CellPress
OPEN ACCESS

Trends in
Plant Science

Feature Review

Expanding genomic prediction in plant breeding: harnessing big data, machine learning, and advanced software

José Crossa [1,2], Johannes W.R. Martini [3], Paolo Vitale [1], Paulino Pérez-Rodríguez [2], Germano Costa-Neto [4], Roberto Fritsche-Neto [5], Daniel Runcie [6], Jaime Cuevas [7], Fernando Toledo [1], H. Li [1], Pasquale De Vita [8], Guillermo Gerard [1], Susanne Dreisigacker [1], Leonardo Crespo-Herrera [1], Carolina Saint Pierre [1], Alison Bentley [9], Morten Lillemo [10], Rodomiro Ortiz [11], Osval A. Montesinos-López [12,*], and Abelardo Montesinos-López [13,*]

With growing evidence that genomic selection (GS) improves genetic gains in plant breeding, it is timely to review the key factors that improve its efficiency. In this feature review, we focus on the statistical machine learning (ML) methods and software that are democratizing GS methodology. We outline the principles of genomic-enabled prediction and discuss how statistical ML tools enhance GS efficiency with big data. Additionally, we examine various statistical ML tools developed in recent years for predicting traits across continuous, binary, categorical, and count phenotypes. We highlight the unique advantages of deep learning (DL) models used in genomic prediction (GP). Finally, we review software developed to democratize the use of GP models and recent data management tools that support the adoption of GS methodology.

Highlights

Statistical machine learning (ML) methods applied to big data and available software have a fundamental role in the democratization of genomic selection (GS) methodology.

Principles behind genomic-enabled prediction and statistical ML tools can significantly increase the efficiency of the GS methodology under big data.

Deep learning (DL) models and methods have been implemented in the context of genomic predictions (GPs), emphasizing the power of this special type of statistical ML tool.

There has been an intense and prolific development of software to be used for the GP models. We provided a brief overview of data management tools generated in recent years to promote the democratization of GS methodology.

Predictive plant breeding: a new frontier of genomic innovation

The practical application of genomic tools in plant breeding has opened up a revolutionary era in agriculture, with fundamental changes in how we approach crop improvement. Today, low-cost and high-throughput genotyping technologies allow breeding programs to generate a vast amount of genomic data to support selection decisions and methods such as genome-wide association studies (GWAS) and **genomic prediction (GP)** (see Glossary) [1] which have become routine in many breeding organizations [2]. Genotypic and phenotypic data characterizing germplasm under varying growing conditions have been accumulated over the years. Moreover, high-throughput phenotyping methods have augmented the number of traits considered, and the extent to which – and resolution at which – phenotypes are measured. With these developments, information technological aspects of breeding, such as data storage and the design of analytical pipelines for knowledge extraction, have moved into the spotlight. Analytical tools need to be able to translate collected data into knowledge, breeding decisions, and ultimately increase genetic gain. A breeding organization's ability to leverage information technology will be a key factor in the competition for the most efficient breeding pipelines.

One of the most compelling applications of GP, especially through the integration of multi-omics approaches, is tackling agriculturally important traits characterized by a complex genetic architecture and low heritability, such as 'yield per hectare'. These traits are significantly influenced

[1]International Maize and Wheat Improvement Center (CIMMYT), Carretera México – Veracruz Km. 45, El Batán, CP 56237, Texcoco, Edo. de México, Mexico
[2]Colegio de Postgraduados, Montecillos, Edo. de México CP 56230, Mexico
[3]Aardevo B.V., Nagele, The Netherlands
[4]Cornell University Ithaca, Ithaca, NY, USA
[5]Louisiana State University, College of Agriculture, Baton Rouge, LA, USA
[6]Department of Plant Sciences at the University of California, Davis, CA, USA
[7]Universidad de Quintana Roo, Chetumal, Quintana Roo, 77019, Mexico

by the genotypic–environment interaction, making traditional breeding approaches less effective in achieving substantial genetic gain.

Conventionally, plant breeding decisions are based on phenotypic observations, and thus often require several growing seasons to evaluate selection candidates before selecting them as the parents of a new generation. In addition, statistical methods such as linear regression models were commonly used due to their simplicity and reliance on well-defined, often linear relationships between variables. These methods, grounded in specific data assumptions, were effective for well-defined problems with linear relationships for traits. GP, also known as **genomic selection (GS)**, marks a paradigm shift leveraging genomic information to predict the performance of individuals for specific traits with statistical **machine learning (ML)** methods. ML includes flexibility and non-linear methods that adapt well to complex datasets without strict assumptions. This approach accelerates the breeding process by shortening the evaluation step, allowing for the selection of individuals at early stages without the need for detailed phenotyping before making selection decisions. Provided that prediction models are predicting genetic merit at sufficient **prediction accuracy (PA)**, this approach has the potential to accelerate the realized genetic gain per time to unprecedented levels. Research questions around predictive breeding include the definition of use-cases for predictions in the breeding scheme, the design, type, and updating of training data, the general data workflows, and the corresponding statistical methods to use. **Deep learning (DL)**, a subset of ML, uses multilayer neural networks to capture intricate patterns in large datasets. Each method suits different data complexities and goals in GP. ML and DL terms are sometimes used interchangeably, especially since DL is technically a type of ML. However, they differ in complexity, computational demand, and suitability for various data types. In GP, choosing between these methods often depends on the nature of the data and prediction goals.

When implementing predictive breeding approaches, breeding organizations cannot discuss which statistical ML methods to use without a conversation on the required resources for data curation, data management, computational power, prediction software, etc. A particular challenge for many breeders and scientists has been a lack of user-friendly software. Consequently, various software packages have surfaced in the last decade to address the diverse needs of researchers and breeders, reflecting the increasing importance of accessible and sophisticated solutions in this domain. Popular tools – such as **genome association and prediction integrated tool (GAPIT)** [3,4], **trait analysis by association, evolution, and linkage (TASSEL)** [5], and **genome-wide complex trait analysis (GCTA)** [6] – facilitate tasks ranging from data quality control to GWAS. Moreover, specialized software like Beagle [7] and BGLR [8] incorporates advanced statistical and ML algorithms for GP. The availability of open-source software ensures accessibility and collaborative development, fostering a dynamic ecosystem for genomic research in plant breeding.

In this comprehensive review, our goal is to delve into the current landscape of applications of statistical ML models in plant breeding. Our objectives encompass five main facets. (i) A general overview of the principles underlying GP and a guide to areas under development. (ii) A discussion of statistical ML methods for non-Gaussian distributed traits, including continuous, binary, categorical, and count-based, and an examination of how these methods play a crucial role in optimizing the selection process within the GS methodology. (iii) DL models in GP: we conduct a thorough survey of DL models implemented in the specific context of GP, and explore the unique contributions and potential advantages of DL techniques in this field. (iv) Software development in GP: we review the existing software developed for GP, and highlight the role of software in democratizing the GP methodology, making it more accessible and user-friendly. (v) Data management tools for **data democratization**: we provide a concise overview of recent advancements in data management tools, and emphasize the role of these tools in promoting the democratization of GS methodologies, ensuring their wider

[8]Research Center for Cereal and Industrial Crops (CREA-CI), CREA – Council for Agricultural Research and Economics, Foggia, Italy
[9]Australian National University, Research School of Biology, Canberra, Australia
[10]Norwegian University of Life Science (NMBU), Department of Plant Science, Ås, Norway
[11]Department of Plant Breeding, Swedish University of Agricultural Sciences (SLU), P.O. Box 190 Sundsvagen 10, SE 23422 Lomma, Sweden
[12]Facultad de Telemática, Universidad de Colima, C. P. 28040, Edo. de Colima, Mexico
[13]Departamento de Matemáticas, Centro Universitario de Ciencias Exactas e Ingenierías (CUCEI), Universidad de Guadalajara, 44430, Guadalajara, Jalisco, Mexico

*Correspondence:
osval78t@gmail.com (O.A. Montesinos-López) and
aml_uach2004@hotmail.com
(A. Montesinos-López).

accessibility and adoption. By addressing these objectives, we offer a comprehensive and up-to-date perspective on the diverse applications and advancements within the intersection of statistical ML and genomic-enabled prediction for plant breeding.

General principles of genomic prediction models: exploring GBLUP, rrBLUP, and LASSO

The relationship between genotype and phenotype is intricate, with millions of genetic variants contributing to the phenotypic variation observed in plants. Statistical ML methods offer a systematic and data-driven approach for disentangling this complexity. ML methods are based on mathematical models in which a rough functional relationship between predictor variables and a dependent variable is set when defining the statistical model. The 'learning' step means specifying parameters that have not been fully determined in the initial model. Instead, their values are chosen such that the model describes training data. A training set consists of measured values for the predictor variables – for instance, single-nucleotide polymorphisms (SNPs) – and the response variable (e.g., 'yield per hectare').

