Technology in Agronomy

REVIEW

Open Access

https://doi.org/10.48130/tia-0025-0002

Technology in Agronomy **2025**, 5: e006

Progress and perspectives on genomic selection models for crop breeding

Dongfeng Zhang[#], Feng Yang[#], Jinlong Li[#], Zhongqiang Liu, Yanyun Han, Qiusi Zhang, Shouhui Pan, Xiangyu Zhao and Kaiyi Wang[*]

Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China

[#] Authors contributed equally: Dongfeng Zhang, Feng Yang, Jinlong Li

[*] Corresponding author, E-mail: wangky@nercita.org.cn

Abstract

Genomic selection, a molecular breeding technique, is playing an increasingly important role in improving the efficiency of artificial selection and genetic gain in modern crop breeding programs. A series of algorithms have been proposed to improve the prediction accuracy of genomic selection. In this review, we describe emerging genomic selection techniques and summarize methods for best linear unbiased prediction and Bayesian estimation of the traditional statistics used for prediction during genomic selection. Moreover, with the rapid development of artificial intelligence, several machine learning algorithms are increasingly being employed to capture the effects of more genes to further improve prediction accuracy, which we describe in this review. We also describe the advantages and disadvantages of traditional models and machine learning models and discuss several crucial factors that could affect prediction accuracy. We propose that additional artificial intelligence techniques will be required for big data management, feature processing, and model innovation to generate a comprehensive model to optimize the prediction accuracy of genomic selection. We believe that improvements in artificial intelligence could accelerate the arrival of Breeding 4.0, in which combining any known alleles into optimal combinations in crops will be fully customizable.

Citation: Zhang D, Yang F, Li J, Liu Z, Han Y, et al. 2025. Progress and perspectives on genomic selection models for crop breeding. *Technology in Agronomy* 5: e006 https://doi.org/10.48130/tia-0025-0002

Introduction

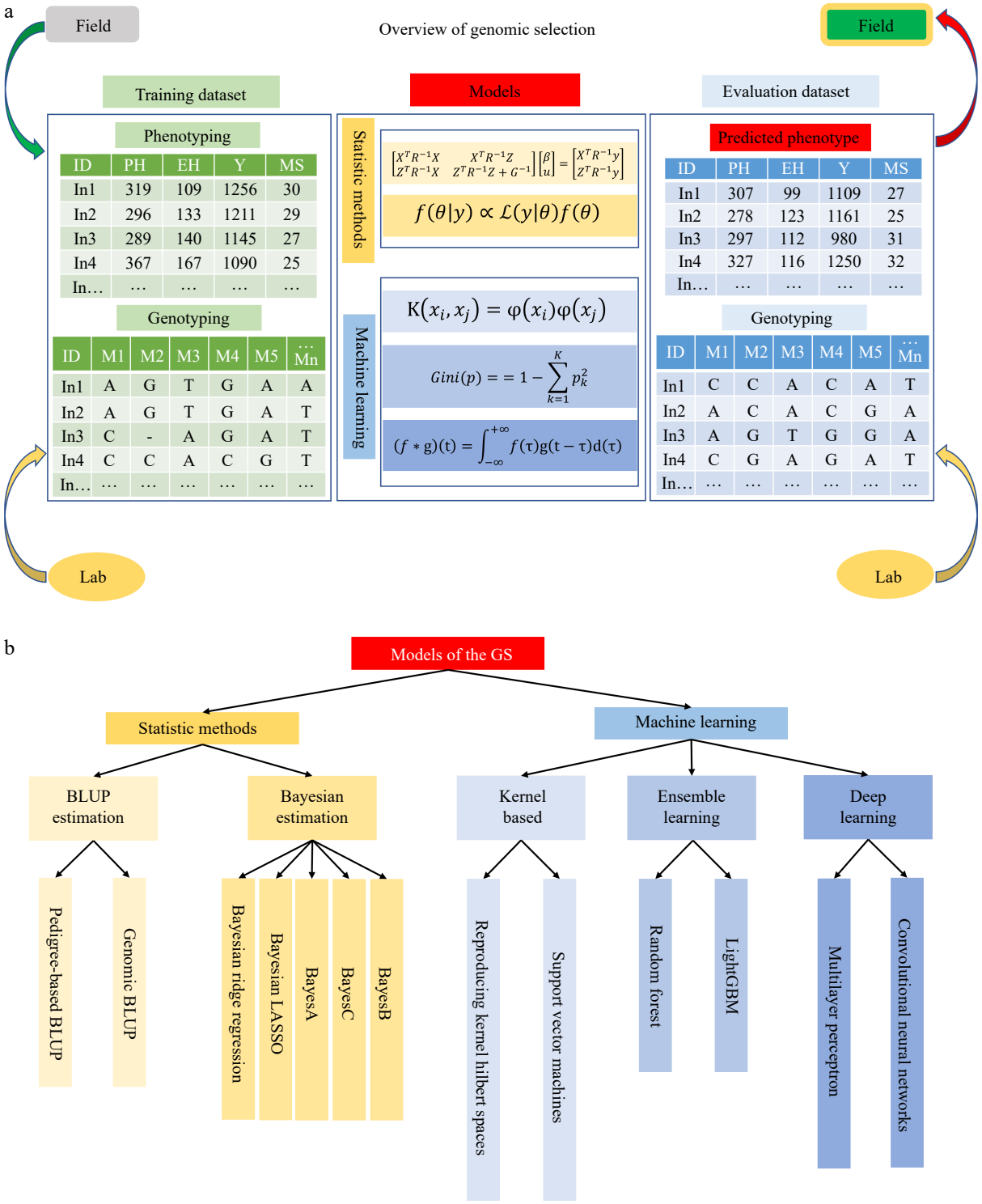
The development of new crop varieties through breeding is an efficient way to enhance crop productivity. Breeders aim to improve breeding traits of interest such as yield, disease resistance, stress tolerance, and nutritional value. Modern crop breeding is both an art and a science, and is vastly different from early crop selection and domestication due to the rapid development of scientific techniques such as biostatistics, genetic engineering, and genomics and to advancements in various modern breeding techniques, including transgenesis, genome editing, speed breeding, and doubled haploid (DH) technology[1]. As a result, modern breeding is entering a novel stage, Breeding 4.0, based on the proposed stages of agriculture from 1.0 to 4.0[1].

Two major strategies are employed in molecular breeding: marker-assisted selection (MAS), and genomic selection (GS)[2]. During MAS, each individual plant is identified based on linked or functional markers that confirm the trait of interest[3]. Owing to the discovery of increasing numbers of major quantitative trait loci (QTLs) and functional genes through genetic research, MAS has been extensively applied in different types of breeding programs to improve traits of interest since its initial use in the 1990s[3]. MAS is an effective way to increase selection efficiency for qualitative traits or categories of traits that are regulated by only a few genes[4]. However, important traits targeted by breeding that are regulated by one or a few major QTLs/genes are extremely rare. Most traits, such as yield and plant height, are quantitative traits that are regulated by multiple loci. As it is challenging to conduct artificial selection using only a few markers to identify the genes underlying traits of interest, a second strategy known as GS (also known as whole-genome prediction) was developed[5].

Guidelines of GS

In breeding programs, much effort was focused on predicting the breeding value of a specific material even before the emergence of GS. One approach that breeders have explored is predicting the performance of the progeny of single crosses. However, predicting the performance of the progeny of single crosses can be complex due to the presence of a genetic network, including factors with dominant or epistatic effects, especially for complex quantitative traits[6]. With the emergence of molecular biology techniques, DNA molecular markers have been developed and used in breeding programs. Researchers have started to incorporate these markers into regression models to estimate the breeding value of breeding materials. The use of DNA markers is effective for increasing the genetic gain, with improvements ranging from 8% to 38% in various studies[3,7]. Simultaneously, with advancements in marker technology, it has become feasible to produce markers rapidly at a lower cost, opening new avenues for incorporating genomic information into breeding programs[8].

GS involves predicting the breeding value of a material using information from a large number of genetic markers distributed across the genome. This technique was formally proposed by Meuwissen et al. in 2001, who proved that it was possible to predict breeding value using Bayesian and BLUP (best linear unbiased prediction) statistical models[5]. GS involves three major steps: construction of the training dataset, optimization of the model and parameters, and prediction of the evaluation dataset. These steps collectively form the basis of the GS workflow (Fig. 1a). However, the development of each step faced various challenges before yielding promising results. For instance, GS was not incorporated into breeding programs in the early 2000s due to the limited number of molecular markers available and the high cost of marker testing[9].



ID	PH	EH	Y	MS
In1	319	109	1256	30
In2	296	133	1211	29
In3	289	140	1145	27
In4	367	167	1090	25
In…	…	…	…	…

ID	M1	M2	M3	M4	M5	… Mn
In1	A	G	T	G	A	A
In2	A	G	T	G	A	T
In3	C	-	A	G	A	T
In4	C	C	A	C	G	T
In…	…	…	…	…	…	…

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z+G^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} X^TR^{-1}y \\ Z^TR^{-1}y \end{bmatrix}$$

$$f(\theta|y) \propto \mathcal{L}(y|\theta)f(\theta)$$

$$K(x_i, x_j) = \varphi(x_i)\varphi(x_j)$$

$$Gini(p) = = 1 - \sum_{k=1}^{K} p_k^2$$

$$(f * g)(t) = \int_{-\infty}^{+\infty} f(\tau)g(t-\tau)d(\tau)$$

ID	PH	EH	Y	MS
In1	307	99	1109	27
In2	278	123	1161	25
In3	297	112	980	31
In4	327	116	1250	32
In…	…	…	…	…

ID	M1	M2	M3	M4	M5	… Mn
In1	C	C	A	C	A	T
In2	A	C	A	C	G	A
In3	A	G	T	G	G	A
In4	C	G	A	G	A	T
In…	…	…	…	…	…	…

Fig. 1 An overview of genomic selection. (a) There are three parts in the genomic selection, including the training dataset, models, and evaluation dataset. The training dataset consists of phenotyping data collected from the field trials and genotyping data tested in the marker lab. The models are trained through two strategies: statistical methods and machine learning. The evaluation dataset is predicted phenotype and genotyping. The materials would be selected according to the predicted phenotyping and then go to field experiments. (b) Summary of models in GS.

