M., Venkateshwarlu C., Ramayya P.J., Raman K.A., Sandhu N., Kumar A. (2017) Molecular Mapping of QTLs Associated with Lodging Resistance in Dry Direct-Seeded Rice (*Oryza sativa* L.). *Front Plant Sci.* Aug 21;8:1431. doi: 10.3389/fpls.2017.01431.

Young S.R., Rose D.C., Karnowski T.P., Lim S.H., Patton R.M. (2015) Optimizing deep learning hyper-parameters through an evolutionary algorithm. *Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments - MLHPC '15*, (New York, New York, USA: ACM Press), 1–5. doi: 10.1145/2834892.2834896.

Zingaretti L. M., Gezan S. A., Ferrão L. F. V., Osorio L. F., Monfort A., Muñoz P. R., et al. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Front. Plant Sci.* 11:25. doi: 10.3389/fpls.2020.00025.

Żmieleńko A., Samelak A., Kozłowski P., Figlerowicz M. (2014) Copy number polymorphism in plant genomes. *Theor Appl Genet.* 127(1):1-18. doi:10.1007/s00122-013-2177-7.

Zhou Y., Zhang Z., Bao Z., Li H., Lyu Y., Zan Y., Wu Y., Cheng L., Fang Y., Wu K., Zhang J., Lyu H., Lin T., Gao Q., Saha S., Mueller L., Fei Z., Städler T., Xu S., Zhang Z., Speed D., Huang S. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature.* 606(7914):527-534. doi: 10.1038/s41586-022-04808-9.