The relatively robust, most basic, and widely used method for GP is a linear model in which genomic markers are assumed to have an additive contribution to the phenotypic observation. The parameters that need to be specified are the marker effects that are determined by fitting the phenotypic observations best, subject to some side conditions. Having specified the required parameters, that is having 'learned' the genotype–phenotype relationship, phenotypes can be predicted from genotypic data. Techniques such as ridge regression – which is analogous to the genomic **best linear unbiased prediction (BLUP)** (**GBLUP**), the Bayesian alphabet [9,10], or the least absolute shrinkage and selection operator (LASSO) [11,12] regression – fall into this class of linear models, among others. Linear models and other methods, such as the ensemble learning random forest, and support vector machines have proved effective in handling high-dimensional genomic data and capturing the subtle relationships between genotypes and phenotypes.

GBLUP

The genetic architecture of agriculturally important traits can range from being monogenic (with a qualitative distribution given by either possessing the relevant allele or not) to highly polygenic with quantitative distribution. For example, think of disease resistances that may be determined by a single resistance gene, by contrast with the usually very quantitative trait 'yield per hectare'. For traits of simple genetic architecture with only a few genes involved, linkage studies on specific populations allow the breeder to identify the underlying alleles. This information can be used to fix the positive alleles in the population, or at least to establish dedicated marker-assisted selection efforts. This scenario is, therefore, not a typical use case for GP. GP addresses the multigenic genetic architecture.

As mentioned earlier, the standard reference model is a linear model:

$$\boldsymbol{y} = \mu + \boldsymbol{M\beta} + \epsilon \qquad (1)$$

In the simplest model with one phenotypic evaluation of each of the n genotypes in the training set, $\boldsymbol{y}$ is an $nx1$ vector of phenotypic records, μ a constant $nx1$ vector providing the overall mean, and $\boldsymbol{M}$ an nxp marker matrix providing the allelic states of the n individuals with respect to the p genetic loci used in the model. In the case of the use of SNPs, the states can be coded as 0, 1, and 2, counting a reference allele. The $px1$ vector $\boldsymbol{\beta}$ of allele effects is the central component that needs to be determined in the training process of the ML method. The last $nx1$ vector ϵ models the (random) error which explains the residual missing to explain $\boldsymbol{y}$ fully. Depending on the exact

Glossary

Best linear unbiased predictor (BLUP): a statistical method used to estimate random effects in linear mixed models. BLUP can also assess genetic values.
Convergent LMM (cLMM): linear mixed models with convergence.
Data democratization: a process designed to enhance data accessibility while ensuring proper governance.
Deep learning (DL): a category of artificial intelligence (AI) that uses neural networks to enable computers to learn from data.
DL GS (DeepGS): deep learning applied to genomic selection.
DL GWAS (DLGWAS): deep learning applied to genome-wide association studies.
Feed forward networks (FFNs): data processing via forward propagation.
GBLUP: Genomic Best Linear Unbiased Prediction – model for genomic prediction.
Generalized linear mixed model (GLMM): extended version of linear mixed models: statistical models for normal and categorical data.
Genome association and prediction integrated tool (GAPIT): R-package.
Genome-wide complex trait analysis (GCTA): statistical method used to estimate the heritability of complex traits or phenotypes based on genetic data.
Genomic prediction (GP): forecasting traits based on genomic data.
Genomic selection (GS): precision breeding through genetic advancements.
Gradient boosting machine (GBM): ensemble learning technique: model for regression.
Item based collaborative filter multi-trait multi-environment (IBCF MTME): a method used in genomic prediction.
LightGBM: Light gradient boosting machine, a method for efficient gradient boosting machine learning framework.
Linear mixed model (LMM): statistical models integrating fixed and random effects.
Machine learning (ML): statistical and computational methods for genomic prediction.
Maximum *a posteriori* threshold GP (MAPT): a model for genomic prediction of ordinal data.
Multilayer perceptron (MLP): a neural network architecture for learning.
Negative binomial (NB): statistical distribution used for count data.

data structure, users often include additional effects, such as year or location effects, genetic interaction (epistasis), or environmental effects or genotype–environment interaction.

This linear model of Equation 1 is the central object for several GP methods. Differences between the methods are often given only by the choice of the side conditions that are introduced to regularize the regression. 'Regularization' refers to restricting the space in which the values of $\boldsymbol{\beta}$ are located or 'likely to be located'. A regularization is required since the number of genotypes n is usually much smaller than the number of predictor variables p. A standard linear ordinary least square regression would therefore not lead to a unique solution, and any of the possible solutions would fit the training data perfectly but provide only little predictive ability for new data. This circumstance is referred to as 'over-fitting', and the regularization can be considered as a mathematical restriction to provide a unique solution for $\hat{\boldsymbol{\beta}}$, but also as a statistical tool to separate 'signal' from 'noise'.

Having this in mind, Equation 1 is called **ridge regression BLUP (rrBLUP)** when the side condition penalizes the squared effect size of the marker effects $\boldsymbol{\beta}$. The unique fit of the training data is then given by minimizing the distance of the theoretical fit to $\boldsymbol{y}$ with the side condition of keeping the sum of squared entries of $\hat{\boldsymbol{\beta}}$ small as well. The difference between several GP methods lies then only in the exact definition of how to restrict $\hat{\boldsymbol{\beta}}$. rrBLUP and GBLUP use both a penalty on the squared effect size of $\hat{\boldsymbol{\beta}}$ and are basically identical, depending on how exactly the penalty weight on $\hat{\boldsymbol{\beta}}$ is defined. The ridge regression approach can also be derived from the prior assumption of Gaussian distributed entries of $\boldsymbol{\beta}$. If effects that are to be estimated are modeled as being random with a specific distribution, the linear model is also called a 'mixed linear model', to highlight that some effects are modeled as being fixed but unknown, and others are random and unknown. The 'random' nature defines a restriction by giving *a priori* a statement which estimates are 'more likely'.

By contrast with rrBLUP, LASSO penalizes the sum of the absolute values of $\hat{\boldsymbol{\beta}}$ instead of the square. This leads to a tendency of estimating many markers to have an effect of zero, while ridge regression – due to the prior assumption – tends to generate Gaussian distributed estimates. The Bayesian alphabet uses different prior distributions for $\boldsymbol{\beta}$. Moreover, LASSO may be advantageous when the number of relevant loci is small and the heritability of the trait is high [13].

Epistasis models, which include interaction effects between loci, have often outperformed the additive model of Equation 1 when predicting wheat yield [14,15]. This observation may be related to interactions between sub-genomes [16] or simply statistical aspects such as marker density [17]. Overall, out of the different regularization methods for a linear model, rrBLUP/GBLUP is therefore the first choice when exploring the potential of GP in a breeding program.

Areas under development

As highlighted, Equation 1 describes the basic model. Depending on the approach of the breeding program, several evaluation sites may be used for each year. Over years, valuable data are accumulated, but the training data are structured by cohorts and phenotypic evaluations with year effects. Such data need to be merged in an appropriate way to be of optimal use for predictions.

Additionally, the dimensions of available data are increasing. Environmental covariates can be included in prediction models ('enviromic prediction'); high-throughput phenotyping data allows the exploration of 'phenomic prediction' [18–21], and available data from metabolomic profiles can also be included in the prediction model [22]. In addition, feature selection combined with data augmentation are also strategies to enhance accuracy in genomic-enabled prediction by minimizing differences between the training and testing populations (Box 1).

Recurrent neural networks (RNNs): artificial neural networks for sequential data tasks.
Reproducing kernel Hilbert space (RKHS): methods that transform the original input of the models before the training process.
Restricted maximum likelihood (REML): a statistical method for variance component estimation.
Ridge regression BLUP R package (rrBLUP): ridge regression using R statistical package.
Super BLUP (sBLUP): a superior version of BLUP.
Support vector regression (SVR): regression using support vector machines.
TGBLUP: threshold genomic best linear unbiased predictor model
Trait analysis by association, evolution, and linkage (TASSEL): methods for trait analyses.

Box 1. Data augmentation enhances within family genomic prediction

The goal of predicting within a family is problematic due to Mendelian sampling variance. Recently, Montesinos-López *et al*. [79] proposed the AB method, which combines the virtues of the adversarial validation (A) method and the Boruta (B) feature selection method. The AB method minimizes the disparity between training and testing distributions. The A method detects the presence and magnitude of the mismatch between the training and testing sets using a binary classifier with the original features (inputs) and its shuffled counterparts and a fictitious response variable, while the Boruta computes feature importance, also using the same fictitious response variables, then with the inverse of the feature importance scores the original features (markers) are weighted, and using them, a weighed genomic relationship matrix can be obtained. The authors reduce the weight assigned to markers that display the most significant differences between the training and testing sets. The AB method built a weighted genomic relationship matrix that is implemented with the Genomic Best Linear Unbiased Predictor (GBLUP) model. Results show that the proposed AB method outperforms the GBLUP by 8.6, 19.7, and 9.8% regarding Pearson´s correlation, mean square error (MSE), and normalized root mean square error (NRMSE), respectively. Their results support the idea that the proposed AB method is a useful tool for improving the accuracy of the prediction of a complete family.

We will now transit to models for non-Gaussian distributions, followed by DL methods that do not require the specific design of additive and interaction effects of the different types of predictor variables.

The importance of non-Gaussian traits in genomic prediction for plant breeding

While much research has focused on traits that are normally distributed, such as yield and height, there is an increasing recognition of the critical importance of non-Gaussian traits, such as ordinal, Poisson, and count data, which are often tied to essential breeding targets. These non-Gaussian traits frequently represent categorical or discrete biological phenomena, such as disease resistance, flowering time, or the number of seeds per pod, each of which profoundly impacts crop productivity and resilience.