The emergence of high-throughput marker testing platforms, such as single-nucleotide polymorphism (SNP) arrays, DNA microarrays, and second- and third-generation sequencing technologies, has enabled breeders to genotype numerous individuals efficiently at low cost, facilitating the widespread adoption of GS in breeding programs[10]. Subsequently, GS has become an increasing focus of research and development in breeding.

With the accumulation of genotypic and phenotypic data, the use of GS for molecular breeding has made significant contributions to breeding within major plant breeding companies, especially for screening DHs and hybrid prediction[11]. Many models continue to be proposed and improved to enhance the accuracy of prediction. Below, we describe various models from statistics and machine learning (ML) and review them in detail (Fig. 1b).

Genomic BLUP (GBLUP) estimation for GS

Before genomic relationships were used for GS, pedigrees were used to predict the breeding values of individual crop plants via BLUP mixed-model equations. It is straightforward to determine phenotypic similarity when the progenies share the same pedigree information, which enhances the efficiency of artificial selection[12]. However, this method cannot be used by breeders when pedigree information is lacking or progenies are derived from the same parents (i.e., a full-sib family). Therefore, scientists explored ways to more accurately assess the kinship of individuals based on their genomic relationships with the advent of DNA molecular markers.

GBLUP were used to predict the breeding values of individual crop plants via BLUP mixed-model equations. BLUP estimation is derived from the linear mixed model and the following mixed-model equation:

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z+\lambda G^{-1} \end{bmatrix}\begin{bmatrix} \beta \\ u \end{bmatrix}=\begin{bmatrix} X^TR^{-1}y \\ Z^TR^{-1}y \end{bmatrix} \quad (1)$$

where, σ_g^2 is the total genetic variance, σ^2 is the random error variance, $\lambda=\frac{\sigma^2}{\sigma_g^2}$, R is the random error variance–covariance coefficient matrix I, and G is the genetic variance–covariance coefficient, which is derived from the estimation of genomic relationships[13,14] as follows:

$$G=\frac{ZZ^T}{2\sum p_i(1-p_i)} \quad (2)$$

where, Z is the matrix of all markers, p_i is the frequency of each marker, and all other parameters are identical. The ability to estimate genomic relationships is a critical measure for improving the prediction accuracy of GBLUP. Therefore, various methods have been introduced for estimating genomic relationships[14,15] including single-step GBLUP (ssGBLUP), which combines pedigree-kinship information with genomic relationship information to estimate genomic relationships[16,17]. Subsequently, additional GBLUP methods have emerged.

In trait-specific marker-derived relationship matrix (TABLUP), another GBLUP method, a trait-specific relationship matrix is built based on the identity by descent (IBD) between both individuals from the locus with the genetic variance in the trait[18]. The equation is as follows:

$$TA_{ij}=\sum_{k=1}^{n}2P_{IBD,ijk}\sigma_{g,k}^2 \quad (3)$$

where, k is the locus, $P_{IBD,ijk}$ is the probability of IBD, i and j are individuals, σ_g^2 is the genetic variance of the trait, and kinship is the sum of values for all loci. Subsequently, Wang et al. proposed two more methods: SUPER BLUP (sBLUP) and compressed BLUP (cBLUP)[19], which were derived from the GWAS (genome-wide association study) theory. In sBLUP, the kinship matrix is designed by selecting significant bins, which are grouped based on the markers associated with traits of interest via GWAS[20]. In cBLUP, all individuals are clustered into several groups according to their genomic relationships. The kinship matrix is the average of the genomic relationships[21]. rrBLUP is a specific model used to implement kinship-based BLUP and SNP-based BLUP. With this method, the genomic relationship is calculated as $K_{RR}=GG^T$, which is equivalent to the SNP-based rrBLUP methods[22,23].

Bayesian estimation for GS

Bayesian ridge regression (BRR)

In BRR, the effect of all SNP markers is assigned to an identical and independent Gaussian prior distribution, and the beta coefficients used to describe the effect are at normal densities: $\beta\sim \mathrm{N}(0,\sigma_\beta^2)$, where the prior distribution of the variance comes from the scaled inverse chi-square distribution, $\sigma_\beta^2\sim\chi^{-2}(\upsilon_\beta,S_\beta)$; the hyperparameters υ_β and S_β will be updated in the model[24,25]. Under this hypothesis, all markers are linked with QTLs. However, in practice, not all markers have a phenotype effect. As a result, other prior distributions have been proposed to solve this problem.

Bayesian LASSO (BL)

The Bayesian form of LASSO infers variants according to the prior density of double-exponential distribution[26,27]. Here, beta coefficients of the effect are at normal densities, $\beta\sim \mathrm{N}(0,\tau_p^2\sigma^2)$. The variance is scaled-inverted chi-square density with degrees of freedom υ and scale S, $\sigma^2\sim\chi^{-2}(\upsilon,S)$. $\tau_j^2\sim \mathrm{Exp}(\lambda)$ is an exponential density with the hyperparameter λ, $\lambda^2\sim Gamma(\alpha_1,\alpha_2)$ is a gamma distribution with two hyperparameters: shape parameter α_1 and rate parameter α_2. There are two other λ prior distributions: $\lambda^2\sim beta(p,\pi)$ and a flat distribution[25,28]. Moreover, Legarra et al. proposed BL2Var, with coefficient τ_j^2: $\tau_j^2\sim \mathrm{IG}\left(\sqrt{\frac{\tilde{\lambda}^2}{\tilde{a}_i^2}},\lambda^2\right)$, and BL1Var, with coefficient τ_j^2: $\tau_j^2\sim \mathrm{IG}\left(\sqrt{\frac{\tilde{\lambda}^2\sigma^2}{\tilde{a}_i^2}},\lambda^2\right)$[29].

BayesA

In BayesA, the variances of each marker effect are different genome-wide, with a scaled inverse chi-square distribution: $\sigma_{\beta_j}^2\sim\chi^{-2}(\upsilon_\beta,S_\beta)$; $\upsilon_\beta=4.012$ and $s_\beta=0.0020$ are the suggested fixed values for the degrees of freedom and scaled parameter, respectively. As a result, a scaled-t density could explain the marginal distribution of marker effects[5]. However, the shrinkage of SNP effects is severely affected by S_β, which should be considered as an unknown parameter with its own prior $S_\beta\sim Gamma(r,s)$; r and s are rate parameter and shape parameter, respectively[30].

BayesC

In BayesC, the variances of all marker effects are identical and independent, with a prior scaled inverse chi-square distribution: $\sigma_\beta^2\sim\chi^{-2}(\upsilon_\beta,S_\beta)$; υ_β and S_β are the degrees of freedom and scaled parameter, respectively. S_β has a prior density: $S_\beta\sim Gamma(r,s)$; r and s are the rate and shape parameters, respectively. The beta coefficients of the effect are normal density: $\beta_j\sim N(0,\sigma_\beta^2)$ with probability π, $\beta_j=0$ with probability $1-\pi$. Additionally, π is the prior density: $beta(a,b)$, when π is suggested as the unknown parameter with a prior density $beta(a,b)$, which is referred to as BayesCπ[30].

BayesB

BayesB combines the hypotheses from BayesA and BayesC. First, the beta coefficient of the effect is normal density: $\beta_j\sim N(0,\sigma_{\beta_j}^2)$ with probability π, and $\beta_j=0$ with probability $1-\pi$. Additionally, π is the prior density: $beta(a,b)$, when π is suggested as the unknown parameter with a prior density $beta(a,b)$, which is referred to as BayesDπ. The variances of each marker effect are different genome-wide, with a scaled inverse chi-square distribution: $\sigma_{\beta_j}^2\sim\chi^{-2}(\upsilon_\beta,S_\beta)$, $S_\beta\sim Gamma(r,s)$; r and s are rate parameter and shape parameter, respectively[5,25].

Several other Bayesian estimations can be used for different prior distributions on marker effects, including BayesU[31], BayesHP, BayesHE[32], BayesR, and emBayesR[33]. Based on the above descriptions, Bayesian theory could provide a series of models for GS given different prior hypotheses. A mixture of prior distributions is used to generate many diverse types of models by combining the different prior hypotheses. The mean and variance of all parameters could be estimated by Markov chain Monte Carlo of Metropolis-Hastings or

Gibbs sampling[34] and the prediction accuracy of these models is comparable to that of other models[35].

Machine learning (ML)

Although the accuracy of phenotypic prediction during GS has improved, it remains challenging to analyze highly complex agronomic traits regulated by numerous genetic loci with minor effects[36]. This makes it difficult to depict the genetic interactions from models via classical BLUP estimation or Bayesian estimation[37]. Furthermore, epistatic effects and imprinting of genetic interactions are common and widely present within biological processes[38]. Consequently, ML methods have been proposed to address the problems arising from genetic interactions using non-linear approaches[39,40]. Much effort has been devoted to developing ML methods to improve the accuracy of GS.

