Bioinformatics Advances, 2023, vbad028
https://doi.org/10.1093/bioadv/vbad028
Advance Access Publication Date: 9 March 2023
Original Article

OXFORD

ISCB-LA Paper

Predicting phenotypes from novel genomic markers using deep learning

Shivani Sehrawat , Keyhan Najafian and Lingling Jin *

Department of Computer Science, University of Saskatchewan, Saskatoon, SK S7N 5C9, Canada

*To whom correspondence should be addressed.

Associate Editor: Thomas Lengauer

Received on December 19, 2022; revised on February 8, 2023; editorial decision on February 23, 2023; accepted on March 8, 2023

Abstract

Summary: Genomic selection (GS) models use single nucleotide polymorphism (SNP) markers to predict phenotypes. However, these predictive models face challenges due to the high dimensionality of genome-wide SNP marker data. Thanks to recent breakthroughs in DNA sequencing and decreased sequencing cost, the study of novel genomic variants such as structural variations (SVs) and transposable elements (TEs) become increasingly prevalent. In this article, we develop a deep convolutional neural network model, `NovGMDeep`, to predict phenotypes using SVs and TEs markers for GS. The proposed model is trained and tested on samples of *Arabidopsis thaliana* and *Oryza sativa* using *k*-fold cross-validation. The prediction accuracy is evaluated using Pearson's Correlation Coefficient (PCC), mean absolute error (MAE) and SD of MAE. The predicted results showed higher correlation when the model is trained with SVs and TEs than with SNPs. `NovGMDeep` also has higher prediction accuracy when comparing with conventional statistical models. This work sheds light on the unappreciated function of SVs and TEs in genotype-to-phenotype associations, as well as their extensive significance and value in crop development.
Contact: lingling.jin@cs.usask.ca

1 Introduction

The genotype of an organism is described by its entire genetic composition. Genotype can also refer to the set of alleles contained within the genome. The phenotype, or observable characteristics, are influenced by the genotype's expression in the cell. Generally, three variables influence a phenotype: the most effective are genes; the inherited epigenetic factors and acquired environmental factors are the other two (Wong *et al.*, 2005).

Because of the success of genome-wide association studies (GWAS), there is a growing interest in using genotype data to predict complex traits like diseases. According to some recent GWAS analyses, only a few common single nucleotide polymorphisms (SNPs) are involved in most diseases and these associated SNPs can only explain a small percentage of the disease susceptibility (Li *et al.*, 2018). Early plant genomic investigations were hampered by technological limitations and a lack of high-quality reference genome assemblies, which prohibited a detailed analysis of both structural variations (SVs) and TEs in plants. Recent improvements in genomic technology, notably long-read sequencing, offer the generation of high-quality plant genome and pangenome assemblies, as well as exposure to a diverse set of SVs for evaluating their possible significance in plant phenotypic diversity (Yuan *et al.*, 2021). Some examples regarding the impacts of SVs in plants regarding gene regulation toward environmental fitness are: SVs may contribute to biotic (Dolatabadian *et al.*, 2017) and abiotic resistance (Gabur *et al.*, 2019), heterosis in maize (Lai *et al.*, 2010), and increased chlorophyll content in *Brassica napus* (Qian *et al.*, 2016). According to a PanSV genome analysis of 100 tomato accessions, SVs altered gene dosage and expression levels, which led to quantitative trait variations in fruit flavor, size and yield (Alonge *et al.*, 2020). Similarly, the transposition of TEs, which are common in most eukaryote genomes, accounts for a considerable amount of genomic variation. As a result, transposable element (TE)-derived molecular markers are useful resources for unraveling genomic variations in both plant and animal (Roy *et al.*, 2015). This work aims to shed light on the unappreciated function of SVs and TEs in genotype-to-phenotype associations, as well as their extensive significance and value in crop development.

Predicting crop phenotypes is an effective step in explaining crop behavior using insights from genome-wide markers. Genomic selection (GS) is a potent method to enhance quantitative traits since it uses genomic-estimated breeding values of individuals generated from genome-wide markers to select candidates for the next breeding cycle (Meuwissen, 2009).

The importance of predicting complex traits using large volumes of genomic data has led to the development of novel machine learning models. The conventional genomic prediction models are Ridge Regression Best Linear Unbiased (rrBLUP) (Endelman, 2011) and Genomic Best Linear Unbiased Prediction (gBLUP) (Clark and Werf, 2013). However, these conventional predictors typically assume that the genotype random effects follow a normal distribution, and

each genotype's contribution to phenotypes is considered as an independent attribute. It is, however, unknown in practice how genotype effects behave and they may not follow a strict distribution. Moreover, these conventional statistical models do not consider nonlinearity between the variables, as SNPs may also interact with other SNPs to cause complex diseases or traits as a result of epistasis. While these models face challenges due to the high dimensionality of genome-wide marker data and interactions between alleles, GS can benefit from Deep Learning (DL), which provides novel approaches to process noisy data (LeCun *et al.*, 2015) and handle nonlinearity (Islam *et al.*, 2018).

For DL models, studies have shown that these methods base their predictions on overall genetic relatedness, rather than on the effects of particular markers (Ubbens *et al.*, 2021). Moreover, DL models offer feature engineering and extraction capabilities (Zheng and Casari, 2018), and therefore have the potential to find hidden patterns in large-scale datasets (Min *et al.*, 2017). As a simplest DL model, in a multi-layer convolutional neural network (CNN), each layer consisting of many neurons, represents complicated relationships among the large datasets. The different neurons in each layer, receive information from the lower hierarchical layers, which is triggered by predefined activation functions. Activation functions bring nonlinearity to the output of each layer to determine the input of the next layer (Montesinos-López *et al.*, 2021). DL-based models have shown promising results in wide variety of tasks such as object detection (Zhao *et al.*, 2019), speech recognition (Deng *et al.*, 2013), natural language processing (Young *et al.*, 2018), etc.

1.1 Existing methods in genomic selection

Broadly speaking, GS methods can be classified into conventional statistical models and DL methods. In this section, we briefly summarize a few models in each category. These models are compared with our proposed `NovGMDeep` model in Section 3 to demonstrate the prediction of phenotypes using SV, TE and SNP markers.

One of the most widely used statistical models for GS is rrBLUP (Endelman, 2011). The model uses the linear regression algorithm which takes the genotype matrix as the input and predicts the phenotype vector. The ridge regression technique is used to estimate the effects of all genotypic markers, and these effects follow a normal distribution with a nonzero variance.

The gBLUP model (Clark and Werf, 2013) is another statistical model for estimating an individual's genetic performance based on its genomic associations. A genomic relationship matrix is used in gBLUP to represent genomic markers. To make predictions, the genotype matrix specifies covariance between individuals based on observed similarity at the genomic level.

The DeepGS model is developed using a deep CNN with architecture including a one-dimensional (1D)-convolutional layer, a 1D-max-pooling layer, combination of a few dropouts and fully connected layers (Ma *et al.*, 2018). The final output represents the predicted phenotypic values for the analyzed individual markers. The DeepGS model is trained using 10-fold cross-validation. They used the backpropagation algorithm (Rumelhart *et al.*, 1986) with the learning rate of 0.01, the momentum of 0.5 and the weight decay set to 0.00001 to optimize the parameters of the model. They trained the model for 6000 epochs. The DeepGS model utilizes hidden variables to collectively represent features in the genome-wide markers while making predictions.