The relevance of non-Gaussian traits is that they present unique challenges and opportunities for breeders. Traits such as disease resistance are often measured on ordinal scales, where severity scores classify a plant's susceptibility or resistance to pathogens. Similarly, many agriculturally important traits follow Poisson or **negative binomial (NB)** distributions, such as the count of disease lesions or the number of tillers per plant. Accurately modeling these traits is crucial because the assumption of normality, inherent in traditional linear models, is violated, potentially leading to biased predictions and reduced selection accuracy.

A common approach to handle non-Gaussian traits is to apply transformations to approximate normality, such as using logarithmic or square root transformations. However, these transformations often have significant drawbacks. However, specialized models offer significant advantages by directly accommodating the unique distributional properties of non-Gaussian traits. For ordinal data, Bayesian logistic ordinal regression (BLOR) and Bayesian **threshold genomic best linear unbiased prediction (TGBLUP)** explicitly model the ordered categorical nature of the trait, preserving the inherent ranking and providing more biologically meaningful interpretations. For count data, Poisson regression models and negative binomial models address the discrete and often over-dispersed nature of these traits without distorting the data structure. Poisson-lognormal models further enhance flexibility by modeling complex variance structures that transformations cannot handle adequately.

Furthermore, **Bayesian regularized neural networks (BRNNs)** extend the capability of GP by accommodating nonlinear relationships and capturing interactions in complex datasets. These models provide robust predictions while maintaining the original scale and distributional characteristics of the trait, leading to more reliable selection outcomes.

Count data represent discrete, non-negative whole numbers, such as the number of reads mapped to a genomic region. Poisson regression models are commonly used to analyze count data in GP.

Poisson regression assumes that the counts follow a Poisson distribution, and the mean and variance are assumed to be the same under the Poisson distribution; for this reason, when there is the presence of over-dispersed distributions like the NB are preferred, since they do not assume that the mean and variance are the same [23]. In GP, a researcher can include genetic markers or other genomic features as predictor variables in a Poisson regression model. Poisson data represent count data but with a specific assumption of constant rate over a given time or space interval. It is often used when modeling events that occur randomly in a fixed time. The Poisson distribution used for modeling count data in the context of GP belongs to a **generalized linear mixed model (GLMM)** with a Poisson distribution for the conditional response variable given the random effects and with a log link function [24]. As stated before, these models can account for both fixed and random effects while modeling the genetic contribution to the count data.

Ordinal data consist of ordered categories, such as disease severity scores or ratings. For example, a disease severity scale may include categories such as mild, moderate, and severe. Ordinal regression models, such as proportional odds models or cumulative logit models, are suitable for modeling ordinal data in GP. These models estimate the odds or probabilities of an individual falling into each category based on genetic markers or other predictors [25]. Binary data represent two categories, such as the presence or absence of a particular genetic variant or the occurrence of a disease. Logistic and probit regression models are commonly used to model binary data in GP. Logistic and probit regression models estimate the probability of an event occurring based on genetic markers or other predictors. By including genetic markers as predictor variables, one can assess their association with the binary outcome of interest [26,27].

The Bayesian TGBLUP model proposed by Montesinos-López *et al.* [28] is a Bayesian version of the classical probit models. It exhibits high competitiveness in terms of prediction performance, as demonstrated by Montesinos-López *et al*. [29] in their comparison to DL and support vector machine methods. However, due to its Bayesian framework utilizing Gibbs sampling, the TGBLUP model necessitates substantial computational resources, as convergence requires a significant amount of time, particularly when applied to large datasets. Montesinos-López *et al.* [30] also introduce the **maximum *a posteriori* threshold GP (MAPT)** model for ordinal traits, which proves to be more efficient than TGBLUP in terms of implementation time; however, it is less efficient than the TGBLUP model in prediction performance. Also, Montesinos-López *et al.* [30] proposed a statistical ML framework that can be used to predict and analyze traits that exhibit an ordinal scale of measurement. While in the BGLR package, under a probit framework, it is possible to deal with binary and ordinal outputs as response variables with a very general framework based on a Bayesian perspective [8].

As an alternative less sensitive to outliers for analyzing ordinal data, Montesinos *et al*. [31] introduced a **Bayesian logistic ordinal regression (BLOR)** model that uses Pólya-Gamma data augmentation for efficient estimation and prediction through a Gibbs sampler, resembling the TGBLUP model as a special case. Simulations and real data confirm its effectiveness for GPs.

The Bayesian regularized neural network (BRNN or BRNNO) proposed by Pérez-Rodríguez *et al.* [32] is a statistical ML approach specifically designed for handling ordinal data. It combines the flexibility of neural networks with Bayesian regularization techniques. The BRNN model is built upon a neural network architecture where Bayesian regularization is applied to prevent overfitting and improve the model's generalization ability. Regularization techniques, such as weight decay or dropout, are incorporated to control the complexity of the network and avoid excessive sensitivity to individual observations. Bayesian inference is used to estimate the parameters of the BRNN model, and it uses appropriate loss functions and activation functions in the neural network. Loss functions, such as the ordinal logistic loss or the proportional odds loss, are used to measure

the discrepancy between the predicted ordinal values and the observed values. By combining the power of neural networks with Bayesian regularization, the BRNN model provides a robust and flexible framework for modeling ordinal data. It allows for capturing complex relationships and patterns in the data while incorporating uncertainty through the Bayesian framework.

As mentioned earlier, the NB distribution is preferred for modeling count data when overdispersion is present, and the Poisson distribution fails to account for this excess of variability. Montesinos-López *et al*. [33,34] proposed a Gibbs sampler for the NB distribution that is not computationally efficient, and other authors have suggested using the Poisson lognormal distribution to model count data and account for overdispersion [35]. In the Poisson-lognormal distribution, the Poisson component describes the actual number of counts observed within a single unit or sample as integer inputs or outputs. The lognormal component of the distribution describes the overdispersion in the Poisson rate parameter; it accounts for the clustering of certain factors, and explains how the average of these factors varies across the population [35]. Incorporating this lognormal component into the predictor of a Poisson model is highly beneficial for accounting for overdispersion since it allows for accommodating a general correlation structure between traits when studying more than one trait. Montesinos-López *et al*. [36] investigated the GP accuracy of a Bayesian Poisson-lognormal for count multi-trait and multi-environment that allows borrowing information between environments and between traits. These authors concluded that the proposed multi-trait multi-environment Poisson-lognormal accounts efficiently for the overdispersion of the data and gave better predictions than the single trait and single environments when traits and environments are correlated.

Traditional regression models, such as linear regression, are not well suited for counting data, as they assume normality and can lead to biased results. Montesinos-López *et al*. [37] proposed a novel approach for predicting count data in genomic-based prediction using a Poisson deep neural network model, which combines the flexibility and nonlinearity of deep neural networks with the Poisson distribution's ability to model count data. The model is trained on genomic features to predict the count outcomes accurately. The Poisson deep neural network model captures the complex relationships between genomic markers and count phenotypes, and it consists of multiple layers of interconnected neurons that learn hierarchical representations of the input data, allowing for the discovery of intricate patterns and associations. The authors compared the performance of the Poisson deep neural network model with other popular count regression models, showing that their proposed model outperforms the alternatives in terms of prediction accuracy and robustness.

However, the models explained so far for count data have difficulties when the response variable exhibits an excess of zeros like some traits, such as disease resistance in plant and animal breeding. To address this challenge, Montesinos-López *et al*. [38] introduced a zero-altered Poisson random forest model. Random forests are ensemble learning methods that combine multiple decision trees to make predictions. In this study, the random forest algorithm is modified to accommodate zero-inflated and over-dispersed count data, which are common in GS-assisted plant breeding. The zero-altered Poisson random forest model incorporates two components: a binary component to model the presence or absence of zeros, and a count component to model the non-zero counts. The binary component identifies whether a given observation is more likely to be a zero or non-zero, while the count component predicts the actual count value. The authors compared the performance of their proposed model with other existing methods, including standard random forests and Poisson regression models. The results showed that the zero-altered Poisson random forest model outperformed the other methods in accurately predicting grain yield, particularly when dealing with zero-inflated and overdispersal count data.

Genomic prediction with deep learning models

More recently, DL tools, which play a central role in developing artificial intelligence (AI) systems, have gained prominence in breeding. Roughly speaking, DL models consist of several layers of artificial neurons which can be activated subsequently. The input data activate the first layer, which then activates the second layer, until the final neuron layer generates the output, which is the dependent variable. Given a predefined set of activation functions, the learning process determines which input activates the first layer of neurons in what way (defining weights) and how the information transits through the other layers. DL architectures include **feed forward networks (FFNs)**, convolutional neural networks (CNNs), **recurrent neural networks (RNNs)**, multimodal architectures, and transformers which excel at capturing intricate patterns and dependencies within genomic sequences [39,40].