Kernel-based models

A kernel function (also known as 'the kernel trick') could be implemented to transform a vector of low-dimensional space into the inner product of a vector of infinite dimension. ML can be used for non-linear classification by implementing features from a low-dimensional space to be mapped into higher-dimensional feature spaces. Kernel functions show highly improved computational efficiency by specifically calculating functions in low-dimensional spaces:

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \tag{4}$$

where x_i and x_j are features of a low-dimensional space, and φ is the function for mapping x_i and x_j to a higher-dimensional feature space. Several kernel functions could be implemented for the computation of φ, including linear kernel, polynomial kernel, sigmoidal kernel, and Gaussian kernel. These methods effectively capture the non-additive genetic effects among individuals for phenotypic prediction in biological statistics. In quantitative genetics, non-additive genetic effects play a critical role in elaborating the basic theory of genetic networks. High-dimensional genetic effects are rarely computed due to limited storage space and computing efficiency. The results of these models are expressed as an $n \times n$ matrix via the kernel function. Additionally, some hyperparameters need to be trained.

Theoretical aspects of the reproducing kernel Hilbert space (RKHS) mixed model were introduced to study the non-linear relationships of genetic interactions through the kernel function in Hilbert spaces in 2006[41]. In 2008, Gianola & Van Kaam expanded this technique, reproducing a mixed model of kernel Hilbert spaces that parsed epistatic variance between the effects of many genetic loci via a linear mixed model[42]. The linear model of the dual formulation of RKHS is:

$$y = X\beta + Zu + K_h\alpha + e \tag{5}$$

where β is the fixed variable; u, α, and e are random variables with independent distributions; and $u \sim N(0, \sigma_g^2 G)$, $\alpha|h \sim N(0, \sigma_\alpha^2 K_h^{-1})$, and $e \sim N(0, \sigma^2 I)$, respectively. Similarly, the regression estimating equation is:

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z & X^T R^{-1} K_h \\ Z^T R^{-1} X & Z^T R^{-1} Z + \left(\frac{\sigma^2}{\sigma_g^2}\right) G^{-1} & Z^T R^{-1} K_h \\ K_h^T R^{-1} X & K_h^T R^{-1} Z & K_h^T R^{-1} K_h + \left(\frac{\sigma^2}{\sigma_\alpha^2}\right) K_h \end{bmatrix} \begin{bmatrix} \beta \\ u \\ \alpha \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \\ K_h^T R^{-1} y \end{bmatrix} \tag{6}$$

where, $R = I$; K_h is the correlation matrix from the results of the kernel function of features.

Subsequently, a general framework was proposed for the genetic evaluation of RKHS regression for pedigree- or marker-based regressions under any genetic model[43]. Additionally, the Bayesian approach was formulated to solve the unknown parameter from the RKHS mixed model[42,44]. RKHS regression outperformed the linear models in predicting the total genetic values of the body weight of chickens[45].

Support vector machines (SVMs) are another set of efficient methods for non-linear classification through the kernel function[46]. The model is:

$$y = X\beta + b \tag{7}$$

where y is the response variable of the sample, β and b are the coefficients to be solved, and X is the feature variable. In the SVM model, the critical thought is a hyperplane that can separate a high-dimensional space into two parts. To guarantee the hyperplane's uniqueness, the points in a space must be at the maximum distance from this hyperplane, i.e., at the maximum margin hyperplane. The transformed formulation of the optimization is:

$$min \frac{1}{2}\|\beta\| + C\sum_{i=1}^{n} \xi_i \tag{8}$$

$$subject\ to\ y_i\left(\beta^T x_i + b\right) \geq 1 - \xi_i \tag{9}$$

$$\xi_i \geq 0 \tag{10}$$

where ξ_i are slack variables and C is a regularization parameter. To achieve variables of the low-dimensional space to map to the high-dimensional space, a primal-dual method was used to solve optimization problems[47,48]. The dual formulation is:

$$\min \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j \varphi(x_i)^T \varphi(x_j) \tag{11}$$

$$subject\ to\ 0 \leq \alpha_i \leq C \tag{12}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{13}$$

where α_i is the variable parameter of dual formulation, $\varphi(x_i)^T \varphi(x_j) = K(x_i, x_j)$, which would be computed by the kernel function. The kernel function might be a polynomial kernel, sigmoidal kernel, Gaussian kernel, or other kernel type. The parameter β is expressed according to the following primal-dual formulation:

$$\beta = \sum_{i=1}^{n} y_i \alpha_i \varphi(x_i) \tag{14}$$

$$b = \frac{1 - y_i \beta^T \varphi(x_i)}{y_i} \tag{15}$$

Finally, the decision function is:

$$y^* = \sum_{i=1}^{n} y_i \alpha_i K(x_i, x_j) + b \tag{16}$$

Applying SVM to GS for trait prediction is a straightforward process and SVM is indeed a competitive and promising strategy for GS in plant breeding[49,50]. On the whole, the decision function of the SVM depends on only a few support vectors, which is an advantage when classifying a small number of samples. Moreover, the complexity of the calculation is low to avoid a 'dimensionality disaster'.

Ensemble learning

Ensemble learning is a technical framework that is not a single ML algorithm per se. In ensemble learning, a learning task is accomplished by building and combining multiple basic machine models that decide the ultimate results. Ensemble learning can be used for classification, regression problems, and feature selection during data mining. This technique could be applied to GS to predict the phenotypic values of new individuals. Three common ensemble learning frameworks are currently used: bagging (also known as bootstrap aggregation), boosting, and stacking.

Random Forest (RF) is a typical bagging ML method (Fig. 2)[51]. Using this method, M training samples are drawn from the N original sample set via the bootstrap method per round. In the training set, some samples may be drawn multiple times, while some samples may not be drawn even once. K rounds of extraction are carried out to obtain K training sets that are independent of each other. K training sets are constructed using different types of decision tree algorithms for classification or regression. For example, each weak learner could be constructed by classification and regression tree[52] using the Gini index: $Gini(p) = 1 - \sum_{k=1}^{K} p_k^2$.

Theoretically, other ML algorithms could generate new weak learners according to the K training sets. Compared to the decision tree, RF has lower accuracy at the beginning of training, but its performance improves with increasing rounds of training. Therefore, RF is an outstanding method for ML and is starting to be used to analyze genetic networks, generating promising results[53]. RF is a robust learner that reduces noise and overfitting in the GS model training process due to the use of nonparametric measures, but a large amount of data is needed to promote the efficiency of the GS model[54]. Additionally, to analyze additive and epistatic effects and thereby improve the accuracy of GS, RF has been modified by combining it with a linear mixed model and a Bayesian model to construct a mixed RF model and Bayesian additive regression trees (BART)[55].

Light gradient boosting machine (LightGBM): There are two major boosting methods: adaptive boosting (AdaBoost)[56], and gradient boosting (GB)[56]. Gradient boosting decision tree (GBDT) was originally developed from GB using a weak classifier CART decision tree[56]. This tool uses weak classifiers to iteratively train and optimize the model, so that it has good prediction accuracy and easily avoids overfitting. LightGBM[57], extreme gradient boosting (XGBoost)[58], and categorical boosting (CatBoost)[59] are thought to be the three best implementations of GBDT, as they outperformed the basic GBDT algorithm framework. XGBoost is widely applied in the industry, LightGBM has more effective computational efficiency, and CatBoost is thought to have better algorithm accuracy.

LightGBM employs some novel optimizations to accelerate calculations and reduce the need for memory storage and is therefore an improved version of XGBoost as well (Fig. 3)[57]. This method extends the pre-sorted algorithm and histogram-based algorithm to preprocess the features of a dataset to reduce its complexity. A novel sampling algorithm, gradient-based one-side sampling (GOSS), was developed to decrease the size of the dataset without decreasing the accuracy of the model.

In this method, the training results are sorted in descending order based on the gradient. The top A data instances are retained as data subset a, and the remaining data are randomly sampled to yield subset b; the new training dataset consists of subsets a and b (Fig. 3). An exclusive feature bundling algorithm was proposed to decrease the number of features by bundling exclusive features into a single feature as a sparse feature matrix. In addition, unlike other boost models, a leaf-wise decision tree growth strategy is employed to split the nodes, resulting in less loss than the level-wise algorithm. Finally, it is easier to perform multi-threaded optimization and control the complexity of the dataset using the lightGBM model. Yan et al. proposed using lightGBM for F1 prediction for GS[60]. The authors demonstrated that lightGBM had higher accuracy than XGBoost, CatBoost, and rrBLUP when the training dataset excluded parental information. However, the accuracy of lightGBM was lower than that of rrBLUP when parental information was added to the training dataset. Nonetheless, the accuracy of the lightGBM model could be affected by the population structure and kinship of the training population. Additionally, lightGBM takes less time and uses less memory than XGBoost, CatBoost, and rrBLUP.

Deep learning (DL)

Deep learning (DL), a popular research direction in artificial intelligence (AI), has developed rapidly in recent years. A series of excellent models for computer vision have been produced using DL, including multilayer perceptron (MLP), convolutional neural networks (CNN), and recurrent neural networks (RNN). Therefore, some attempts have been made to apply DL to biology to enhance GS and other techniques[61]. Here, we focus on MLP and CNN algorithms.