The quantitative phenotypic prediction problem was treated by G2PDeep as a regression problem (Zeng *et al.*, 2021). Data on zygosity and SNPs are fed into the model. The model is made up of a dual-CNN layer and a fully connected neural network. The encoded genotypes are transferred by dual-CNN, which has two concurrent series of CNN layers with the kernel sizes of 4 and 20, followed by another CNN layer of size 4 to enhance marker representation. Both CNN streams are aggregated and fed into the following CNN layer to complete the feature extractor section of the model. In the end, fully connected layers with 512 and 1 neurons serve as regression blocks for predicting phenotypes in the model.

1.2 Genomic selection using novel genomic markers

The above-discussed methods only consider the SNP variants as genomic markers in predicting the phenotypes. For standard linear models, the number of genome-wide SNP features largely exceeds the sample size, leading to overfitting. To mitigate these limitations, instead of SNPs, we utilize SVs and TEs to represent genomic variations in two separate experiments and explore their effectiveness in GS. We have provided an overview of SVs and TEs in Section 1 and explained their importance in phenotypic prediction in plants.

Our study illustrates the need to look beyond SNPs to understand evolutionary processes and how SVs and TEs can help us understand variation within species or during early divergence. The genotype-to-phenotype predictions identified using SVs and TEs will be useful to investigate *Arabidopsis thaliana* and *Oryza sativa* evolution and trait architecture. We develop `NovGMDeep`, a 1D deep CNN, to predict the different phenotypes from novel genomic markers—SVs and TEs. Unlike the statistical models, our model learns the complex relationships between genome-wide markers and phenotypic traits from the training data. With the advanced DL technology, the model evades the overfitting of the data using the convolutional, pooling and dropout layers hence decreasing the complexity of dimensional genomic markers.

This article is organized as follows. In Section 2, first, we discuss our data sources, second, we present our proposed DL model, and third, we present the data representation for the proposed model in detail and its architecture and optimization strategies. Then, in Section 3, we discuss the prospective results by comparing the results with all of the mentioned statistical and DL models. Finally, we draw some conclusions and discuss how predicting phenotypes using SVs and TEs has more impact than using SNPs in Section 4.

2 Materials and methods

2.1 Data sources

2.1.1 *Arabidopsis thaliana* dataset with SV markers

The flowering plant, *A. thaliana*, is an ideal plant species for researching genotype–phenotype–environment interactions because their naturally inbred strains allow for repeated phenotyping and have adapted genotypes under a variety of controlled circumstances. *Arabidopsis thaliana* is quite appealing in scientific research for surveying the molecular and phenotypic effects at the species level (Weigel and Mott, 2009). It is found in a wide variety of environments around the world, and representative accessions have been extensively phenotypically and genetically characterized. The 1001 Genomes 'A Catalog of *A. thaliana* Genetic Variation' (Quadrana *et al.*, 2016) is a database where whole-genome sequencing has been performed on over 1000 *A. thaliana* accessions which makes the GS and GWAS in this species possible.

The full VCF variant files containing the structural variants data for the 1301 *A. Thaliana* samples are publicly available on European Variation Archive (PRJEB38975) (Göktay *et al.*, 2021). They were used in this study, as well as the SNPs and indels for 1135 accessions of *A. thaliana* (Alonso-Blanco *et al.*, 2016). The SV dataset includes several types of SVs, such as deletions, duplications and inversions. The lengths of these variations span from 50 bp to 10 kb.

The phenotypes of matched samples were downloaded from AraPheno, the 1001 genomes project (Togninalli *et al.*, 2020). Flowering time, rosette leaf number, cauline axillary branch number and stem length were used with their definitions listed as follows.

1. Flowering time: Generally scored as days from seeds in the soil until the first open flower.
2. Rosette leaf number: A shoot system morphology trait which is the number of leaves in the shoot system of *A. thaliana*.
3. Cauline axillary branch number: A shoot axis morphology trait which is the amount or pattern of branches arising from the shoot axis.

4. Stem length: A stem morphology trait often measured from the soil surface to the highest point on the stem.

2.1.2 *Oryza sativa subsp. Japonica* dataset with TE markers

Oryza sativa, rice is a monocotyledonous flowering plant in the *Poaceae* family that is one of the world's most significant agricultural plants, providing the primary source of nutrition for half of the world's population. *Oryza sativa subsp. Japonica* is one of three primary rice subspecies: *indica*, *javanica* and *japonica*. The grains of *O. sativa* are short and high in amylopectin, causing them to stick together when cooked.

The TE markers computed from the 176 *O. sativa* accessions were used in this study (Yan *et al.*, 2022). The SNPs for the same number of accessions were also obtained in this study in comparison to TE markers.

Four phenotypes are used in the study: panicle length, spikelet number, days to heading and plant height measured for the same accessions (Yano *et al.*, 2016).

1. Panicle length: The length measured from a terminal component of the rice tiller which is an inflorescence called a panicle.
2. Spikelet number: The number of panicle rice spikelets, which develop into grains.
3. Days to heading: Characterized together by the vegetative growth phase. It is the period from germination to panicle initiation and the reproductive phase of rice development, meaning the time from panicle initiation to heading.
4. Plant height: Defined as the shortest distance between the upper boundary (the highest point) of the main photosynthetic tissues (excluding inflorescences) and the ground level.

2.2 NovGMDeep: a deep learning model for genomic selection using novel genomic markers

We present, NovGMDeep, a deep 1D CNN to predict phenotypes in this study. The model is trained separately with three different types of data: SVs, TEs and SNPs. We evaluate the model by calculating Pearson's Coefficient of Correlation (PCC) among the predicted and observed phenotype values. The results show that SVs and TEs are more informative for phenotype prediction than SNPs.

2.3 Genomic data representation

Genotypes in plant and animal species are represented as either haploid or diploid. A *haploid* genotype comprises a single set of chromosomes for generating the phenotype whereas a diploid genotype contains two sets of chromosomes, each of which is required for phenotype generation. Genotypes are usually represented as '0|0' or '0/0', where '|' and '/' indicates the phased and unphased genotypes, respectively. To gain a complete picture of genetic variation, phasing entails separating maternally and paternally inherited copies of each chromosome into haplotypes. A genotype with at least one set of explanatory haplotypes is referred to as a phased genotype. Unphased genotypes are those for which no set of explaining haplotypes have been identified (Neigenfind *et al.*, 2008). The genotype is considered to be *homozygous* if all the haplotypes for a specific site have the same value. Otherwise, the genotype is considered *heterozygous*.