This description already points out the advantages and disadvantages of DL compared with simpler models established for GPs. As a first advantage, DL is very flexible in capturing genotype–phenotype relations. It is not restricted to a specific class of relationships, as are, for instance, non-linear models. Moreover, another advantage of DL methods is that they can automatically identify intricate patterns – for instance, genotype-by-environment interactions – in the data and can extract and pronounce relevant features. By contrast, linear models require the user to define such relevant aspects specifically when setting up the model. This capability is very attractive since the genotype–phenotype map is inherently non-linear, and when using additional layers of data – such as metabolomic data, microbiome, or other intermediate traits such as multispectral reflectance data – the specific relevance and role of each type of data may not be clear.

Disadvantages of the DL methods compared with simpler methods are that DL methods require, in general, very large datasets for obtaining good performance; they are computationally demanding, since a complex tuning process is required for optimal implementation [41,42], the (biologic) interpretability of the determined parameters is challenging, and there is a threat of overfitting or predicting phenotypes instead of genetic merit. As a simple example for the latter, imagine a model including environmental data that shows high predictive ability in cross-validations on a given dataset including genotypic data, phenotypic data, and the environmental conditions under which the crop was growing. The user needs to make sure that the high predictive ability is not a result of predicting the yield from the environmental conditions but that the genetic potential is elucidated. Despite potential challenges, DL holds the promise of uncovering hidden genetic features that traditional methods might overlook, thereby broadening our understanding of the genetic basis of traits.

Deep learning for plant genomics

DL focuses on training artificial neural networks with multiple layers to learn representations of data. These networks, known as deep neural networks, can automatically extract complex patterns from input data [43]. DL has demonstrated success in various domains, including image recognition, natural language processing, speech recognition, gene expression prediction, protein structure prediction, and disease risk prediction [39,44]. DL models can capture intricate relationships between genetic variants and phenotypic traits, leading to more accurate predictions and a better understanding of underlying genetic mechanisms [45,46].

In a recent article by Azodi *et al.*, the authors analyzed and compared the performance of neural networks versus gradient tree boosting machines using vast data across six plant species, each characterized by different marker densities and varying training population sizes [47]. These researchers assessed six linear and six non-linear algorithms. Key findings were that (i) applying feature selection before training neural networks became crucial when the number of markers

exceeded the training sample size, (ii) no single algorithm consistently outperformed the others in all species-trait scenarios, (iii) aggregate predictions from multiple algorithms (ensemble algorithms) demonstrated robust and reliable performance overall, and (iv) the performance of nonlinear algorithms exhibited greater variability depending on the trait. The authors concluded that artificial neural networks did not emerge as the top-performing model, but specific strategies – such as implementing feature selection and using seeded starting weights – enhanced their accuracy, making them competitive with other methods.

In the context of plant genomics, DL can be applied to analyze large-scale genomic datasets and make predictions about complex plant traits. There is also great potential for DL to be leveraged for predictions using whole-genome sequencing data, which are currently difficult to handle using standard computational analysis tools. DL models can automatically learn relevant features from genomic data without explicit feature engineering. This helps to capture complex relationships between genetic variants and traits. DL models can analyze DNA or RNA sequences to identify regulatory elements, predict the functional effects of genetic variants, or classify sequences based on their phenotypic impact [48]. Gene expression analysis uses RNNs or attention-based models to analyze gene expression time series data, predict gene expression levels, or identify gene regulatory networks [49]. This is likely to revolutionize the usage of whole-genome sequencing data for GP in the future. DL can be trained to predict complex traits, such as grain yield, disease resistance, or stress tolerance in plants, using genomic and other data inputs. As the resolution and scale of the data increase (e.g., sequencing and gene expression data), the accuracy of selection for complex traits is likely to be enhanced. This will enable breeders to identify promising plant lines earlier in the breeding process.

DL approaches for plant genomics require large datasets, intense computational resources, and domain expertise to ensure reliable results. Additionally, the interpretation of DL models in plant genomics is an ongoing area of research, as understanding the underlying genetic mechanisms from learned representations can be challenging. Overall, the integration of GP and DL holds great potential to accelerate crop improvement and advance our understanding of plant biology [50,51].

Several researchers have observed that GS in plants has not shown a clear superiority over DL in terms of prediction power compared with conventional GBLUP prediction models [41,42]. However, the authors mentioned that there is evidence that DL captures nonlinear patterns more efficiently than conventional genome-based methods, and that it is able to integrate data from different sources. Nevertheless, DL is not free of problems in its application to GP and plant genomics. For example, the high chances of overfitting (that could be solved by a stronger regularization and by the Bayesian paradigm), the difficulty of the biological interpretation of the results, and the intense computing capacity required to be used by practitioners with small-capacity laptops.

Montesinos-López *et al.* [41], listed the most recent advances on the use of DL (as a special form of ML) in GS of animal and plant breeding, including research on DL for GP of plant traits under multi-environment trials, and benchmarked with other ML. Authors have investigated the use of DL and other ML for continuous, ordinal, count, and binary traits for single as well as multiple traits [37,52–55]. In general, results showed that when excluding the genotype–environment interaction (G×E) from the model, DL models overcome the parametric GBLUP but not otherwise.

The training process for DL is challenging due to (i) the numerous hyper-parameters that need to be tuned, and (ii) imperfect tuning that can result in biased predictions. Trying to solve this problem, Montesinos-López *et al.* [56] proposed a simple method for calibrating that is computationally faster than the old computing calibration method. The new calibration method had a higher GP

ability than the conventional GBLUP. The authors used **multilayer perceptron (MLP)** and CNN where the genomic information was incorporated into the MLP as a relationship matrix and to CNN as a genomic image. Both MLP and CNN gave very competitive results compared with GBLUP [57].

Recently, a deep neural network for GP (DNNGP) for multi-omic data based on a multilayer hierarchical structure was introduced and compared with its GP accuracy versus GBLUP, light **gradient boosting machine (GBM)** (**LightGBM**), **support vector regression (SVR)**, **DL GS (DeepGS)** and **DL GWAS (DLGWAS)** [58]. The advantage of DNNGP over the other methods resides in its multilayer hierarchical structure allowing learning of features from the raw data, avoiding overfitting and enhancing the convergence rate by means of a batch normalization layer and early stopping and rectified linear activation function. The computation time of the DNNGP is competitive compared with the other methods, and for large datasets the proposed DNNGP is superior in prediction accuracy to all the other five models.

Wang *et al*. [58] pointed out five advantages of the DNNGP. One of these is its generality, as it can be applied to several omics data with a multilayer hierarchical structure that learns features from raw data and avoids overfitting. DNNPG can be efficiently used on small datasets with very competitive results as compared with other methods, and is faster than other methods for large datasets. The authors remarked that, in several instances, it overcame the prediction accuracy of GBLUP.

Few ML linked genomics and phenomics for GP using different approaches to link genomic and image data. DL neural networks have been developed to increase the GP accuracy of unobserved phenotypes while simultaneously accounting for the complexity of G×E. However, unlike conventional GP models, DL has not been investigated for when genomics is linked with phenomics. Montesinos-López *et al*. [59] utilized a multimodal DL method and compared its GP with those of GBLUP, GBM, and SVR. Multimodal DL provided better GP accuracy than the results obtained by the other models. However, GP accuracy obtained for other years indicated that the GBLUP model was slightly superior to the multimodal DL. This multimodal DL method [59] is novel and presents a strong degree of generalization, as several modules can potentially be incorporated and concatenated to produce an output for a multi-input data structure (Box 2). Genomic data provides information about an individual's genetic makeup, while phenomics data encompasses observable traits or characteristics. The integration of image data can provide detailed phenotypic information, such as plant morphology or disease symptoms, which can further enhance GP.

Software for GP

One of the pioneering R packages for genome-based prediction was introduced by de los Campos *et al.* [60]. Following this, Pérez *et al*. [61] introduced Bayesian linear regression (BLR), which allowed for the incorporation of molecular markers, pedigree information, and other covariates in high-dimensional linear regression models. The BLR R package offered the convenience of jointly analyzing markers and pedigree data, and the authors also delved into essential aspects such as assessing GP accuracy through random cross-validation and selecting optimal hyperparameters for Bayesian models. Another significant milestone was the development of the rrBLUP R package by Endelman [62], which facilitated ridge-regressions (whole-genome regressions) and **linear mixed models (LMMs)** with two variance components using maximum likelihood or **restricted maximum likelihood (REML)** methods.

Pérez and de los Campos [8] extended the original BLR R-package to a more comprehensive and versatile package called Bayesian generalized linear regression (BGLR), further expanding

Box 2. Multimodal deep learning

The most applied deep learning (DL) architectures in genomic selection (GS) are the multi-perceptron and the convolutional neuronal networks (CNN) [37,41,42,53–56]. In DL methods, thousands of single-nucleotide polymorphisms (SNPs) can be used to train a model with a vast number of parameters, or the squared root matrix or the Cholesky factor of the genomic relations matrix can also be used.

