Trends in Plant Science

CellPress OPEN ACCESS

Feature Review

Machine learning algorithms translate big data into predictive breeding accuracy

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

Statistical machine learning (ML) extracts patterns from extensive genomic, phenotypic, and environmental data. ML algorithms automatically identify relevant features and use cross-validation to ensure robust models and improve prediction reliability in new lines. Furthermore, ML analyses of genotype-by-environment (G×E) interactions can offer insights into the genetic factors that affect performance in specific environments. By leveraging historical breeding data, ML streamlines strategies and automates analyses to reveal genomic patterns. In this review we examine the transformative impact of big data, including multi-trait genomics, phenomics, and environmental covariables, on genomic-enabled prediction in plant breeding. We discuss how big data and ML are revolutionizing the field by enhancing prediction accuracy, deepening our understanding of G×E interactions, and optimizing breeding strategies through the analysis of extensive and diverse datasets.

Highlights

The genomic prediction (GP) approach that uses genotypic and phenotypic data to predict the genomic estimated breeding value (GEBV) of individuals has been widely adopted by both public and private breeding organizations. GP models can predict the performance of plant germplasm in different environments by correctly modeling genotype × environment interactions (G×E) across multiple traits.

Machine learning (ML) algorithms can help breeders to determine the most effective parental selection, mating designs, population sizes, and selection intensities to maximize selection gain at a given budget while minimizing the loss of genetic diversity.

Neural networks (NNs) have great potential in improving the accuracy of GP models in the context of 'big data'. These algorithms can identify complex patterns and relationships between genotypes and phenotypes, leading to more precise predictions of important plant traits.

The impact of data-driven strategies and ML techniques

Valuation and selection processes are crucial in plant breeding for identifying desirable traits such as disease resistance, drought and heat tolerance, and high grain yield. Testing these selected cultivars across different environments through **multi-environment trials** (**METs**; see Glossary) assists in understanding their performance and stability.

Accurate and early predictions have a pivotal role in plant breeding [1]. With advances in statistical modeling and data analysis, breeders can now predict the performance of breeding lines or cultivars with greater precision. This allows more informed and accurate decisions and also reduces the time and resources required for developing superior cultivars. Early predictions assist breeders in focusing on the most promising genotypes, thereby accelerating the breeding cycle and enhancing the efficiency of the breeding process. Nevertheless, data collected from METs are intrinsically complex owing to structural patterns, nonstructural noise, and relationships among genotypes, environments, and genotypes and environments considered jointly, namely **genotype × environment (G×E) interactions** [1]. Pattern implies that cultivars respond to specific environments (location, years, location–year combinations) in a systematic and interpretable manner, whereas noise suggests that the responses are unpredictable and uninterpretable.

Genomic markers have revolutionized plant breeding by enabling precise selection of desirable traits at the DNA level. This accelerates the breeding process, increases accuracy in predicting plant performance, reduces costs, and enhances the development of pest/stress-resistant and high-yield cultivars, thus making plant breeding faster, more efficient, and more effective [2].

[1]Louisiana State University, College of Agriculture, Baton Rouge, LA, USA
[2]Colegio de Postgraduados, Montecillos, CP 56230, Estado de México, Mexico
[3]International Maize and Wheat Improvement Center (CIMMYT), Carretera México- Veracruz Km 45, El Batán, Texcoco, CP 56237, Estado de México, Mexico
[4]Department of Statistics and Operations Research and Distinguished Scientist Fellowship Program, King Saud University, Riyadh 11451, Saudi Arabia
[5]Facultad de Telemática, Universidad de Colima, CP 28040 Estado de Colima, Mexico

Genomic prediction (GP) estimates the genetic value of an individual for a trait from its genomic data and has changed practices in plant and animal breeding since the landmark publications of Bernardo *et al.* [3] for maize hybrid prediction and Meuwissen *et al.* [4]. When used for selection (i.e., **genomic selection, GS**), it can accelerate the genetic gains of breeding programs by increasing selection precision, by reducing costs of experimental validation of selection candidates, or by reducing the time needed to make selection decisions. The latter aspect has the highest potential for accelerating breeding by identifying promising individuals at early stages for further crossing (in a shorter timeframe), thus increasing genetic gains per unit of time. GS has been quickly adopted in dairy cattle breeding where the practice of working with **estimated breeding values (EBVs)** had already become routine, especially through the work of Henderson [5] and Quaas [6]. After discussions on how GS could be transferred from animal to plant breeding programs [7], GS has been tested for many crops such as cassava, chickpea, maize, rice, and wheat [8–11]. Today it is widely adopted in different ways across public and private sector plant breeding organizations [9,12,13]. The concept of separating 'population improvement' activities from 'product development' [14] can be interpreted as a structural adaptation of plant breeding programs to make the best use of GS.