2.4 SV marker representation

Arabidopsis thaliana is a diploid species and has phased genotypes as '0|0' and '1|1' which means homozygous for reference allele and homozygous for alternate allele, respectively. Although SVs are smaller in numbers than SNPs, which account for almost a million variant sites per genome, they are more informative, as they account for a higher number of nucleotide differences due to their size. Different types of SVs such as inversions, duplications and deletions can capture the genomic variations better than the SNPs. We propose a strategy as shown in formula 1 to represent the SVs. The SV genotypic data for inversions, duplications and deletions are transformed using one-hot encoding as binary arrays. In one-hot encoding, given a dataset with different features, the encoder finds the unique values per feature and transforms the data into a binary one-hot vector (Pedregosa *et al.*, 2011). Different arrays for different types of SVs were chosen so that the details of genomic variations were well-represented and understandable by the model. For the SNP and indel data, the same one-hot encoding strategy was used, 0|0 for reference allele, 1|1 for alternate allele and .|. indicating the missing genotypes. As shown below, INV represents inversions, DUP represents duplications and DEL represents deletions:

$$\begin{aligned} 0|0(INV) &\rightarrow [1,0,0,0,0,0] \\ 1|1(INV) &\rightarrow [0,1,0,0,0,0] \\ 0|0(DUP) &\rightarrow [0,0,1,0,0,0] \\ 1|1(DUP) &\rightarrow [0,0,0,1,0,0] \\ 0|0(DEL) &\rightarrow [0,0,0,0,1,0] \\ 1|1(DEL) &\rightarrow [0,0,0,0,0,1] \\ .|.(MissingValue) &\rightarrow [0,0,0,0,0,0] \end{aligned} \tag{1}$$

2.5 TE marker representation

Oryza sativa is a diploid species and has unphased genotypes. The TE and SNP markers in the raw dataset were coded as integers 1, 0 and −1 which means the reference genotype (1/1), the first (most common) alternative genotype (0/1) and the second most frequent alternative genotype (0/0), respectively. A similar strategy is used to encode TE and SNP variants using one-hot-encoded binary arrays as shown in formula (2).

$$\begin{aligned} 1/1 &\rightarrow [1,0,0] \\ 0/1 &\rightarrow [0,1,0] \\ 0/0 &\rightarrow [0,0,1] \\ ./. &\rightarrow [0,0,0] \end{aligned} \tag{2}$$

For the both the *A. thaliana* and *O. sativa* datasets, the data are represented in the format of a 3D matrix $A(x \times y \times z)$ as the input for the proposed model shown in Table 1.

2.6 Model architecture

The proposed deep CNN model has four 1D convolutional layers, a single 1D max-pooling layer, a flatten layer and one dropout layer followed by a fully connected layer (Fig. 1).

As the backbone of the model, we use four 1D convolutional layers of the following sizes. The first layer includes 16 filters of size five. For the next layers, we double the number of filters in each layer, i.e. 32, 64 and 128. The filter size is five for the first two layers and for the last two layers it is reduced to three. If the input to the convolution layer is represented as X, filter as F, bias as B and the

Table 1. Data representation of different markers used in the study as the input for the NovGMDeep model

Dataset	Number of accessions (x)	Number of markers (y)	Size of one-hot-encoded array
Arabidopsis Thaliana (SV)	914	155 440	6
Arabidopsis Thaliana (SNP)	923	500 000	2
Oryza Sativa (TE)	176	6074	3
Oryza Sativa (SNP)	176	493 882	3

Note: In the *A. thaliana* (SNP) dataset, only 500 000 SNP genotypes were selected randomly from the 10 million genotypes to save computational resources.

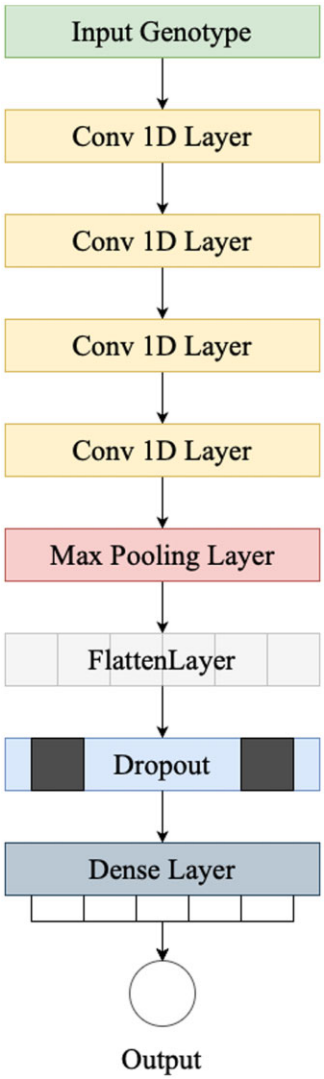

Fig. 1. NovGMDeep architecture. The first convolutional layer has 16 filters of kernel size five. The second layer comes with 32 filters of kernel size five. The third one has 64 filters of kernel size three. And the last convolutional layer has 128 filters of kernel size three. The max-pooling layer has a pool size of two and a stride of two. At last, there is one fully connected layer as the regressor with one output neuron

output of the convolution layer as Y, the convolution operation between X and F including bias is defined in Equation (3).

$$Y(N_i) = F * X(N_i) + B \tag{3}$$

where N is the batch size and * is the cross-correlation (Bracewell, 1965) operator. For all the convolutional layers, ReLU activation function (Agarap, 2018) (Equation 4), LecunNormal kernel initializer (Hanin and Rolnick, 2018) and L1L2 kernel regularizer (Krishnan *et al.*, 2011) are used with same padding.

$$ReLU\Big(Y(N_i)\Big) = \max\Big(0, Y(N_i)\Big) \tag{4}$$

To reduce the high dimensionality of the prominent features of the last convolutional layers, a 1D max-pooling layer is used after the convolutional layers with a window size of two and a stride of two. It minimizes the amount of the input to the following layer by taking only the maximum values as the most important features, which cut the size of the input into the following layer in half (Albawi *et al.*, 2017). Then the extracted feature map from the model's backbone is flattened using a flatten layer and fed into the fully connected layer which is our phenotype predictor. Fully connected layers link every node in the input to every node in the output, but they do not capture spatial information. Due to the high number of input features which results in a large-size feature map, the dropout layer with the drop ratio of 0.6 is used before the fully connected layer to reduce computational complexity and avoid overfitting. We use Adam (Kingma and Ba, 2014) optimizer with a learning rate of 0.0003. The Adam optimization algorithm is applied to iteratively adjust the network weights based on training samples. Moreover, we utilized mean absolute error (MAE) as the loss function. The NovGMDeep is developed with Keras (Gulli and Pal, 2017), an open-source software library that provides a Python interface for artificial and deep neural networks. The source code of NovGMDeep is publicly available on GitHub (https://github.com/shivanisehrawat4/NovGMDeep.git).

Due to the large number of features, four convolutional layers were used to extract the most representative feature maps for the linear predictor. On the one side, with a small number of training samples, increasing the number of layers made the model more complex and led to overfitting. On the other side, decreasing the number of layers as well could not provide the predictor with representative feature maps. Even with four convolutional layers, we had a huge output feature map, resulting in a complex linear predictor. Therefore, we needed to use a dropout layer to decrease the complexity of the predictor and avoid overfitting. We made the above-mentioned choices based on observing the stability of the models' performances during the model development and final evaluation. Using a 3-fold cross-validation strategy, all three developed models should ideally show stable training behavior by eventually improving the training and validation performance, leading to high-performing models.