In plant breeding research, several statistical machine learning (ML) methods have been developed and studied for assessing the genomic prediction (GP) accuracy of unobserved phenotypes; only a few methods have linked genomics and phenomics. A DL neural network has been developed to increase the GP accuracy of unobserved phenotypes while simultaneously accounting for the complexity of genotype × environment interaction (G×E). As opposed to conventional GP models, DL has not been investigated when genomics and phenomics are linked. Montesinos-López *et al*. [59] incorporate DL with genomics and images (as covariables) tested for several traits (Figure I). The fitted models were: genomic best linear unbiased prediction (GBLUP), gradient boosting machine (GBM), support vector regression (SVR), and the DL method. Results indicated that for 1 year, DL provided better GP accuracy than results obtained by the other models (Figures II and III). However, GP accuracy obtained for other years indicated that the GBLUP model was slightly superior to the DL. The GP of various traits for drought and irrigation in 3 years shows slight increases in GP accuracy of DL over GBLUP for some years but not for the 3 years under drought conditions. Results from the study of Montesinos-López *et al.* [59] show that the multimodal DL method has a robust degree of generalization with other very data-specific DL methods previously reported. It is also important to stress that the BGLR package is a very vigorous parametric statistical software for GP accuracy. The DL method used in this study is novel and presents a reasonable degree of generalization and an important accuracy for predicting new years. One reason is that, instead of concatenating all feature types and using them to feed the created network, the outputs are combined to generate the output value for each type of information and individual neuronal network.

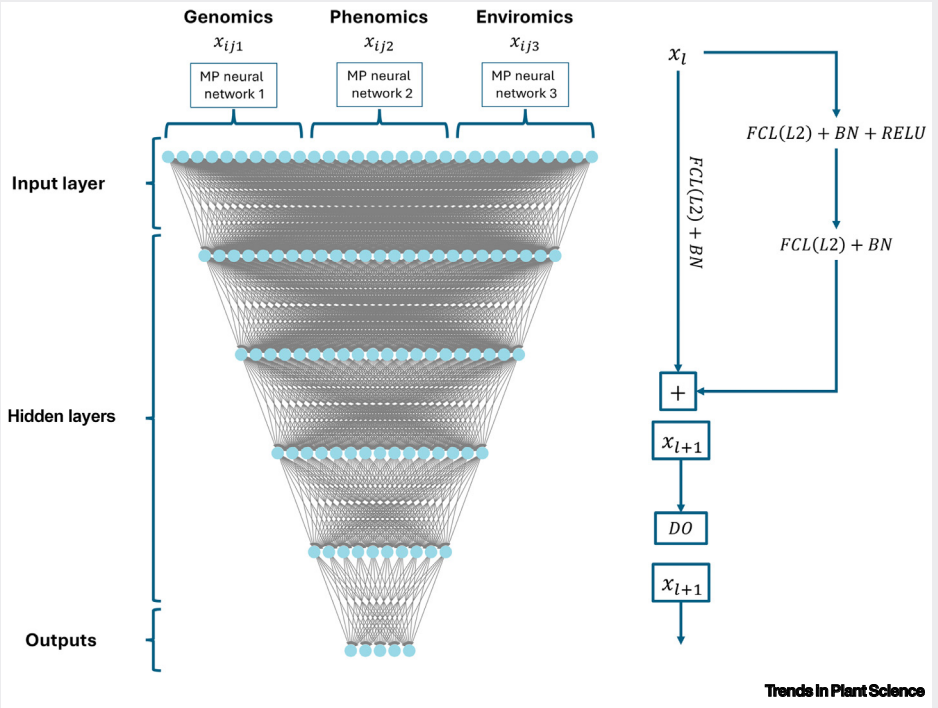

Figure I. Multimodal deep learning model (MMDL) with three modalities (type of input). A stacked residuals network (RestNet) is composed of two sequence-dense layers (FCL) applied in each multilayer perceptron neuronal network. FCL(L2)+BN+RELU are the successive applications of a fully connected layer (FCL) with L2 regularization, batch normalization layer, and relu activation. The meaning for FCL(L2)+BN is similar, while DO indicates the application of the dropout regularization. Batch normalization (BN) also works as a regularizer to speed up the training process [80]. The concatenated outputs of three networks are used as an input in the output layer with one neuron, linear activation function, and L2 regularization for its weights: concatenate outputs of all three (multilayer perceptron deep neuronal network) MP Neuronal Networks+FCL+L2. Figure extracted from Montesinos-López *et al.* [59].

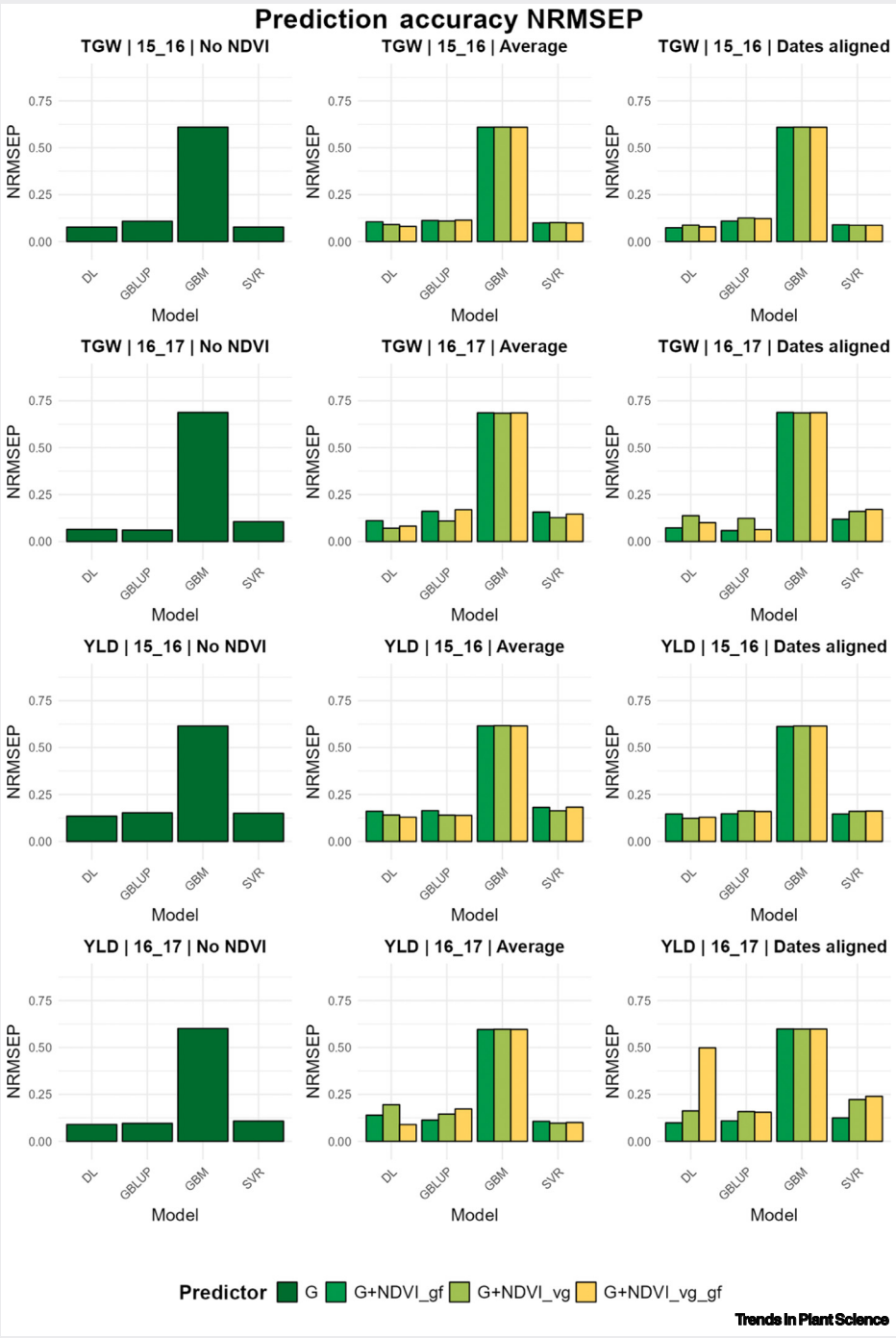

Figure II. DS1.normalized root mean squared error of prediction (NRMSEP) for the prediction of a complete year using information from the other year (leave one environment out, LOO) for traits TGW (thousand grain weight) and YLD (grain yield) using the models fitted with Bayesian Genomic Linear Regression (BGLR; R Software) and with deep learning (DL) (Python Software) with the predictors Genomic Matrix (G; NDVI, Normalized Difference Vegetation Index; no use of NDVI), which correspond to the genomic matrix, G + NDVIs_gf, G+NDVIs_vg (vg, vegetative stage), and G+NDVIs_vg_gf (grain filling stage) for the NDVIs covariate in any of its two types NDVI Averages and Dates aligned. Figure extracted from Montesinos-López *et al.* [59].