Like all statistical learning methods, GP requires training data as a standard that comprise at least genotypic and phenotypic information of individuals in a **training population (TRN)** on which a statistical model is trained. The model being trained means that the values of unspecified parameters are determined to fit the training data in the best way while respecting potential secondary conditions. The trained model can then be used for the prediction of genetic values of (un)observed individuals in a **test population (TST)**. The standard reference method is the genomic best linear unbiased prediction (GBLUP), which is a **linear mixed model (LMM) that uses** genomic marker data to derive a genomic relationship matrix that is plugged into the mixed model equations [15]. The linear model approach has been expanded to the use of different prior distributions in the Bayesian paradigm [16,17], different types of regularization [18], or different types of relationship matrices that may be motivated by statistical techniques, such as reproducing kernel Hilbert space (RKHS) [19,20] or by biological mechanisms such as epistasis [21,22]. Recently, other non-linear methods such as random forests [23,24], support vector machines [25], and artificial **neural network (NN)** models have been used for GP [26–28]

Prediction performance is highly influenced by several factors, including the tuning procedure for optimizing the TRN design, marker data quality, heritability, genetic architecture of the trait, and the relationship (or mismatch) between the TRN and the TST sets [29–31]. The statistical ML model attempts to capture the association between patterns in the genomic markers and the phenotypes of the individuals. Once the model is trained, the prediction of the genetic potential of an individual is based solely on its genomic data. However, trait-assisted GS could also be very useful when secondary traits with genomic information are used for improving the prediction of focal (primary) traits. In addition to exploiting parental relationships (by markers or pedigree) for improving the prediction of unobserved cultivars, **high-throughput phenotyping (HTP)** and phenomics through image analyses, as well as the inclusion of environmental covariables (EVs) (Figure 1) for studying G×E, are valuable tools for integrating additional information and increasing prediction accuracy [32]. Provided that the prediction is sufficiently accurate, selection decisions can be made earlier in a breeding program, and the time needed for a breeding cycle consisting of crossing, evaluation, and selection of parents for new crosses can be reduced.

An important example of a novel approach is 'phenomic (HTP) prediction' [33,34] which uses low-cost and high-throughput methods for non-target traits to predict relevant characteristics of selection candidates. GP is based on the genetic information that an organism bears and

[6]Cornell University Ithaca, New York, NY, USA
[7]Aardevo B.V., Nagele, The Netherlands
[8]Department of Plant Sciences, University of California Davis, Davis, CA, USA
[9]Departamento de Matemáticas, Centro Universitario de Ciencias Exactas e Ingenierías (CUCEI), Universidad de Guadalajara, 44430 Guadalajara, Jalisco, Mexico
[10]Norwegian University of Life Science (NMBU), Department of Plant Science, Ås, Norway
[11]Universidad de Quintana Roo, Chetumal, Quintana Roo, 77019, Mexico
[12]Australian National University, Research School of Biology, Canberra, NSW, Australia
[13]Department of Plant Breeding, Swedish University of Agricultural Sciences (SLU), PO Box 190 Sundsvagen 10, SE 23422 Lomma, Sweden

*Correspondence:
Alison.Bentley@anu.edu.au (A. Bentley) and rodomiro.ortiz@slu.se (R. Ortiz).