2.7 Performance metrics

A regression model's performance or competence must be stated as an error in those predictions. As we are predicting numerical phenotypes in this study, such as plant height, we are more interested in how closely the predictions came to the actual values rather than whether the model accurately anticipated the value. This is precisely what error does, summarizing on average how well predictions matched expected values. For assessing and summarizing the effectiveness of our proposed model, MAE metric was employed.

The MAE metric is used to calculate the validation loss on the test set. MAE refers to the magnitude of difference between the prediction of observation and the true value of that observation.

$$MAE = \frac{1}{N}\sum_{i=0}^{N}|P_i - O_i|, \tag{5}$$

where N represents the total number of individuals present in the training set and P_i and O_i denote the values for the ith predicted and ith observed phenotypes.

The SD is a measure of how spread out numbers are.

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \tag{6}$$

where x is a value in the dataset, μ is the mean of the dataset and N is the number of data points in the population.

The model predicts quantitative phenotype as a real number which is the model's output. The predicted phenotypes are later compared with the ground-truth values to evaluate the model performance. As the evaluation metric, we use PCC between the predicted and observed phenotypes. Five assumptions must be met before we calculate the PCC between two variables (Onwuegbuzie and Daniel, 1999). As far as our datasets are concerned, all of the conditions are true and the detailed analysis can be found in Appendix. PCC (Glen, 2021), normally denoted by ρ, is a measure of the linear correlation between two variables (X and Y), whose values range from -1 to 1, with 1 indicating total positive correlation, 0 indicating no correlation and -1 indicating a total negative

correlation. It is defined as the covariance of the two variables divided by the product of their SDs.

$$\rho = \frac{cov(X, Y)}{\sigma_X \sigma_Y}, \quad (7)$$

where $cov(X, Y)$ is the covariance of X and Y, σ_X is the SD of X, and σ_Y is the SD of Y. PCC is a model-free method, and it, therefore, shows the nature of the data without depending on any of the existing models. It is commonly used in statistical analysis for the evaluation of the model and measuring the relationship between two variables.

3 Results

We trained the NovGMDeep model on both SVs and SNPs variations of the accessions and their associated phenotypes for *A. thaliana*. We also trained the model on TE and SNP data of *O. sativa*. We split all the samples into training and testing subsets with the ratios of 80% and 20%, respectively. For the sake of model development, we applied 3-fold cross-validation (Hastie *et al.*, 2009) solely on the training sets of both datasets. Before data preprocessing and model development, the test sets were left out to avoid any information leaks. Thereafter, the same preprocessing strategy used for training sets was used for the test sets, which were separately preprocessed and merely used for model evaluation. In the model development step, we save the three best-obtained models, each developed in each of the three rounds of cross-validation. Then, the three best models were used in the evaluation phase to calculate the PCC value on the test set which can be seen in Figure 2. Later the average of the three obtained PCC values was finally used as the main results (shown in Tables 2 and 3) to evaluate the overall model performance.

PCC was used to find the correlation between the observed and predicted phenotypes. The correlation between the two quantitative variables assessed for the same samples is depicted by scatter plots in Figure 3 The horizontal axis displays the true values of the phenotypes, while the vertical axis shows the values of the predicted phenotypes. Each value in the data is represented by a point on the graph. The strength of the association between the two variables is a key aspect of a scatterplot. The slope conveys information about the strength of the relationship between true and predicted values. When the slope is 1, the correlation between the two comparable variables is the strongest. Figure 3 shows the results between ground truth and predicted phenotypic values for NovGMDeep using SVs and SNPs, respectively, for predicting flowering time of *A. thaliana*. The PCC graphs were created on the testing samples and show the accuracy of the proposed model in predicting phenotypes. The orange line in these graphs represents the trend of the relationship between the observed and predicted phenotypes (blue points) by the NovGMDeep model. This linear regression line shows the best fit between predicted and true values. As shown in Figure 3, the predicted values which are close to the fitting line indicate a strong correlation, and values far from the fitting line indicate a weak correlation between the true and predicted values.

The aforementioned two statistical methods, rrBLUP and gBLUP, were evaluated with the same data to compare their overall prediction performance with NovGMDeep. As shown in Table 2, phenotypes predicted by NovGMDeep show the strongest correlations with observed phenotypes. Therefore, NovGMDeep outperforms the other statistical models in predicting phenotype with SVs data. Moreover, phenotypes predicted by NovGMDeep using SVs

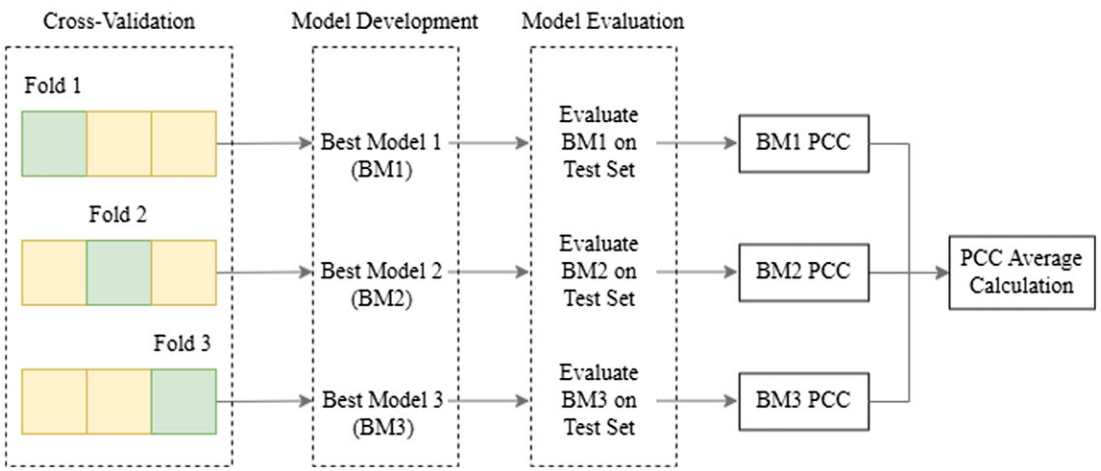

Fig. 2. The process of calculating the average PCC value from 3-fold cross-validation for the *A. thaliana* and *O. sativa* datasets. The three best models were saved in the model development phase in the three rounds of cross-validation. Then, the three best models were used in the model evaluation phase to calculate the PCC value on the test set. The average value of obtained PCC values is later reported as the final result

Table 2. PCCs between the observed and predicted phenotypes on the testing set

	Flowering time		Rosette leaf number		Branch number		Stem length	
Model	PCC (SVs)	PCC (SNPs)	PCC (SVs)	PCC (SNPs)	PCC (SVs)	PCC (SNPs)	PCC (SVs)	PCC (SNPs)
rrBLUP	0.664	0.520	0.653	0.547	0.675	0.513	0.690	0.502
gBLUP	0.657	0.516	0.662	0.512	0.683	0.530	0.681	0.514
DeepGS	–	0.542	–	0.559	–	0.544	–	0.563
G2Pdeep	–	0.561	–	0.557	–	0.571	–	0.569
NovGMDeep	**0.788**	**0.594**	**0.742**	**0.563**	**0.761**	**0.581**	**0.759**	**0.572**

Notes: SVs and SNPs are used to capture the genomic variations among the accessions of *A. thaliana*. Results show the average of the three PCC values taken from the 3-fold cross-validation results of the model. Top performers are highlighted as bold.