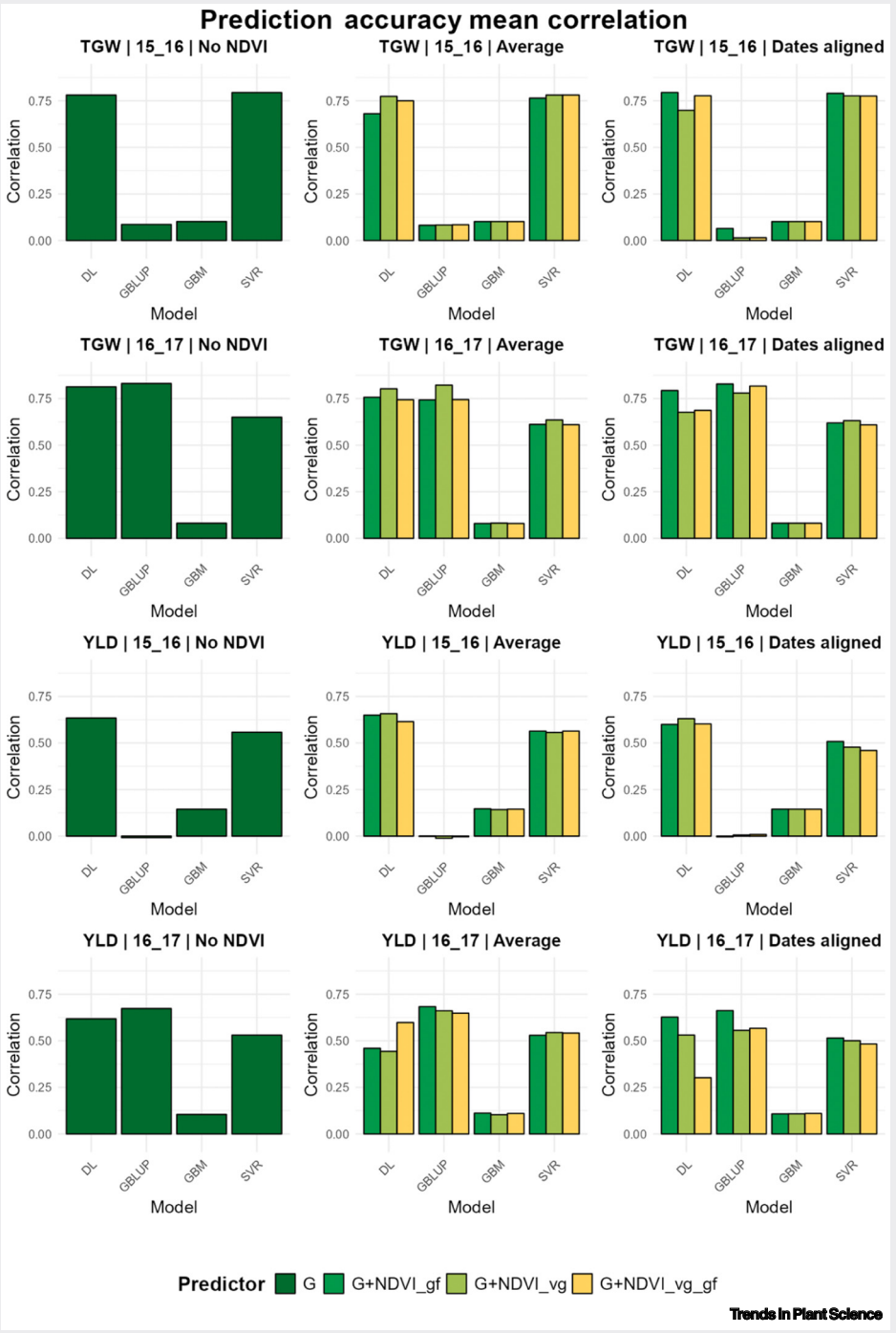

Figure III. DS1.correlation mean (Cor Mean) for the prediction of a complete year using information from the other year (leave one environment out, LOO) for traits TGW (thousand grain weight) and YLD (grain yield) using the models fitted with Bayesian Genomic Linear Regression (BGLR; R Software) and with deep learning (DL) (Python Software) with the predictors Genomic Matrix (G; NDVI, Normalized Difference Vegetation Index; no use of NDVI), which correspond to the genomic matrix, G + NDVIs_gf (grain filling stage), G+NDVIs_vg (vegetative stage), and G+NDVIs_vg_gf for the NDVIs covariate in any of its two types NDVI averages and dates aligned. Figure extracted from Montesinos-López *et al.* [59].

the capabilities for GP tasks. These R-based tools have significantly contributed to the advancement of genomic-enabled prediction in the field of plant breeding and genomics research. From this software, it is possible to democratize a wide range of genomic models and methods in a unified computing software for data analysis, and for this reason it became popular in dealing with multiple data types (e.g., genomic markers and environmental covariables). Additionally, the BGLR package includes various Bayesian regression models, parametric variable selection, shrinkage methods, and semi-parametric procedures, which also support both continuous and categorical response traits. The BGLR algorithm is based on a Gibbs sampler with efficient routines implemented in C programming language, and serves as the main framework for adapting more complex genomic models, such as G×E with pedigree and environmental covariates. It is also used for assessing marker effect × environment interaction.

Pérez-Rodríguez and de los Campos [63] extended the BGLR package to fit multi-trait models. The software allows researchers to fit a model with an arbitrary number of random effects and assign different prior distributions to marker effects (e.g., Gaussian, Spike-Slab). The software also incorporates routines to model variance–covariance matrices (e.g., diagonal, factor analytic, recursive).

A new package for fitting multi-trait GP models is MegaLMM [64]. MegaLMM tackles the challenge of scaling multi-trait LMMs to hundreds or thousands of traits at once to accommodate phenomic data or data from METs with large numbers of trials. It is particularly useful when a set of genotypes is measured for many but not all possible traits (sparse data). MegaLMM is based on a Bayesian latent factor model to regularize estimates of among-trait covariance matrices caused by genetic or environmental factors in a similar way that marker effects are regularized in single-trait GP.

An active area of research is the development of software for fitting LMM. The lme4 is the standard package for fitting linear and generalized LMMs in the R package, but it lacks the ability to define correlations between individuals or groups in genetic analyses. To address this, a new package called lme4GS has been introduced for R [65] which is focused on fitting LMMs with covariance structures defined by the user, bandwidth selection, and GP. The new package is focused on GP models used in GS and can fit LMMs using different variance–covariance matrices. Several examples of GS models are presented using this package as well as the analysis using real data.

Other packages [66] important in fitting LMM are the sommer [66], capable of fitting LMM with multiple random effects accounting for known or unknown covariance structures, and rTASSEL [67]. This latter is a modern R interface for the TASSEL software [5], among the most important software for gene discovery by association mapping (e.g., GWAS), in which a GP plugin was introduced in 2015 to include supervised approaches, such as ridge regression and GBLUP. This software has the advantage of working directly with variant call format (VCF) files, which speed up the process of SNP quality control, computation of relationship matrices, and dimensionality reduction techniques – for example, principal component analysis (PCA), multidimensional scaling (MDS) – as well making it easier to draw linkage disequilibrium maps, explore phylogenetic trees, and fit LMM into a single and memory-efficient computational platform. In practical terms, an ML platform for GP could take advantage of this software to accommodate modeling approaches directly from the DNA sequence to trait variations.

Moreover, the R package GAPIT serves as a widely adopted tool for conducting both GWAS and GP analyses, using a range of diverse models. In its initial version, GAPIT version 1, users were empowered to execute several models, including GLM (generalized linear model), LMM,

compressed LMM (CLMM), and GBLUP [3]. In subsequent iterations like GAPIT version 2, advancements continued with the integration of enriched factored spectrally transformed LMM (FaST-LMM), enhanced **convergent LMM (cLMM)**, and settlement of LMMs under progressively exclusive relationship (SUPER) [4]. Recently, the most recent version of the GAPIT package (version 3) was released in 2021 by Wang and Zhang [68]. Notably, this version included several new options for GWAS application, such as the fixed and random model circulating probability unification (FarmCPU), Bayesian information and linkage-disequilibrium iteratively nested keyway (BLINK), as well as GP models such as compressed BLUP (cBLUP) and **super BLUP (sBLUP)**.

Recently, a new R library for GS, called BWGS, was released by Charmet and colleagues [69]. This R package includes most of the steps for GS application, such as missing data imputation, dimension reduction (discarding uninformative markers), model training using 15 different models: for example, GBLUP, EGBLUP, the Bayesian alphabet, **reproducing kernel Hilbert space (RKHS)**, RF, and SVM. The package also allows us to perform a random cross-validation by using a set of genotyped and phenotype lines for model testing and the GEBVs estimation for a set of unphenotyped lines. To extend the computational capabilities to deal with multi-trait and multi-environment data, other initiatives have been established, including Bayesian approaches and DL and enviromics. One example is the R package accounting for the **item based collaborative filter multi-trait multi-environment (IBCF MTME):** an algorithm developed by O.A. Montesinos-López *et al*. [70].

The implementation of the Bayesian generalized kernel regression method in R for GP is important for efficiently capturing complex nonlinear patterns that conventional linear regression models cannot handle. The software presented by A. Montesinos-López *et al.* [71] performed these tasks and is also powerful for leveraging environmental covariables, including G×E interaction prediction. The authors gave the bases for constructing seven kernel methods, linear, polynomial, sigmoid, Gaussian, exponential, Arc-cosine I, and Arc-cosine L. Furthermore, the authors provide illustrative examples for implementing exact kernel methods in single-environment, multi-environment, and multi-traits frameworks as well as the implementation of sparse kernel methods in a multi-environment framework.