MLP is a deep neural network mechanism and technology. The simplest MLP is a neural network with a three-layered structure containing an input layer, a hidden layer, and an output layer (Fig. 4)[62]. Multiple hidden layers can be added to the network. The mathematical expressions are as follows:

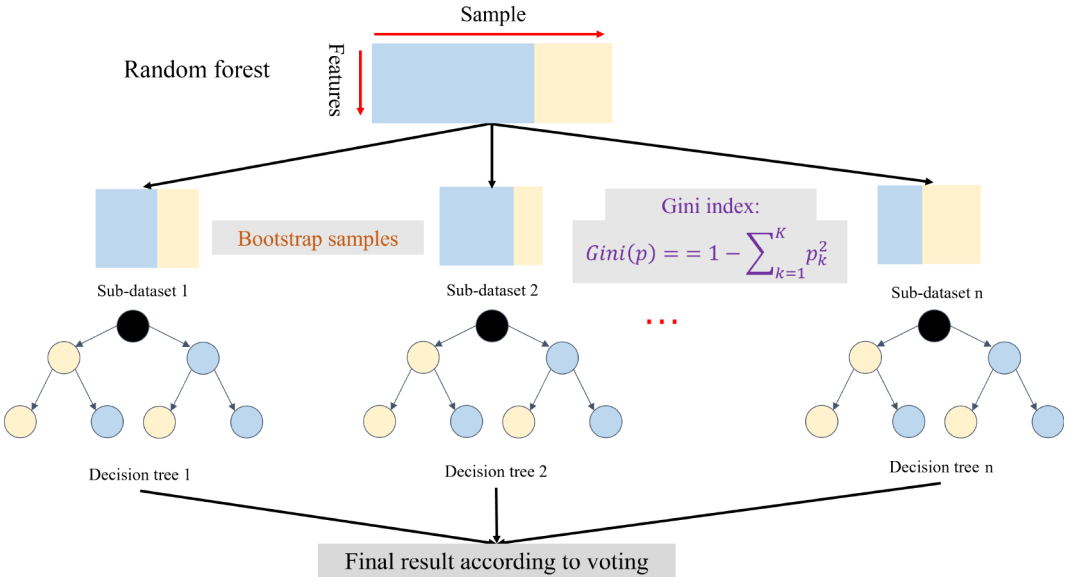


Fig. 2 An overview of the random forest[51]. The random forest includes the bootstrap samples and weak learners based on the decision tree with the Gini algorithm.

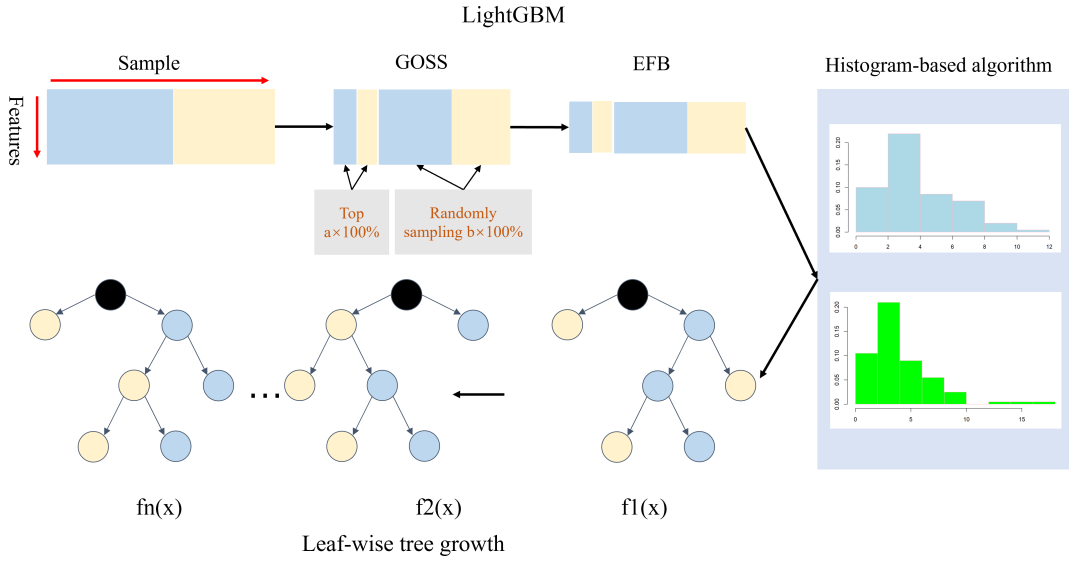


Fig. 3 An overview of LightGBM[57]. LightGBM includes the GOSS, EFB, histogram-based feature selection, and leaf-wise tree growth of the decision tree.

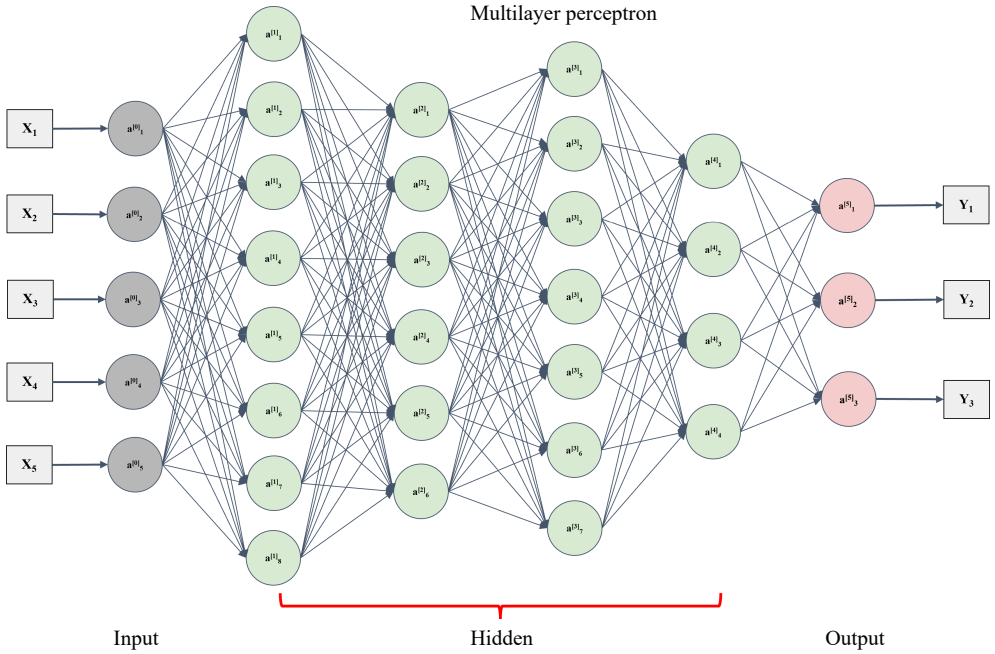


Fig. 4 An architecture of multilayer perceptron[62]. The multilayer perceptron includes one layer (a0 layer) with respect to input data and one layer (a5 layer) with respect to the output. The hidden layer could consist of many layers (from a1 to a4).

$$H^1 = \sigma(XW^1 + b^1) \tag{17}$$

$$O = \sigma(H^1 W^2 + b^2) \tag{18}$$

$$Loss\ function: min\frac{1}{2}(y-O)^2 \tag{19}$$

Here, X is the $n \times o$ matrix for the feature, W^1 is the $o \times p$ matrix for the trained parameters, H^1 is the $n \times p$ matrix for the first hidden layer, b^1 is the $1 \times p$ vector for the bias parameters, W^2 is the $p \times m$ matrix for the training parameters, O is the $n \times m$ matrix for the output variables, b^2 is the $1 \times m$ vector for bias variables, and σ is the activation function. All parameters can be optimized via forward propagation and backpropagation according to the loss function. Bellot et al. used MLP to predict complex human traits and determined that one hidden layer is better than multiple hidden layers[62]. Nevertheless, another report indicates that the optimal number of hidden layers depends on the dataset[63,64]. In general, no more than three hidden layers are normally required in GS projects[65].

CNN plays a crucial role in DL. The predecessors of this tool date back to 1980[66]. Nevertheless, the first formal CNN construction model, which was proposed by LeCun et al. in 1998, contains three types of layers: convolutional layers, pooling layers, and a fully connected layer (Fig. 5)[67]. The authors also described a complete backpropagation optimization algorithm. Obtaining a convolutional kernel from convolutional layers, which focuses only on local features, is a critical step for CNN. The size of the convolution kernel determines the scope of view. The convolution of the kernel and the corresponding areas could be computed as the inner product using the equation $(f \times g)(t) = \int_{-\infty}^{+\infty} f(\tau) g(t-\tau) d(\tau)$. Pooling layers are periodically inserted between successive convolutional layers. The function gradually reduces the spatial size and decreases the features to reduce the need for computing resources by