Table 3. PCCs between the observed and predicted phenotypes on the testing set

	Days to heading		Plant height		Spikelet number		Panicle length	
Model	PCC (TEs)	PCC (SNPs)	PCC (TEs)	PCC (SNPs)	PCC (TEs)	PCC (SNPs)	PCC (TEs)	PCC (SNPs)
rrBLUP	0.439	0.281	0.391	0.275	0.387	0.294	0.405	0.209
gBLUP	0.417	0.212	0.402	0.266	0.371	0.282	0.387	0.215
DeepGS	–	−0.004	–	−0.015	–	−0.171	–	−0.165
G2Pdeep	–	0.023	–	−0.011	–	−0.036	–	−0.182
NovGMDeep	**0.436**	**−0.014**	**0.406**	**0.002**	**0.372**	**−0.169**	**0.397**	**−0.148**

Notes: TEs and SNPs are used to capture the genomic variations among the accessions of *O. sativa*. Results show the average of the three PCC values taken from the 3-fold cross-validation results of the model. Top performers are highlighted as bold.

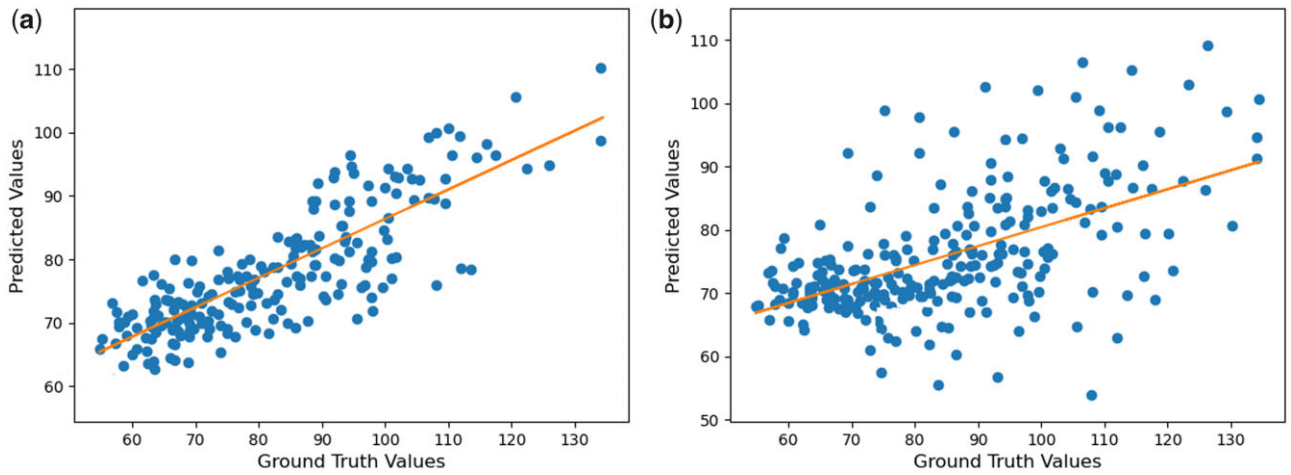

Fig. 3. PCC analysis of *A. thaliana* for predicting flowering time using (a) SV and (b) SNP datasets. (a) The orange line represents the trend of the relationship between predicted and ground-truth values (blue points). The graph shows that predicted values by the model positively follow the ground truth values with a high confidence level (minimum fluctuation) in a positive direction. (b) The graph shows that predicted values by the model positively follow the ground truth values but with a low confidence level (more fluctuation)

Table 4. MAE and SD of MAE calculated by `NovGMDeep` model on the test sets for predicting the flowering time of *A. thaliana* using SV and SNP dataset

	SVs		SNPs	
Phenotypes	MAE	SD	MAE	SD
Flowering time	7.64	6.51	13.71	11.56
Rosette leaf number	8.23	7.02	14.89	12.47
Branch number	7.89	6.52	14.84	12.62
Stem length	7.92	6.71	14.31	13.09

Table 5. MAE and SD of MAE calculated by `NovGMDeep` model on the test sets for predicting the flowering time of *O. sativa* using TE and SNP dataset

	TEs		SNPs	
Phenotypes	MAE	SD	MAE	SD
Days to heading	8.65	6.41	16.84	18.51
Plant height	8.01	6.57	15.92	19.74
Spikelet number	9.26	7.22	19.46	20.25
Panicle length	9.51	6.38	16.78	19.02

and TEs data show stronger correlations with observed phenotypes than the ones predicted using SNPs data. This shows the effectiveness of using SV and TE marker data instead of SNPs for phenotype prediction. Though, TE data showed the lowest PCC values compared to the two statistical models because of the mere 176 samples as there are not enough training samples for the DL model (Table 3).

The `NovGMDeep` model was also compared with the aforementioned DL models. However, the comparison was possible just on the SNP data as these models are specially designed for that. We tried the G2Pdeep, which is a web-based framework where users can upload their datasets and predicts phenotypes by creating a DL model according to the data. Because of the extremely high-dimensional SNP dataset we used in this article, the framework did not work well for us and, therefore, we ran the code uploaded into the repository that is publicly available in GitHub. DeepGS package was also run on the command line. PCC values were calculated for all the phenotypes of *A. thaliana* and *O. sativa* in Tables 2 and 3, respectively.

The results in the Table 2 showed that `NovGMDeep` performed better than the other existing DL models. Also, it can be seen that G2Pdeep has a better performance than DeepGS. However, the results for DL models suggest that statistical models are better for datasets with fewer samples as seen in Table 3.

The MAE loss was watched to select the best model during the training process of 150 epochs with early stopping. The model with the lowest MAE was selected to evaluate the model on the test set as shown in Tables 4 and 5. The SD of all the MAE values on the test set was also calculated to see the deviation among those values. The MAE measures the average absolute error over the dataset while the SD measures how far the absolute error on each training point is from the MAE. A low SD means that errors across the dataset tend