A new R-based software package called sparse kernel methods (SKM) for implementing six popular supervised ML algorithms (generalized boosted machines, generalized linear models, support vector machines, random forests, Bayesian regression models, and deep neural networks) was developed by Montesinos-López *et al.* [72]. SKM also offers the option to use sparse kernels. The primary focus of SKM is user simplicity, providing an easy-to-understand format that encompasses the most important aspects of these six algorithms. Additionally, the package includes a function for computing seven different kernels: linear, polynomial, sigmoid, Gaussian, exponential, arc-cosine I, and arc-cosine L (with L = 2, 3, etc.), along with their sparse versions. These kernels enable users to create kernel machines without modifying the statistical ML algorithm. The SKM package makes the computation of the sparse versions of the seven basic kernels. This functionality is crucial for reducing the computational resources required to implement kernel ML methods without a significant loss in prediction performance. To evaluate the performance of SKM, the authors experimented using a genome-based prediction framework with maize and wheat datasets. Although initially designed for GP problems, the SKM package is not limited to such applications and can be utilized in various domains.

A new study comparing the performance of three ML methods using multi-trait GP under GBLUP, partial least squares, and random forest was published by Montesinos-López *et al*. [73]. Among the three ML methods, random forest performed the best when using predictors,

environments, genomics, and their interaction. The authors also highlighted the availability of these three ML methods in the SKM library, which includes various single-trait and multi-trait statistics using the SKM library. In relation to the use and properties of the SKM library, Montesinos-López *et al*. [74] recently introduced a complete guideline for using the SKM R library and explained how to implement statistical ML [74] available in this library for GP. The general guide included details of the functions necessary to implement the various ML methods and thus facilitate their use by practitioners and analysts.

A recently developed tool, known as characterization and integration of driven omics (CHiDO), enables the integration of diverse omics datasets in a user-friendly interface [75]. This innovative software is a noncoding application designed to promote the democratization of multi-omics selection within breeding programs. By using CHiDO, breeders can analyze and model multi-omics data such as genomics, phenomics, and enviromics along with their interactions, thereby enhancing the prediction of complex traits.

A critical step in plant breeding programs is parental selection. Finding the best parents for starting new breeding pipelines might be deeply challenging due to the number of traits that breeders take into consideration, especially if those traits often show trade-offs. A novel R package called IPLGP was recently generated to provide a new GP approach to identify superior parents by using multi-trait selection [76]. This approach is initiated by crafting a selection index by utilizing normalized GEBVs and incorporating subjectively assigned weights for each trait. Subsequently, parental selection can be performed based only on the GEBVs by choosing the lines with the best selection index, based only on the genetic diversity (GD) maximizing the D-score, or combining both GEBVs and GD. Having identified the optimal parent, a simulation cascade ensues, encompassing pivotal steps of the breeding pipeline such as crosses, self-pollination, and selection. Finally, the genetic gain for each trait is estimated as the difference between the GEBVs of the parent and the GEBVs of the last generation.

Selecting multiple traits becomes increasingly challenging when traits are negatively correlated or when some traits have missing data. To address these complexities, the multi-trait parental selection (MPS) R package has recently been introduced [77]. This package leverages Bayesian optimization algorithms along with three distinct loss functions (Kullback–Leibler, energy score, and multivariate asymmetric loss) to identify parental candidates with optimal trait combinations. The application demonstrates that the MPS package is a powerful tool for selecting superior parents through multi-trait genomic selection, empowering breeders to make data-driven decisions and achieve high-performance offspring across multiple traits.

After selecting the parent lines, breeders across all crops face an equally challenging task: deciding which crosses to make. In plant breeding, this decision is particularly complex due to inherent trade-offs among target traits, the combining ability of each genotype, and the need to consider the future genetic variance of the offspring. To address this complexity, the R package PopVar provides breeders and quantitative geneticists with a powerful tool to make informed decisions [78]. It enables accurate predictions of key metrics, including the population mean (μ), genetic variance (V_G), the mean performance of the top 10% of the offspring (superior progeny mean), and the genetic correlation across multiple traits in the predicted biparental populations.

Advanced data management through envirotyping and graphical haplotypes

This version highlights the advanced nature of the tools, their role in combining genomic and environmental data, and their relevance to precision breeding efforts. It adds a broader scope to capture the innovative and practical implications of these methods.

The practical haplotype graph (PHG) utilizes graph-based pangenomes to impute high-density SNPs and haplotypes from sparse genotyping data, making GP more accessible and cost-effective. Implemented in species like sorghum, wheat, and cassava, PHG enables low-cost genotyping while maintaining or improving prediction accuracy compared with traditional methods. For example, sorghum studies demonstrated no accuracy loss when imputing SNPs via PHG, while cassava showed higher GP accuracy with PHG compared with BEAGLE. The R-based interface (rPHG) enhances its usability for breeding programs, allowing efficient genotyping and improved imputation quality.

The EnvRtype R package integrates environmental data with genomics for GP, offering tools for remote sensing, environmental profiling, and kernel-based enrichment of GP models. Drawing on NASA POWER and SoilGrids data, it efficiently processes eco-physiological variables for crops such as maize, wheat, rice, and eucalyptus. Future developments aim to deliver unsupervised envirotyping, linking environmental markers to genomic variants and optimizing breeding trials using genetic algorithms (GAs).

Advanced methods, such as environmental-to-phenotype associations (EPAs), use historical environmental trends to model location similarities and genotype reaction norms. These approaches enhance site selection and future environmental predictions, combining genomic, environmental, and empirical relationships into a unified G×E kernel for predictive modeling.

Concluding remarks

In this review we outlined the principles of GP and discussed how statistical ML methods enhance GP efficiency as a prediction methodology. We also examined the advantages and disadvantages of the statistical ML tools developed in the past year for predicting continuous, binary, categorical, and count traits. Additionally, we focused on DL models, highlighting their successful applications in genomic selection while also addressing their limitations and strengths. Also, we explored the software developed in recent years aimed at democratizing the GP methodology, and we reviewed the data management tools available for the broader application of GS methodology (see Outstanding questions).

Acknowledgments

We acknowledge the financial contributions of the Accelerated Breeding (ABI) WIN01.05.19 (Breeding pipelines 1 and 3) and WIN01.04.16 (Application of Genotyping).

Declaration of interests

The authors declare no competing interests.

References

1. Meuwissen, T.H.E. *et al.* (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829
2. Gholami, M. *et al.* (2021) A comparison of the adoption of genomic selection across different breeding institutions. *Front. Plant Sci.* 12, 728567
3. Lipka, A.E. *et al.* (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28, 2397–2399
4. Tang, Y. *et al.* (2016) GAPIT Version 2: an enhanced integrated tool for genomic association and prediction. *Plant Genome* 9. https://doi.org/10.3835/plantgenome2015.11.0120
5. Bradbury, P.J. *et al.* (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635
6. Yang, J. *et al.* (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82
7. Ayres, D.L. *et al.* (2012) BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* 61, 170–173
8. Pérez, P. and De Los Campos, G. (2014) Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495
9. Gianola, D. *et al.* (2009) Additive genetic variability and the Bayesian alphabet. *Genetics* 183, 347–363
10. Gianola, D. (2013) Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 194, 573–596
11. Santosa, F. and Symes, W.W. (1986) Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Stat. Comput.* 7, 1307–1330
12. Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* 56, 267–288

Outstanding questions

What challenges exist in genomic prediction for continuous, binary, ordinal, and count traits, and what models have been proposed?

Despite limited progress in GS for mixed phenotypes (e.g., continuous, binary, ordinal, and count traits), challenges remain. For count data, statistical models are available but are often impractical in GS due to the large *n* (number of observations) and small *p* (number of parameters) setup, as well as high computational demands. There is a need for models suitable for large *p* and smaller *n*, such as the Bayesian mixed negative binomial (BMNB) model, which incorporates genotype-by-environment (G×E) interactions using full conditional Gibbs sampling.

What are the limitations of Bayesian GS models using Poisson and negative binomial distributions, particularly regarding computational demands and their ability to capture nonlinear interactions?

Bayesian GS models employing Poisson and negative binomial distributions have been developed but face significant computational challenges due to the reliance on nonanalytical Gibbs samplers. Furthermore, these linear models often fail to effectively capture nonlinear interactions, which are crucial in complex traits.

What are the challenges of using Bayesian TGBLUP for ordinal traits, such as disease resistance?

Ordinal traits, such as disease resistance, present unique challenges. The Bayesian TGBLUP is a suitable approach but is computationally demanding due to its reliance on Gibbs sampling for parameter estimation.

What are the advantages and limitations of maximum a posteriori (MAP) estimation for large datasets and how can its inability to estimate parameter uncertainty be addressed?