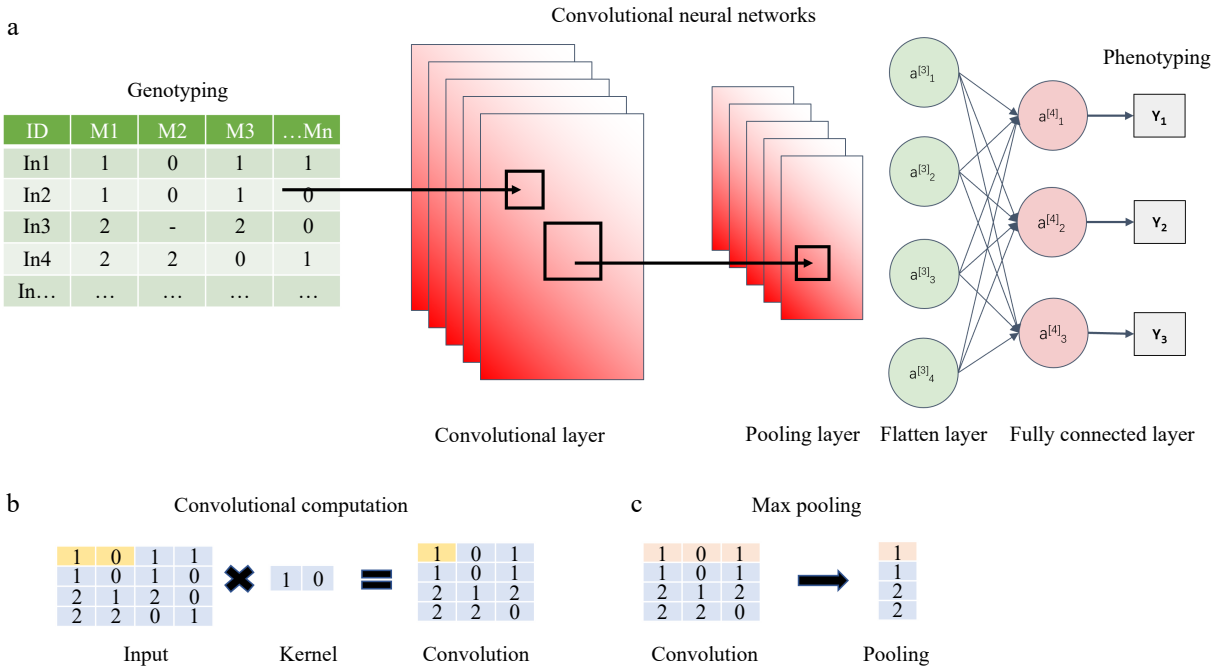


Fig. 5 An architecture of convolutional neural networks[67]. (a) General CNN algorithm, including convolutional layer, pooling layer, and fully connected layer. (b) Explanation of the convolutional computation. (c) Max pooling method.

downsampling without affecting the results. A max-value algorithm is usually used to implement downsampling.

The kernel could be set to different dimensions based on circumstances: 1D, 2D, or 3D, all of which are expressed as a weight parameter matrix to be trained in the CNN. A 1D convolutional kernel could be applied for the SNP feature, as the genotyping of each individual is expressed as a 1D vector based on previous studies, such as DeepGS[68], DLGWAS[69], and DNNGP[70]. While all these models aim to enhance genomic prediction, they exhibit significant differences in their architectural designs and methodologies. DeepGS and DLGWAS both employ CNN structures; however, DLGWAS introduces a dual-stream design that significantly enhances the flexibility of feature processing. In contrast, DNNGP adopts a more conventional approach focused on efficient feature extraction, which may limit its capacity to capture complex data relationships due to its relatively simpler architecture. Meanwhile, there is a new focus on 2D expression, which remains to be tested. The convolutional layer performs dimensionality reduction. An activation function is then used to implement a non-linear approach. All these models are MLP or CNN models that do not use many layers. Nevertheless, these DL models that use convolutional layers were not shown to be more accurate than traditional models based on previous results[64].

Comparison of ML-based GS algorithms and statistical algorithms

ML algorithms have been extensively applied to various large and high-dimensional datasets. These algorithms not only allow for the inclusion of numerous markers from high-throughput sequencing but are also suitable for different types of omics data, such as gene expression data, and functional annotation of proteins. ML models also provide important results, facilitating the identification of markers with the most significant effects on the trait of interest[60]. Therefore, ML is highly flexible for feature processing. The biological process from gene to phenotype is highly complex, including transcription, translation, and gene networks. Traditional linear models cannot accurately capture all of these complex relationships. In contrast, many ML methods can recognize non-linear relationships between genetic markers and phenotypes, offering improved modeling capabilities[40]. However, ML algorithms, such as DL models, require significant computational resources and large training datasets due to the many parameters to be estimated among the potential factors[70]. Additionally, ML models, particularly DL architectures, are often black boxes, making it challenging to interpret marker effects or to understand the biological mechanisms underlying the resulting predictions. It is essential to select the appropriate ML method based on the specific characteristics of the dataset and the objectives of the GS. In general, ML model tuning should be performed to evaluate the performance and suitability of each method for a particular trait and dataset (Table 1).

Application scenarios of GS in crop breeding

Screening of DHs, backcross materials, and selfing materials

Maize DH lines can be produced in a high-throughput manner in the laboratory, greenhouse, or winter nursery via yearly management. However, it is challenging to screen and develop DHs. It is not feasible to plant all DHs in the field due to the excessive workload and low-efficiency selection process. Consequently, GS is a critical step for methods involving DH[71]. Similarly, numerous F2/F3 or BC1/BC2 materials of maize, rice, and wheat are generated on a large scale during breeding, although the genotypes of these

Table 1. Comparison of ML-based GS algorithms and statistical algorithms.

Comparison item	ML-based GS algorithms	Statistical algorithms
Data handling capacity	Process high-dimensional datasets, handle omics data	Limited to traditional markers
Non-linear relationship	Capture non-linear relationships and enhance model performance	Struggle with non-linear relationships
Computational resources	Require significant computational resources	Require fewer resources
Interpretability	Act as black boxes, difficult to interpret	Provide transparent models
Applicability	Offer flexible processing, require tuning	Suit linear relationships

materials are not stable and some loci are segregating. However, phenotypic prediction of these materials still increases the efficiency of the breeding process[72,73].

Hybrid prediction

With the hybrid materials of maize or rice in place, breeders can use this model to predict the performance of different hybrid combinations[60]. The genomic information from the selected parental lines is used to estimate the expected performance of the hybrid progeny for the target traits. These predictions help breeders make informed decisions about which hybrid combinations are likely to display superior performance and should be advanced in the breeding program.

Prediction of simulated progenies of parental lines

During inbred line selection, the two parental lines from maize, rice, or wheat must be selected to perform a single cross. A general rule is that the two parents will generate lines that have improved versions of their phenotypic traits. However, parents that have complementary advantaged genetic backgrounds tend to produce superior progeny. As a result, it is difficult to decide which combination is best when many parental lines are available. In modern breeding, each elite line could be genotyped using high-density markers, and a breeding population containing 300–400 individuals could be simulated by the genetic algorithm based on the markers of the parental lines. Each phenotypic trait of these individuals could be predicted by the trained models[74,75]. A predicted phenotypic value could be obtained for each possible combination from two parental lines. All combinations could then be compared to facilitate decision-making based on the predicted breeding values.

Prediction accuracy

The prediction accuracy for various traits is relatively variable across BLUP estimation models, Bayesian estimation models, and ML[76,77]. The genomic estimated breeding value from GBLUP of the realized (additive) relationship matrix is equivalent to that from the marker-based rrBLUP strategy[22,78]. Additionally, PBLUP is less accurate than GBLUP in most situations, as GBLUP can capture the effects from both within- and between-family genetic variation[79]. There is a growing number of prior distributions and hypotheses based on Bayesian regression models, with varying levels of accuracy: the accuracy is higher when the genetic architecture of the traits complies with the prior distribution[80].

There is a dialectical relationship between GS models and QTLs. For example, BayesB is more accurate when traits are controlled by many genes or major QTLs[5]. Meher et al. reported that the Bayesian alphabet model is better when the traits are controlled by a few genes/QTLs with relatively large effects. To assess this issue in more detail, we used a dataset of 487 wheat individuals with 30,548 markers. The dataset included various types of wheat such as cultivars, landraces, and synthetic hexaploids, and we analyzed several models, including rrBLUP, BRR, BL, BayesA, BayesB, MLP, and CNN (Fig. 6). In some traits, BayesA and BayesB had more accurate predictions than other models when the traits were determined by multiple small-effect QTLs and a few large-effect QTLs[80] (Fig. 6a). However, for some complex traits such as yield, the BayesA and BayesB are not always better (Fig. 6b).

Additionally, during the construction of models for use in GS, data preprocessing has a critical effect on the accuracy of prediction. An appropriate data preprocessing technique pipeline could help ensure the integrity and usefulness of the data, resulting in improved accuracy for GS. This preprocessing could include steps such as screening and imputation of genotyping data and filtering out missing phenotypic data and outliers[81].

The heritability of traits is critical for prediction accuracy as well. There is a close relationship between predictability and heritability. Bayesian regression models are more accurate than BLUP regression for traits with high heritability (Fig. 6a & b), whereas traits with lower heritability should be predicted using BLUP regression models[82]. However, if the heritability of traits is lower, the prediction accuracy can be overestimated. Consequently, GS for traits with low heritability is still quite challenging.

The size of the training population is a crucial factor for model accuracy. In general, the larger the population, the greater the prediction accuracy, but this association weakens after a certain threshold[60,70]. Therefore, the cost of the experiment and prediction accuracy must be balanced. A representative training population of sufficient size could lead to better generalization of a model. Higher accuracy is obtained when there is a very close relationship between the training dataset and the testing dataset, such as when both are part of the same full-sib family. Insufficient or biased training data can result in poor accuracy for GS.

Genotyping provides genetic information in the form of markers or variants, which are linked with the gene of interest and can be used as features in GS models. Theoretically, the prediction accuracy depends on the number of markers used (Fig. 6c & d). Additionally, incorporating genetic features derived from genotyping into the modeling process could potentially improve the prediction accuracy for GS[83], such as PCA, Min-Max Normalization, or others (Fig. 6e & f). However, strategies used for feature processing must be chosen carefully, as their effects on prediction ability are not always as expected (Fig. 6e & f). In practice, markers cannot explain all genetic variation; a haplotype block feature strategy could be a suitable way to explain some portion of additional genetic variation[84]. Consequently, combining different genetic features could improve the accuracy of a model.