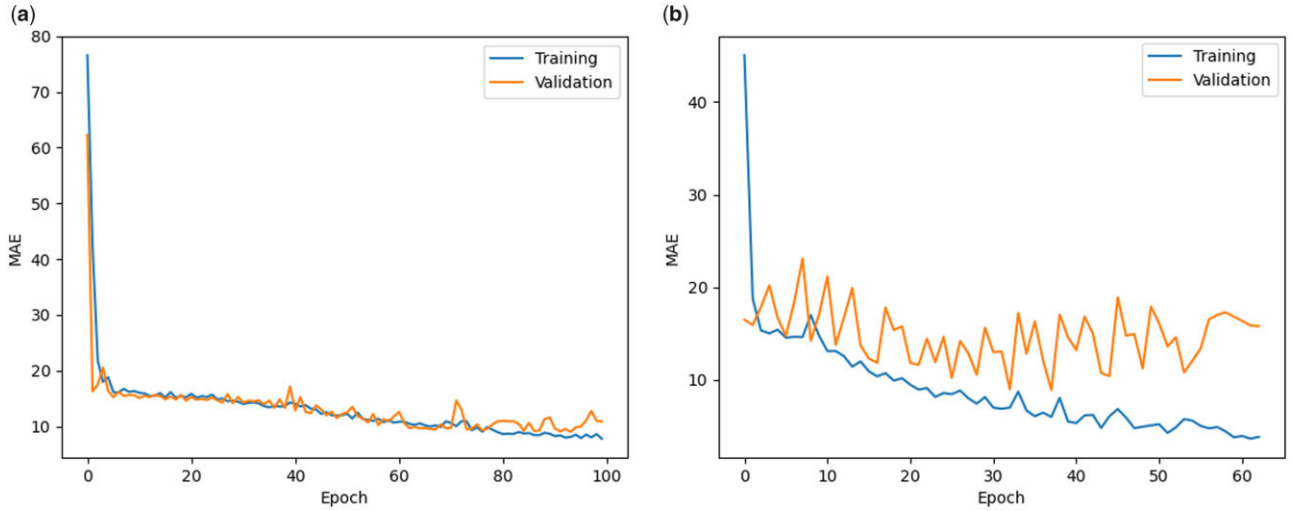

Fig. 4. Training and validation losses of NovGMDeep model using (**a**) SV and (**b**) SNP dataset for predicting flowering time of *A. thaliana*. The first few epochs illustrate that the model is learning quickly as both the training and validation losses are decreasing faster. Later it shows the potential of the model in handling the co-variate shift between sets as the training and validation losses are decreasing steadily now

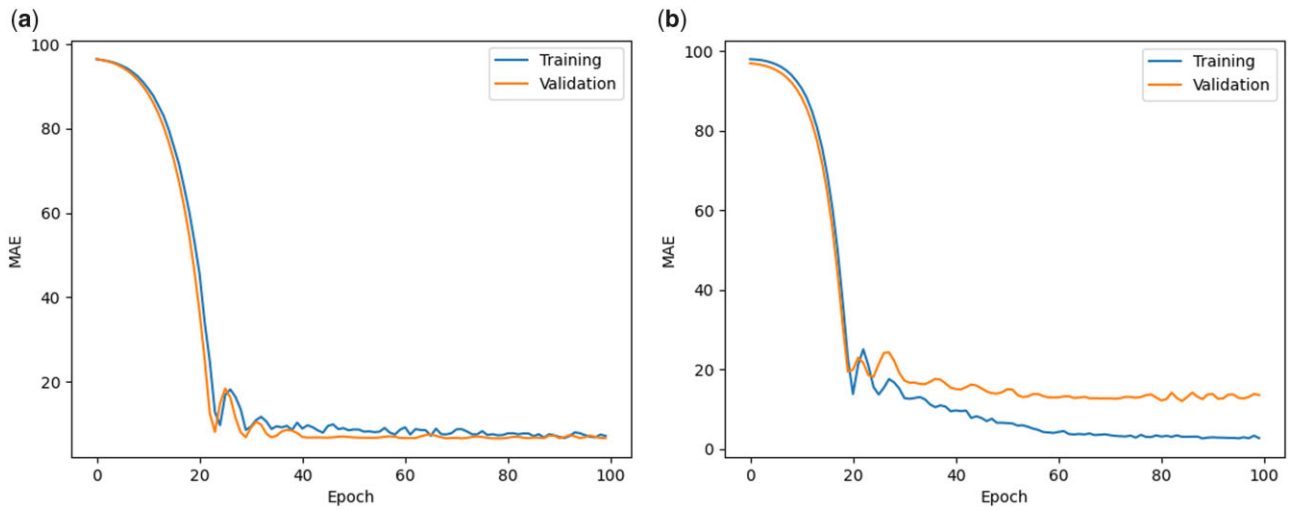

Fig. 5. Training and validation losses of NovGMDeep model using (**a**) TE and (**b**) SNP dataset for predicting days to heading of O. *sativa*. The graph shows that the model is learning slowly at the start as it took 20 epochs for it to be at a steady state. Later it shows the potential of the model in handling the co-variate shift between sets as the training and validation losses are decreasing steadily now

to have similar values close to the mean. A high SD tells that the errors are spread over a bigger range. This can provide insight into the model: a model with a low MAE indicates a good 'average' performance over the dataset, and if that model also has a low SD then it tells that the performance is not only good on average, but also uniformly on the dataset. The low values of MAE and SD of MAE for the SV and TE markers show higher prediction accuracy of NovGMDeep model over the SNPs.

Figures 4 and 5 demonstrate a typical training process for all the datasets. The trend shows that the model keeps learning the discriminative features of the data for the first few epochs as both the training and validation losses are decreasing to a significant amount at the start. For the rest of the epochs, the flattened part of the graph illustrates that the model is capable to handle the co-variate shift between the training and validation set because now the training and validation losses are decreasing steadily. The closeness of the training and validation curves demonstrates that the model is neither overfitted nor underfitted in Figures 4(a) and 5(a). In Figures 4(b) and 5(b) the training curve is lower than the validation curve, which shows that the model is overfitted. An underfitted model will have a high training and high validation loss while an overfitted model will have an extremely low training loss but a high validation loss. A large number of features (SNPs) expands the hypothesis space, making the data more sparse and this might also lead to overfitting problems.

4 Discussions and conclusions

GS has currently brought a revolution in applications of breeding programs in plants and livestock (Hill, 2014). Novel prediction algorithms and methods for predicting complex traits of large genotypic data have increasingly become an essential need of breeders. High-performing DL techniques have served the purpose of the principal need for breeders to improve their practices. DL as an advanced technique of machine learning algorithms, has the potential to find hidden patterns in huge datasets. This work aimed to develop a DL model and evaluate its accuracy in predicting the specific traits from genome-wide SV and TE markers.

In this article, the applications of DL and its relationship with GS have been explored. Although the literature suggests DL

algorithms as a solid method for predicting complex phenotypes, there are a few limitations to the usage of DL techniques. In the existence of large-scale datasets, the most important consideration for high prediction accuracy is the model's architecture which requires great knowledge of DL techniques. A well-constructed network in any model is the key source for the high performance of the model. Another crucial consideration is the choice of applying convolutional, fully connected, dropout and sampling layers with distinct sets of hyperparameters that handle distinct characteristics of the data (Angermueller *et al.*, 2016). Interpreting the biological relevance of the data is also a helpful consideration to minimize the limitations of DL techniques in bioinformatics. The small number of DL models for GS applications demonstrates the enormous potential for these models to enhance early candidate genotype selection and enhance knowledge of the intricate biological mechanisms underlying the link between genotypes and phenotypes. This potential is partially explained by the way such models are constructed, which provides them the capacity to recognize more intricate data patterns.