MAP estimation offers a computationally efficient alternative to Bayesian methods for large datasets. However, it does not provide parameter uncertainty estimates. Techniques such as cross-validation

13. Wimmer, V. *et al.* (2013) Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics* 195, 573–587
14. Crossa, J. *et al.* (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724
15. Martini, J.W.R. *et al.* (2016) Epistasis and covariance: how gene interaction translates into genomic relationship. *Theor. Appl. Genet.* 129, 963–976
16. Cuevas, J. *et al.* (2024) Modeling within and between subgenomes epistasis of synthetic hexaploid wheat for genome-enabled prediction of diseases. *Genes (Basel)* 15, 262
17. Schrauf, M.F. *et al.* (2020) Phantom epistasis in genomic selection: on the predictive ability of epistatic models. *G3 Genes Genomes Genet.* 10, 3137–3145
18. Rincent, R. *et al.* (2018) Phenomic selection is a low-cost and high-throughput method based on indirect predictions: proof of concept on wheat and poplar. *G3 Genes Genomes Genet.* 8, 3961–3972
19. Robert, P. *et al.* (2022) Phenomic selection: a new and efficient alternative to genomic selection. *Methods Mol. Biol.* 2467, 397–420
20. Loladze, A. *et al.* (2024) Use of remote sensing for linkage mapping and genomic prediction for common rust resistance in maize. *Field Crop Res.* 308, 109281
21. Crossa, J. *et al.* (2021) The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. *Front. Plant Sci.* 12, 651480
22. Westhues, M. *et al.* (2017) Omics-based hybrid prediction in maize. *Theor. Appl. Genet.* 130, 1927–1939
23. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.* 11, R106
24. Stroup, W.W. (2016) *Generalized Linear Mixed Models*, CRC Press
25. Agresti, A. (2010) *Analysis of Ordinal Categorical Data*, Wiley
26. Albert, J.H. and Chib, S. (1993) Bayesian Analysis of Binary and Polychotomous Response Data. *J. Am. Stat. Assoc.* 88, 669–679
27. Hosmer, D.W. *et al.* (2013) *Applied Logistic Regression*, Wiley
28. Montesinos-López, O.A. *et al.* (2015) Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding. *G3 Genes Genomes Genet.* 5, 291–300
29. Montesinos-López, O.A. *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 Genes Genomes Genet.* 9, 601–618
30. Montesinos-López, A. *et al.* (2020) Maximum *a posteriori* threshold genomic prediction model for ordinal traits. *G3 Genes Genomes Genet.* 10, 4083–4102
31. Montesinos-López, O.A. *et al.* (2015) Genomic-enabled prediction of ordinal data with Bayesian logistic ordinal regression. *G3 Genes Genomes Genet.* 5, 2113–2126
32. Pérez-Rodríguez, P. *et al.* (2020) Genome-based prediction of Bayesian linear and non-linear regression models for ordinal data. *The Plant Genome* 1–13
33. Montesinos-López, O.A. *et al.* (2015) Genomic prediction models for count data. *J. Agric. Biol. Environ. Stat.* 20, 533–554
34. Montesinos-López, A. *et al.* (2016) Genomic Bayesian prediction model for count data with genotype × environment interaction. *G3 Genes Genomes Genet.* 6, 1165–1177
35. Williams, M.S. and Ebel, E.D. (2012) Methods for fitting the Poisson-lognormal distribution to microbial testing data. *Food Control* 27, 73–80
36. Montesinos-López, O.A. *et al.* (2017) A Bayesian Poisson-lognormal model for count data for multiple-trait multiple-environment genomic-enabled prediction. *G3 Genes Genomes Genet.* 7, 1595–1606
37. Montesinos-López, O.A. *et al.* (2020) A multivariate Poisson deep learning model for genomic prediction of count data. *G3 Genes Genomes Genet.* 10, 4177–4190
38. Montesinos-López, O.A. *et al.* (2021) A zero altered Poisson random forest model for genomic-enabled prediction. *G3 Genes Genomes Genet.* 11, jkaa057
39. LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
40. Goodfellow, I. *et al.* (2016) *Deep Learning*, MIT Press
41. Montesinos-López, O.A. *et al.* (2021) A review of deep learning applications for genomic selection. *BMC Genomics* 22, 19
42. Montesinos-López, O.A. *et al.* (2021) Deep-learning power and perspectives for genomic selection. *Plant Genome* 14, e20122
43. Shrestha, A. and Mahmood, A. (2019) Review of deep learning algorithms and architectures. *EEE Access* 7, 53040–53065
44. Abiodun, O.I. *et al.* (2018) State-of-the-art in artificial neural network applications: a survey. *Heliyon* 4, e00938
45. Telenti, A. *et al.* (2018) Deep learning of genomic variation and regulatory network data. *Hum. Mol. Genet.* 27, R63–R71
46. Eraslan, G. *et al.* (2019) Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* 20, 389–403
47. Azodi, C.B. *et al.* (2019) Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3 Genes Genomes Genet.* 9, 3691–3702
48. Angermueller, C. *et al.* (2016) Deep learning for computational biology. *Mol. Syst. Biol.* 12, 878
49. Sun, N. *et al.* (2023) Single-nucleus multiregion transcriptomic analysis of brain vasculature in Alzheimer's disease. *Nat. Neurosci.* 26, 970–982
50. Pérez-Enciso, M. and Zingaretti, L.M. (2019) A guide for using deep learning for complex trait genomic prediction. *Genes* 10, 553
51. Zingaretti, L.M. *et al.* (2020) Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Front. Plant Sci.* 11, 25
52. Montesinos-López, O.A. *et al.* (2018) Prediction of multiple-trait and multiple-environment genomic data using recommender systems. *G3 Genes Genomes Genet.* 8, 131–147
53. Montesinos-López, A. *et al.* (2018) Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3 Genes Genomes Genet.* 8, 3813–3828
54. Montesinos-López, O.A. *et al.* (2019) Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods. *Front. Plant Sci.* 10, 1311
55. Montesinos-López, O.A. *et al.* (2019) New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. *G3 Genes Genomes Genet.* 9, 1545–1556
56. Montesinos-López, O.A. *et al.* (2021) A new deep learning calibration method enhances genome-based prediction of continuous crop traits. *Front. Genet.* 12. https://doi.org/10.3389/fgene.2021.798840
57. Galli, G. *et al.* (2022) Automated machine learning: a case study of genomic 'image-based' prediction in maize hybrids. *Front. Plant Sci.* 13, 845524
58. Wang, K. *et al.* (2023) DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Mol. Plant* 16, 279–293
59. Montesinos-López, A. *et al.* (2023) Multimodal deep learning methods enhance genomic prediction of wheat breeding. *G3 Genes Genomes Genet.* 13, jkad045
60. De Los Campos, G. *et al.* (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182, 375–385
61. Pérez, P. *et al.* (2010) Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. *Plant Genome* 3, 106–116
62. Endelman, J.B. (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255
63. Pérez-Rodríguez, P. and de Los Campos, G. (2022) Multitrait Bayesian shrinkage and variable selection models with the BGLR-R package. *Genetics* 222, iyac112
64. Runcie, D.E. *et al.* (2021) MegaLMM: mega-scale linear mixed models for genomic predictions with thousands of traits. *Genome Biol.* 22, 213
65. Caamal-Pat, D. *et al.* (2021) lme4GS: an R-package for genomic selection. *Front. Genet.* 12, 680569
66. Covarrubias-Pazaran, G. (2016) Genome-assisted prediction of quantitative traits using the r package sommer. *PLoS One* 11, e0156744
67. Monier, B. *et al.* (2022) rTASSEL: An R interface to TASSEL for analyzing genomic diversity. *J. Open Source Softw.* 7, 4530
68. Wang, J. and Zhang, Z. (2021) GAPIT Version 3: boosting power and accuracy for genomic association and prediction. *Genomics Proteomics Bioinforma.* 19, 629–640

and bootstrapping can mitigate this limitation in prediction scenarios. Common implementations of MAP estimation include (1) numerical methods (e.g., Newton's method), (2) the Expectation-Maximization (EM) algorithm (avoiding the need for derivatives), and (3) Monte Carlo techniques (e.g., simulated annealing).

69. Charmet, G. *et al.* (2020) BWGS: an R package for genomic selection and its application to a wheat breeding programme. *PLoS One* 15, e0222733
70. Montesinos-López, O.A. *et al.* (2018) An R package for multitrait and multienvironment data with the item-based collaborative filtering algorithm. *Plant Genome* 11. https://doi.org/10.3835/plantgenome2018.02.0013
71. Montesinos-López, A. *et al.* (2021) A guide for kernel generalized regression methods for genomic-enabled prediction. *Heredity* 126, 577–596
72. Montesinos López, O.A. *et al.* (2022) A general-purpose machine learning R library for sparse kernels methods with an application for genome-based prediction. *Front. Genet.* 13. https://doi.org/10.3389/fgene.2022.887643
73. Montesinos-López, O.A. *et al.* (2022) A comparison of three machine learning methods for multivariate genomic prediction using the sparse kernels method (SKM) library. *Genes (Basel)* 13, 1494
74. Montesinos López, O.A. *et al.* (2023) Statistical machine-learning methods for genomic prediction using the SKM library. *Genes (Basel)* 14, 1003
75. González, F. *et al.* (2024) Introducing CHiDO – A No Code Genomic Prediction software implementation for the characterization and integration of driven omics. *Plant Genome*, Published online October 24, 2024. http://doi.org/10.1002/tpg2.20519
76. Chung, P.-Y. and Liao, C.-T. (2022) Selection of parental lines for plant breeding via genomic prediction. *Front. Plant Sci.* 13, 934767
77. de J. Villar-Hernández, B. *et al.*, eds (2024) *A Bayesian optimization R package for multitrait parental selection*. *Plant Genome* 17, e20433
78. Mohammadi, M. *et al.* (2015) PopVar: a genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. *Crop Sci.* 55, 2068–2077
79. Montesinos-López, O.A. *et al.* (2023) A marker weighting approach for enhancing within-family accuracy in genomic prediction. *G3 Genes Genomes Genet.* 14, jkad278
80. He, K. *et al.* (2016) Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778