Perspectives on GS

In previous sections, we summarized GS using models based on statistical analysis or ML. Much effort has focused on improving the prediction accuracy of all these models; however, breeding traits, especially yield, are highly complex due to the control exerted by gene regulatory networks, and innovative new models must be developed. What is the future direction of GS algorithms?

With the continuous improvement of commercial breeding systems, the types of breeding populations have become increasingly diverse, including many DHs from bi-parental populations and sets of breeding populations or lines derived from different stages of self-crosses in breeding programs and hybrid trials across multiple locations and over multiple years[85,86]. Multi-population joint prediction models are urgently needed to address this issue. With the development of high-throughput sequencing techniques and advances in biotechnology, the cost of processing single samples has decreased, making the production of multi-omics data, including genomics[87], transcriptomics[88], and proteomics[89] data easier and more convenient. The construction and preprocessing of large-scale multi-omics datasets, including screening, filtering, and integration, are challenging using traditional methods[90,91]. With the continued collection, accumulation, and analysis of envirotypic data, enviromics is increasingly being used to explore genotype-by-environment interactions based on spatial and temporal variability at multidimensional scales[85,92]. This analysis provides insights into the environmental drivers of the distribution of elite germplasm, facilitates the screening of breeding materials, and enhances the precise evaluation of plant varieties, ultimately leading to advanced breeding processes.

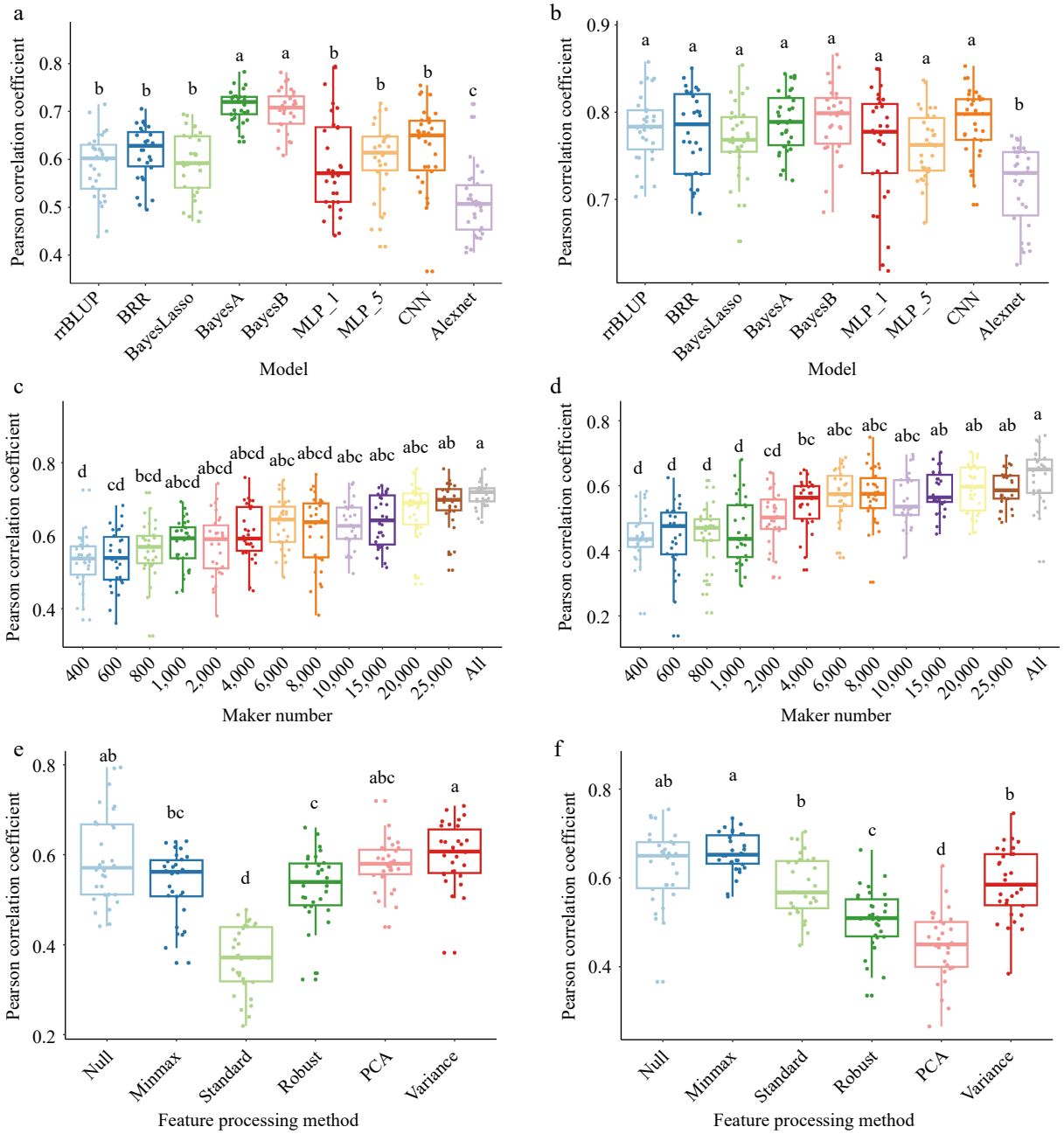


Fig. 6 Comparison of factors on GS models prediction ability. (a) and (b) comparison of nine GS algorithms on wheat based on the Pearson correlation coefficient of the model prediction ability. (a) Plant height with two major QTLs and heritability is 75.7% in 2014 and 76.5% in 2015, (b) yield with five major QTLs and heritability is 70.1% in 2014 and 85.6% in 2015. 'MLP_1' and 'MLP_5' denote the one and five hidden layers in multilayer perceptron algorithm; 'CNN' means the one convolutional layer, one pooling layer, one fully connected layer; 'Alexnet' is based on the Alexnet architecture model. (c) and (d) Impact of marker numbers on prediction accuracy. (c) Thirteen ways of markers set were randomly selected from 30548 markers to validate the prediction accuracy through the BayesA model. (d) Thirteen ways of markers set were randomly selected from 30548 markers to validate the prediction accuracy through the MLP_1 (multilayer perceptron algorithm with one hidden layer) model. (e) and (f) Impact of feature processing on prediction accuracy. (e) Six ways of feature processing were used to validate the prediction accuracy through the MLP_1 model. (f) Six ways of feature processing were used to validate the prediction accuracy through the CNN model. 'Null' is all markers; 'minmax' is min-max scaling; 'standard' is z-score normalization; 'robust' is robust feature processing; 'PCA' is principal component analysis; 'variance' is the variance scaling. Validation of all models is conducted by five-fold cross validation and repeat 30 times. The least significant difference (LSD) is used as the significance test with threshold of 0.05.

Complex quantitative traits, such as yield, are regulated by multiple genes and their interactions. These traits pose significant challenges for current predictive models due to their genetic complexity. Each gene contributes a small effect to the overall phenotype, and these effects accumulate to determine the final trait expression. Consequently, to capture interactions between genetic factors and specific environments, much more novel and innovative models must be explored. Three specific models for genomic selection (GS) based on CNN architecture have been reported: DeepGS[68], DLGWAS[69], and DNNGP[70]. These methods have proven to be competitive with others and outperform traditional methods in some respects due to their ability to handle high-dimensional data. We believe that AI holds tremendous potential applications for GS. Some researchers have started to use RNA-seq

data to enhance the efficiency of selection by integrating gene expression data into GS models using kernel-based methods, which have been used to capture complex genetic interactions and non-linear relationships between genetic variants and phenotypic traits in animals[93].

However, applying these models directly for GS might not be straightforward, since they are primarily designed for computer vision (CV), speech recognition, and natural language processing (NLP), especially Bidirectional Encoder Representations from Transformers (BERT)[94] and Generative Pre-trained Transformer (GPT) based on transformer architecture models. Genomic data are different from text data, making it necessary to preprocess and represent the genomic sequences in a format suitable for the models. This may involve encoding variants, genomic regions, or other relevant genetic features appropriately[70,83]. Directly applying ML models is not better than Bayesian models and rrBLUP, like AlexNet, is not very well suited for GS applications[63] (Fig. 6a & b). It is important to handle the sequential nature of genetic data and to ensure that the representations capture the relevant information.

Other technologies for applying these models to GS could also be considered: the underlying transformer architecture and the transfer learning principles they employ could be adapted for GS with the appropriate modifications. DeepCCR[95] improves contextual comprehension by integrating BiLSTM layers, which are particularly beneficial for the interpretation of sequential data. In comparison, SoyDNGP[96] employs a three-dimensional input matrix, enabling it to capture intricate genotypic variations and provide richer feature density than the one-dimensional vectors used in other models. GPformer[97] utilizes innovative attention mechanisms to enhance the representation of SNP relationships, thereby enhancing predictive accuracy. Meanwhile, Dual-Extraction Modeling (DEM)[98] stands out with its dual-extraction mechanism, effectively integrating multi-omics data and improving performance through noise reduction and enhanced feature separability. Collectively, the advancements exemplified by GPformer, SoyDNGP, and DEM reflect a trend toward developing more complex and integrated processing architectures that effectively address the multifaceted challenges of genomic prediction. Pre-training the models on large-scale genomic data or related tasks could be beneficial for adapting the pre-training process to GS, capturing meaningful patterns from the pre-training data, and assisting in downstream GS tasks. After pre-training, it is essential to fine-tune the models for specific GS tasks. The fine-tuning process helps adapt the models to specific tasks, and the relationships between genetic variants and performance can be deduced by incorporating domain-specific knowledge into the models. This could be achieved by incorporating prior knowledge about the biology and genetic pathways of the organism, or by using specialized loss functions that account for the specific requirements of GS[92].