In this study, we developed NovGMDeep as a method to predict phenotypes based on different types of SVs (inversions, duplications and deletions) and TEs. This work was motivated by the importance of SVs in genome evolution and that SVs could better capture variants among genomes than SNPs (Stankiewicz and Lupski, 2010). With the proposed model, it is still difficult to link phenotypes to genetic SV and environmental variables. Also, the authors of Yan *et al.* (2022) concluded through GWAS that TE markers have a similar ability to discover association patterns to SNP markers. Therefore, the model is not only for predicting phenotypes from SVs and TEs but also shows that SVs can be more beneficial in GS than SNPs.

Both statistical models (rrBLUP and gBLUP) have the lowest PCC values compared to the NovGMDeep model for the *A. thaliana* data. This is potentially because simple regressions cannot capture the complex relationships between genotypes and phenotypes. For the *O. sativa* data, the PCC value is very low for TEs and even negative for SNPs because of the low sample size to train the DL model. The results for the TE markers are better than the SNPs which makes the former more useful. Different model architectures impact the results significantly. Finding a dataset with enough samples and related phenotypic labels to train a neural network to be efficient and broadly applicable is the primary challenge that must be overcome to employ DL techniques for phenotype prediction. There are several datasets with significant numbers of sequenced genomes that are available, but the majority of these databases lack reliable phenotypic data for the linked samples. The cost of sequencing sufficient depth of coverage for SV and TE calls is also high. The search for datasets with more than 176 samples and associated phenotypes for TEs was unsuccessful as of the time of writing this article. It was challenging to develop a generalizable prediction model for TEs due to the issues with overfitting and short validation sets caused by the use of datasets with a small sample size (Keshari *et al.*, 2020; Lawrence *et al.*, 1997; Montesinos-López *et al.*, 2018). Due to a large number of SNP markers across the genome used as features, overfitting increases in the model. This is potentially because the number of SNPs is significantly larger than the number of samples, causing the 'small-*n*-large-*p* problem'. The performance comparison between NovGMDeep and other DL models, such as DeepGS (Ma *et al.*, 2018) and G2Pdeep (Zeng *et al.*, 2021), was only possible with SNP data. NovGMDeep outperformed the other two DL models for *A. thaliana* whereas the low sample size problem persisted for *O. sativa*.

The study also shows that the nature of data, network architecture and the type of the trait which is being predicted play an influential role in model training and prediction. It is important to focus on the individuals with high phenotypic trait values so that they can serve as a selected asset for the different breeding programs for other vital crops. In future work, we will explore finding top-ranked individuals with high phenotypic values by incorporating all types of genomic data, such as SVs, TEs and SNPs together. Moreover, environmental data, such as weather conditions, can be taken into consideration in phenotype prediction. Overall, the NovGMDeep model was well implemented in the prediction of phenotypes using SVs genotypic markers and could be added to the toolkits of crop breeders. Also, the results showed the importance of the selection of algorithms and hyperparameters for genotype-to-phenotype predictions. In conclusion, this study paves the way for the novel use of SVs and TEs in the field of GS.

Author contributions

Shivani Sehrawat (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [equal], Resources [lead], Software [equal], Validation [lead], Writing—original draft [lead]), Keyhan Najafian (Investigation [supporting], Methodology [equal], Software [equal], Validation [supporting], Writing—original draft [equal], Writing—review and editing [supporting]) and Lingling Jin (Conceptualization [lead], Data curation [equal], Formal analysis [equal], Funding acquisition [lead], Investigation [equal], Methodology [equal], Project administration [lead], Supervision [lead], Validation [equal], Writing—review and editing [lead]).

Software and data availability

The data and source code underlying this article are freely available and can be accessed with links as follows.

The phenotypes for *A. thaliana* were downloaded from https://arapheno.1001genomes.org/study/12/ and https://arapheno.1001genomes.org/study/38/

The TE markers for *O. sativa* were downloaded from https://data.cyverse.org/dav-anon/iplant/home/yanhaidong1991/opt_genos_fltmissing_fltmaf_fltukn_TE.txt and the SNPs were downloaded from https://data.cyverse.org/dav-anon/iplant/home/yanhaidong1991/176GATK.indelsnpsFiltered0.05M0.25.txt.

The G2Pdeep web-based framework can be found at https://g2pdeep.org.

The standalone G2Pdeep and DeepGS packages can be found at https://github.com/kateyliu/DL_gwas and https://github.com/cma2015/DeepGS, respectively.

Funding

This research is supported by Discovery grant to LJ from the Natural Sciences and Engineering Research Council of Canada.

References

Agarap,A.F. *et al.* (2018) Deep Learning using Rectified Linear Units (ReLU). *arXiv preprint arXiv:1803.08375*.

Albawi,S. *et al.* (2017) Understanding of a convolutional neural network. In: *2017 International Conference on Engineering and Technology (ICET)*, Antalya, Turkey, pp. 1–6. IEEE.

Alonge,M. *et al.* (2020) Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, **182**, 145–161.

Alonso-Blanco,C. *et al.* (2016)) 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, **166**, 481–491.

Angermueller,C. *et al.* (2016) Deep learning for computational biology. *Mol. Syst. Biol.*, **12**, 878.

Bracewell,R. *et al.* (1965) *Pentagram Notation for Cross Correlation. The Fourier Transform and Its Applications*. Vol.**46**. McGraw-Hill, New York, p. 243.

Clark,S.A. and Werf,J. (2013) Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. In: Gondro,C. *et al.* (eds) *Genome-Wide Association Studies and Genomic Prediction*. Springer, Berlin, Germany. pp. 321–330.

Deng,L. *et al.* (2013) New types of deep neural network learning for speech recognition and related applications: an overview. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, pp. 8599–8603. IEEE.

Dolatabadian,A. *et al.* (2017) Copy number variation and disease resistance in plants. *Theor. Appl. Genet.*, **130**, 2479–2490.

Endelman,J.B. (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*, **4**, 250–255.

Gabur,I. *et al.* (2019) Connecting genome structural variation with complex traits in crop plants. *Theor. Appl. Genet.*, **132**, 733–750.

Glen,S. (2021) Correlation coefficient: simple definition, formula, easy steps. *StatisticsHowTo.Com*. https://www.Statisticshowto.Com/probability-and-statistics/correlation-coefficient-formula/ (3 August 2020, date last accessed).

Göktay,M. *et al.* (2021) A new catalog of structural variants in 1,301 *A. thaliana* lines from Africa, Eurasia, and North America Reveals a signature of balancing selection at defense response genes. *Mol. Biol. Evol.*, **38**, 1498–1511.

Gulli,A. and Pal,S. (2017) *Deep Learning with Keras*. Packt Publishing Ltd, Birmingham, UK.

Hanin,B. and Rolnick,D. (2018) How to start training: the effect of initialization and architecture. In: *Advance Neural Information Processing Systems*, Vol. 31, Montréal, Canada.

Hastie,T. *et al.* (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. **2**. Springer, Berlin, Germany.

Hill,W.G. (2014) Applications of population genetics to animal breeding, from wright, fisher and lush to genomic prediction. *Genetics*, **196**, 1–16.