Breeding 4.0 reflects the future ability to combine any known alleles into optimal combinations, potentially revolutionizing the field. This stage of agriculture will rely on the development of highly advanced genetic manipulation technologies that can be used to obtain ideal genetic combinations for specific traits. In future breeding pipelines, large-scale multimodal data will be created and developed, such as DNA/protein sequences; text annotation of multi-omics data; images, audio files, and video files from phenomic analysis; sensor readings; and telemetry data[99]. Additionally, numerous public databases, data-sharing platforms, and research papers will be channeled into the pipelines. We believe that GS can serve as a bridge from AI to Breeding 4.0 with innovations in ML, NLP, CV, and other novel technologies. Such enhancements will extend the use of AI in breeding programs.

Table 2. Summary of the performance between the ML and traditional methods.

Crop	Population size	Marker no.	Performance	Ref.
Wheat	2,374	39,758	GBLUP ≥ MLP	[63]
Wheat	250	12,083	GBLUP ≥ MLP	[63]
Wheat	693, 670, 807	15,744	GBLUP ≥ MLP	[63]
Maize	309	158,281	GBLUP ≥ MLP	[63]
Wheat	767, 775, 964, 980, 945, 1,145	2,038	GBLUP ≥ MLP ≈ SVM	[64]
Maize	2,267	19,465	MLP > Lasso	[100]
Maize	4,328	564,692	GBLUP ≈ BayesR ≈ SVM	[49]
Barley	400	50,000	Transformer ≈ BLUP	[83]
Maize	8,652	32,559	LightGBM > rrBLUP	[60]
Wheat	2,000	33,709	LightGBM ≈ DNNGP > GBLUP	[70]
Maize	1,404	6,730,418	SVR ≈ DNNGP > GBLUP	[70]
Wheat	599	1,447	SVR ≈ DNNGP > GBLUP	[70]

In summary, the above factors could enhance the prediction accuracy of GS models. These advanced analytical methods can handle big genomic data, identify complex patterns, and enhance prediction accuracy. As computational tools continue to improve, they could enhance the efficiency and effectiveness of genomic prediction models. By considering a wider range of molecular information, researchers can gain a deeper understanding of the biological mechanisms underlying traits and develop more accurate prediction models for GS.

Conclusions

Many innovative models based on ML are applied to GS to improve the efficiency of breeding. However, not all ML methods significantly enhance prediction accuracy compared to traditional methods, such as BLUP estimation and Bayesian estimation. All of these methods are useful under the appropriate scenarios, and there is currently no unified strategy for GS. Consequently, all types of models for enhancing the accuracy of GS should be considered on a case-by-case basis based on previous studies (Table 2, Fig. 6a & b). Compared to ML, the development of GS models is in its early stages, but GS will serve as a tool for enhancing breeding programs. AI technologies can analyze vast repositories of breeding data, scientific literature, and expert knowledge to discover new patterns, relationships, and insights. These findings can be shared and transferred to breeders, scientists, and stakeholders, fostering collaboration, innovation, and accelerated progress in the breeding community. AI will play a crucial role in accelerating the transition to Breeding 4.0 by facilitating the creation of large-scale multimodal datasets, complex predictive modeling, and precise decision-making processes in crop breeding programs.

Author contributions

The authors confirm contribution to the paper as follows: literature review and manuscript writing: Zhang D; manuscript editing: Yang F, Li J, Wang K; manuscript review: Yang F, Li J, Liu Z, Han Y, Zhang Q, Pan S, Zhao X; project supervision: Wang K. All authors reviewed the results and approved the final version of the manuscript.

Data availability

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Acknowledgments

The work was funded by the Key Research and Development Program of Jiangsu Province, China (BE2022337).

Conflict of interest

The authors declare that they have no conflict of interest.

Dates

Received 15 December 2024; Revised 28 February 2025; Accepted 10 March 2025; Published online 9 April 2025

References

1. Wallace JG, Rodgers-Melnick E, Buckler ES. 2018. On the road to Breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Annual Review of Genetics* 52:421–44
2. Jonas E, de Koning DJ. 2013. Does genomic selection have a future in plant breeding? *Trends in Biotechnology* 31:497–504
3. Lande R, Thompson R. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743–56
4. Ribaut JM, Hoisington D. 1998. Marker-assisted selection: new tools and strategies. *Trends in Plant Science* 3:236–39
5. Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–29
6. Ertiro BT, Zeleke H, Friesen D, Blummel M, Twumasi-Afriyie S. 2013. Relationship between the performance of parental inbred lines and hybrids for food-feed traits in maize (*Zea mays* L.) in Ethiopia. *Field Crops Research* 153:86–93
7. Meuwissen THE, Goddard ME. 1996. The use of marker haplotypes in animal breeding schemes. *Genetics Selection Evolution* 28:161
8. Whittaker JC, Thompson R, Denham MC. 2000. Marker-assisted selection using ridge regression. *Genetics Research* 75:249–52
9. Heffner EL, Sorrells ME, Jannink JL. 2009. Genomic selection for crop improvement. *Crop Science* 49:1–12
10. Eathington SR, Crosbie TM, Edwards MD, Reiter RS, Bull JK. 2007. Molecular markers in a commercial breeding program. *Crop Science* 47:S-154–S-163
11. Li J, Cheng D, Guo S, Chen C, Wang Y, et al. 2023. Genome-wide association and genomic prediction for resistance to southern corn rust in DH and testcross populations. *Frontiers in Plant Science* 14:1109116
12. Hayes BJ, Daetwyler HD, Bowman P, Moser G, Tier B, et al. 2009. Accuracy of genomic selection: comparing theory and results. *Association for the Advancement of Animal Breeding and Genetics* 18:34–37
13. VanRaden PM. 2007. Genomic measures of relationship and inbreeding. *INTERBULL Bulletin* 37:33–36
14. VanRaden PM. 2008. Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91:4414–23
15. Astle W, Balding DJ. 2009. Population structure and cryptic relatedness in genetic association studies. *Statistical Science* 24:451–71
16. Aguilar I, Misztal I, Legarra A, Tsuruta S. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *Journal of Animal breeding and Genetics* 128:422–28
17. Legarra A, Christensen OF, Aguilar I, Misztal I. 2014. Single Step, a general approach for genomic selection. *Livestock Science* 166:54–65
18. Zhang Z, Liu J, Ding X, Bijma P, de Koning DJ, Zhang Q. 2010. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS One* 5:e12648
19. Wang J, Zhou Z, Zhang Z, Li H, Liu D, et al. 2018. Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity* 121:648–62
20. Wang Q, Tian F, Pan Y, Buckler ES, Zhang Z. 2014. A SUPER powerful method for genome wide association study. *PLoS One* 9:e107684
21. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, et al. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* 42:355–60
22. Endelman JB. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome* 4:250–55
23. Lorenzana RE, Bernardo R. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics* 120:151–61
24. De Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193:327–45
25. Pérez P, de los Campos G. 2014. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198:483–95
26. Park T, Casella G. 2008. The Bayesian Lasso. *Journal of the American Statistical Association* 103:681–86
27. De Los Campos G, Naya H, Gianola D, Crossa J, Legarra A, et al. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182:375–85
28. Mutshinda CM, Sillanpää MJ. 2010. Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. *Genetics* 186:1067–75
29. Legarra A, Robert-Granié C, Croiseau P, Guillaume F, Fritz S. 2011. Improved Lasso for genomic selection. *Genetics Research* 93:77–87
30. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. 2011. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186
31. Pong-Wong R, Woolliams JA. Bayes U: a genomic prediction method based on the horseshoe prior. *Proc. 10th World Congress of Genetics Applied to Livestock Production, Vancouver, BC. Canada, 2014*. 3 pp
32. Shi S, Li X, Fang L, Liu A, Su G, et al. 2021. Genomic prediction using Bayesian regression models with global–local prior. *Frontiers in Genetics* 12:628205
33. Wang T, Chen YPP, Bowman PJ, Goddard ME, Hayes BJ. 2016. A hybrid expectation maximisation and MCMC sampling algorithm to implement Bayesian mixture model based genomic prediction and QTL mapping. *BMC Genomics* 17:744
34. Cheng H, Qu L, Garrick DJ, Fernando RL. 2015. A fast and efficient Gibbs sampler for BayesB in whole-genome analyses. *Genetics Selection Evolution* 47:80
35. Azevedo CF, de Resende MDV, Fonseca e Silva F, Viana JMS, Valente MSF, et al. 2015. Ridge, Lasso and Bayesian additive-dominance genomic models. *BMC Genetics* 16:105
36. Vieira IC, Dos Santos JPR, Pires LPM, Lima BM, Gonçalves FMA, et al. 2017. Assessing non-additive effects in GBLUP model. *Genetics and Molecular Research* 16:gmr16029632
37. Piepho HP, Möhring J, Melchinger AE, Büchse A. 2008. BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161:209–28
38. Zuk O, Hechter E, Sunyaev SR, Lander ES. 2012. The mystery of missing heritability: genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America* 109:1193–98
39. Ma C, Xin M, Feldmann KA, Wang X. 2014. Machine learning–based differential network analysis: a study of stress-responsive transcriptomes in *Arabidopsis*. *The Plant Cell* 26:520–37
40. Abdollahi-Arpanahi R, Gianola D, Peñagaricano F. 2020. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution* 52:12
41. Gianola D, Fernando RL, Stella A. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173:1761–76
42. Gianola D, Van Kaam JBCHM. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178:2289–303
43. De Los Campos G, Gianola D, Rosa GJM. 2009. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of Animal Science* 87:1883–87
44. De los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics Research* 92:295–308
45. Long N, Gianola D, Rosa GJM, Weigel KA, Kranis A, González-Recio O. 2010. Radial basis function regression methods for predicting quantitative traits using SNP markers. *Genetics Research* 92:209–25

46. Cortes C, Vapnik V. 1995. Support-vector networks. *Machine Learning* 20:273−97
47. Chang CC, Lin CJ. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27
48. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. 1998. Support vector machines. *IEE Intelligent Systems and their Applications* 13:18−28
49. Zhao W, Lai X, Liu D, Zhang Z, Ma P, et al. 2020. Applications of support vector machine in genomic prediction in pig and maize populations. *Frontiers in Genetics* 11:598318
50. Targhi MVA, Jafarabadi GA, Aminafshar M, Kashan NEJ. 2019. Comparison of non-parametric methods in genomic evaluation of discrete traits. *Gene Reports* 15:100379
51. Breiman L. 2001. Random forests. *Machine Learning* 45:5−32
52. Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. *Classification and regression trees*. New York: Chapman and Hall/CRC. 368 pp. doi: 10.1201/9781315139470
53. Naderi S, Yin T, König S. 2016. Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. *Journal of Dairy Science* 99:7261−73
54. Sarkar RK, Rao AR, Meher PK, Nepolean T, Mohapatra T. 2015. Evaluation of random forest regression for prediction of breeding value from genomewide SNPs. *Journal of Genetics* 94:187−92
55. Waldmann P. 2016. Genome-wide prediction using Bayesian additive regression trees. *Genetics Selection Evolution* 48:42
56. Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 29:1189−232
57. Ke G, Meng Q, Finley T, Wang T, Chen W, et al. 2017. LightGBM: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 30:3149−57
58. Chen T, Guestrin C. 2016. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, 2016*, San Francisco, California, USA: Association for Computing Machinery. pp. 785–94. doi: 10.1145/2939672.293978
59. Dorogush AV, Ershov V, Gulin A. 2018. CatBoost: gradient boosting with categorical features support. *arXiv* 00:1810.11363
60. Yan J, Xu Y, Cheng Q, Jiang S, Wang Q, et al. 2021. LightGBM: accelerated genomically designed crop breeding through ensemble learning. *Genome Biology* 22:271
61. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, et al. 2019. A primer on deep learning in genomics. *Nature Genetics* 51:12−18
62. Bellot P, De Los Campos G, Pérez-Enciso M. 2018. Can deep learning improve genomic prediction of complex human traits? *Genetics* 210:809−19
63. Montesinos-López A, Montesinos-López OA, Gianola D, Crossa J, Hernández-Suárez CM. 2018. Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3* 8:3813−28
64. Montesinos-López OA, Martín-Vallejo J, Crossa J, Gianola D, Hernández-Suárez CM, et al. 2019. A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding. *G3* 9:601−18
65. Pérez-Enciso M, Zingaretti LM. 2019. A guide on deep learning for complex trait genomic prediction. *Genes* 10:553
66. Fukushima K. 1980. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36:193−202
67. Lecun Y, Bottou L, Bengio Y, Haffner P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86:2278−324
68. Ma W, Qiu Z, Song J, Li J, Cheng Q, et al. 2018. A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* 248:1307−18
69. Liu Y, Wang D, He F, Wang J, Joshi T, et al. 2019. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Frontiers in Genetics* 10:1091
70. Wang K, Abid MA, Rasheed A, Crossa J, Hearne S, et al. 2023. DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Molecular Plant* 16:279−93
71. Wang N, Wang H, Zhang A, Liu Y, Yu D, et al. 2020. Genomic prediction across years in a maize doubled haploid breeding program to accelerate early-stage testcross testing. *Theoretical and Applied Genetics* 133:2869−79
72. Juliana P, Singh RP, Braun HJ, Huerta-Espino J, Crespo-Herrera L, et al. 2020. Genomic selection for grain yield in the CIMMYT wheat breeding program—status and perspectives. *Frontiers in Plant Science* 11:564183
73. Tessema BB, Liu H, Sørensen AC, Andersen JR, Jensen J. 2020. Strategies using genomic selection to increase genetic gain in breeding programs for wheat. *Frontiers in Genetics* 11:578123
74. Chung PY, Liao CT. 2020. Identification of superior parental lines for biparental crossing via genomic prediction. *PLoS One* 15:e0243159
75. Chung PY, Liao CT. 2022. Selection of parental lines for plant breeding *via* genomic prediction. *Frontiers in Plant Science* 13:934767
76. Sun X, Qu L, Garrick DJ, Dekkers JCM, Fernando RL. 2012. A fast EM algorithm for BayesA-like prediction of genomic breeding values. *PLoS One* 7:e49157
77. Jiang S, Cheng Q, Yan J, Fu R, Wang X. 2020. Genome optimization for improvement of maize breeding. *Theoretical and Applied Genetics* 133:1491−502
78. Hayes BJ, Visscher PM, Goddard ME. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research* 91:47−60
79. Joshi R, Skaarud A, Alvarez AT, Moen T, Ødegård J. 2021. Bayesian genomic models boost prediction accuracy for survival to *Streptococcus agalactiae* infection in Nile tilapia (*Oreochromus nilioticus*). *Genetics Selection Evolution* 53:37
80. Meher PK, Rustgi S, Kumar A. 2022. Performance of Bayesian and BLUP alphabets for genomic prediction: analysis, comparison and results. *Heredity* 128:519−30
81. Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, et al. 2014. Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* 15:740
82. Jia Z. 2017. Controlling the overfitting of heritability in genomic selection through cross validation. *Scientific Reports* 7:13678
83. Jubair S, Tucker JR, Henderson N, Hiebert CW, Badea A, et al. 2021. GPTransformer: a transformer-based deep learning method for predicting Fusarium related traits in barley. *Frontiers in Plant Science* 12:761402
84. Zhang H, Wang X, Pan Q, Li P, Liu Y, et al. 2019. QTG-Seq accelerates QTL fine mapping through QTL partitioning and whole-genome sequencing of bulked segregant samples. *Molecular Plant* 12:426−37
85. Crossa J, Fritsche-Neto R, Montesinos-Lopez OA, Costa-Neto G, Dreisigacker S, et al. 2021. The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. *Frontiers in Genetics* 12:651480
86. Weyen J. 2021. Applications of doubled haploids in plant breeding and applied research. In *Doubled Haploid Technology*, ed. Segui-Simarro JM. New York, NY: Humana. Volume 2287. pp. 23–39. doi: 10.1007/978-1-0716-1315-3_2
87. Wang N, Yuan Y, Wang H, Yu D, Liu Y, et al. 2020. Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Scientific Reports* 10:16308
88. Rich-Griffin C, Stechemesser A, Finch J, Lucas E, Ott S, et al. 2020. Single-cell transcriptomics: a high-resolution avenue for plant functional genomics. *Trends in Plant Science* 25:186−97
89. Liu Y, Lu S, Liu K, Wang S, Huang L, et al. 2019. Proteomics: a powerful tool to study plant responses to biotic stress. *Plant Methods* 15:135
90. Jamil IN, Remali J, Azizan KA, Nor Muhammad NA, Arita M, et al. 2020. Systematic Multi-Omics Integration (MOI) approach in plant systems biology. *Frontiers in Plant Science* 11:944
91. Khaki S, Khalilzadeh Z, Wang L. 2020. Predicting yield performance of parents in plant breeding: a neural collaborative filtering approach. *PLoS One* 15:e0233382
92. Harfouche AL, Jacobson DA, Kainer D, Romero JC, Harfouche AH, et al. 2019. Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. *Trends in Biotechnology* 37:1217−35

93. Liang M, An B, Chang T, Deng T, Du L, et al. 2022. Incorporating kernelized multi-omics data improves the accuracy of genomic prediction. *Journal of Animal Science and Biotechnology* 13:103
94. Devlin J, Chang MW, Lee K, Toutanova K. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv* 00:1810.04805
95. Ma X, Wang H, Wu S, Han B, Cui D, et al. 2024. DeepCCR: large-scale genomics-based deep learning method for improving rice breeding. *Plant Biotechnology Journal* 22:2691–93
96. Gao P, Zhao H, Luo Z, Lin Y, Feng W, et al. 2023. SoyDNGP: a web-accessible deep learning framework for genomic prediction in soybean breeding. *Briefings in Bioinformatics* 24:bbad349
97. Wu C, Zhang Y, Ying Z, Li L, Wang J, et al. 2023. A transformer-based genomic prediction method fused with knowledge-guided module. *Briefings in Bioinformatics* 25:bbad438
98. Ren Y, Wu C, Zhou H, Hu X, Miao Z. 2024. Dual-extraction modeling: a multi-modal deep-learning architecture for phenotypic prediction and functional gene mining of complex traits. *Plant Communications* 5:101002
99. Yan J, Wang X. 2023. Machine learning bridges omics sciences and plant breeding. *Trends in Plant Science* 28:199–210
100. Khaki S, Wang L. 2019. Crop yield prediction using deep neural networks. *Frontiers in Plant Science* 10:621