Islam,M.M. *et al.* (2018) Deep learning models for predicting phenotypic traits and diseases from omics datas. In: Aceves-Fernandez, M.A. (ed.) *Artificial Intelligence-Emerging Trends and Applications*. IntechOpen, London, Greater London, UK.

Keshari,R. *et al.* (2020) Unravelling small sample size problems in the deep learning world. In: *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, New Delhi, India, pp. 134–143. IEEE.

Kingma,D.P. and Ba,J. (2014) ADAM: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krishnan,D. *et al.* (2011) Blind deconvolution using a normalized sparsity measure. In: *CVPR 2011 Colorado Springs*, pp. 233–240. IEEE.

Lai,J. *et al.* (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.*, **42**, 1027–1030.

Lawrence,S. *et al.* (1997) Lessons in neural network training: overfitting may be harder than expected. In: *Aaai/Iaai*. Citeseer, Palo Alto, California, USA, pp. 540–545.

LeCun,Y. *et al.* (2015) Deep learning. *Nature*, **521**, 436–444.

Li,B. *et al.* (2018) Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet.*, **9**, 237.

Ma,W. *et al.* (2018) A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta*, **248**, 1307–1318.

Meuwissen,T.H. (2009) Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genet. Sel. Evol.*, **41**, 1–9.

Min,S. *et al.* (2017) Deep learning in bioinformatics. *Brief. Bioinf.*, **18**, 851–869.

Montesinos-López,A. *et al.* (2018) Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3*, **8**, 3813–3828.

Montesinos-López,O.A. *et al.* (2021) A review of deep learning applications for genomic selection. *BMC Genom.*, **22**, 1–23.

Neigenfind,J. *et al.* (2008) Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. *BMC Genom.*, **9**, 1–26.

Onwuegbuzie,A.J. and Daniel,L.G. (1999) Uses and misuses of the correlation coefficient, ERIC.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Qian,L. *et al.* (2016) Deletion of a Stay-Green gene associates with adaptive selection in brassica napus. *Mol. Plant.*, **9**, 1559–1569.

Quadrana,L. *et al.* (2016) The *Arabidopsis thaliana* mobilome and its impact at the species level. *eLife*, **5**, e15716.

Roy,N.S. *et al.* (2015) Marker utility of transposable elements for plant genetics, breeding, and ecology: a review. *Genes Genom.*, **37**, 141–151.

Rumelhart,D.E. *et al.* (1986) Learning representations by back-propagating errors. *Nature*, **323**, 533–536.

Stankiewicz,P. and Lupski,J.R. (2010) Structural variation in the human genome and its role in disease. *Annu. Rev. Med.*, **61**, 437–455.

Togninalli,M. *et al.* (2020) AraPheno and the AraGWAS catalog 2020: a major database update including RNA-Seq and knockout mutation data for *Arabidopsis thaliana*. *Nucleic Acids Res.*, **48**, D1063–D1068.

Ubbens,J. *et al.* (2021) Deep neural networks for genomic prediction do not estimate marker effects. *Plant Genome.*, **14**, e20147.

Weigel,D. and Mott,R. (2009) The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol.*, **10**, 1–5.

Wong,A.H. *et al.* (2005) Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Hum. Mol. Genet.*, **14** (**Suppl_1**), R11–R18.

Yan,H. *et al.* (2022) Exploring transposable element-based markers to identify allelic variations underlying agronomic traits in rice. *Plant Comm.*, **3**, 100270.

Yano,K. *et al.* (2016) Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat. Genet.*, **48**, 927–934.

Young,T. *et al.* (2018) Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.*, **13**, 55–75.

Yuan,Y. *et al.* (2021) Current status of structural variation studies in plants. *Plant Biotechnol. J.*, **19**, 2153–2163.

Zeng,S. *et al.* (2021) G2pdeep: a web-based deep-learning framework for quantitative phenotype prediction and discovery of genomic markers. *Nucleic Acids Res.*, **49**, W228–W236.

Zhao,Z.Q. *et al.* (2019) Object detection with deep learning: a review. *IEEE Trans. Neural Netw. Learn. Syst.*, **30**, 3212–3232.

Zheng,A. and Casari,A. (2018) *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc., Sebastopol, California, USA.

Appendix: Detailed analysis of the assumptions of PCC

Five assumptions must be met before we calculate the PCC between two variables (Onwuegbuzie and Daniel, 1999). As far as our datasets are concerned, all of the conditions are true. First, while measuring the two variables, we should consider the interval or ratio level. All the phenotypes used in the article can describe by using intervals, for example, plant height, branch number, stem length etc. on the real number line. These all are physical measures which fall into general continuous category.

Second, there must be a linear relationship between both variables. All the phenotypes when plotted using scatter plot, fell roughly along a straight line, which can also be seen in Figure 3. The main thing to consider here is that the variables should not exhibit some other type of relationship, like quadratic.

Third, both variables should have a roughly normal distribution. We have plotted the values of flowering time for *A. thaliana* and days to heading for *O. sativa*. Figures A1 and A2 show a bell-shaped curve for both of these phenotypes that indicates that the variables follow a normal distribution.

Fourth, each observation in the dataset needs to have a pair of related values. It simply means that to calculate the correlation between the variables, each observation in the dataset has one measurement for the observed value and one measurement for the predicted value. This one was easy to check as `NovGMDeep` predicted phenotypes for all the corresponding ground truth values.

Fifth, the dataset should not contain any extreme outliers. An extreme value in the dataset substantially changes the PCC between the two variables. We did not find any extreme outlier in the dataset of predicted and observed phenotypes.

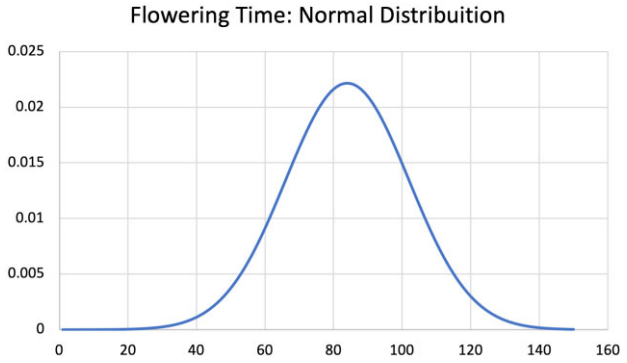

Fig. A1. The bell-shaped curve show the values of flowering time are normally distributed

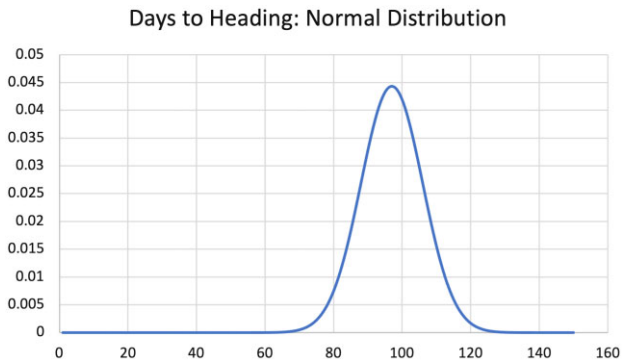

Fig. A2. The bell-shaped curve show the values of days to heading are normally distributed