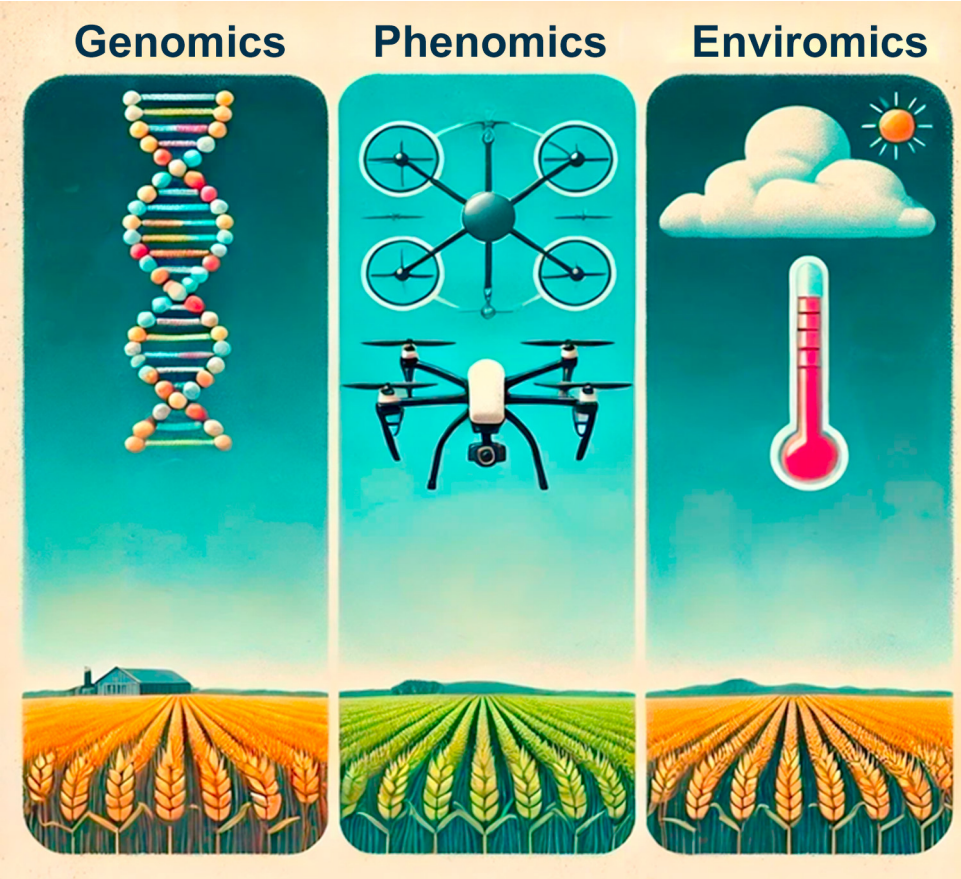

Figure 1. Components of modern plant breeding include not only phenotypic data collected from observed field cultivar trials but also genomic (molecular markers), phenomic (images from drones, airplanes, satellites), and enviromic (temperature, sun radiation, precipitation, soil humidity) data.

can pass on to its offspring; in other words, GP is centered on the genetic potential of future generations whereas phenomic prediction focuses on the observable traits of the organism at a specific point in time, and thus provides insights into the present phenotypic state.

With 'big data', more complex statistical ML models such as NNs are gaining importance. Functional annotation of genetic variants uses ML for learning complex patterns from genomic data and predicting the potential functional consequences of genetic variants [35]. Deep learning NN architectures such as **convolutional neural networks (CNNs)** and **recurrent neural networks (RNNs)** have been used to analyze genomic sequences, regulatory elements, and epigenetic features, thereby aiding in the identification of functional variants [36,37]. ML models have played a significant role in predicting gene expression levels based on genomic features [38]. We provide here a comprehensive review of the latest developments in ML for GP in plant breeding, and highlight key approaches, challenges, and future directions.

Bases for modern plant breeding

With the advent of cheaper and more accurate genotyping and phenotyping techniques, as well as the ability to collect multi-omic data and large-scale environmental information, the integration

Glossary

Arc-cosine kernel (AK): a mathematical function used in ML, particularly for measuring the similarity between datapoints in high-dimensional spaces.
Bayesian multi-output regressor stacking (BMORS): combines predictions from multiple regression models using a Bayesian approach to provide a robust and probabilistic framework for handling uncertainties in multi-output scenarios.
Coefficient of parentage (CoP): measures the genetic relationship between two individuals in a pedigree, indicating the proportion of genes they share due to common ancestors.
Convolutional neural networks (CNNs): designed for processing structured grid data and images. Utilizes convolutional layers to learn hierarchical features automatically and adaptively from the input, thus being highly effective for image recognition and computer vision tasks.
Crop growth models (CGMs): mathematical representations used to simulate and predict the growth and development of crops over time by incorporating environmental factors and management practices.
Estimated breeding value (EBV): predicts the genetic contribution one individual will pass on to its offspring. It is calculated based on the performance both of the individual and of its relatives, and often considers environmental factors. The idea is to separate the genetic component from the environmental effects.
Fivefold cross validation (5FCV): a common technique in ML for model assessment. The dataset is randomly partitioned into five subsets, and the model is trained and evaluated five times, each time using a different subset as the testing set and the remaining subset as the training set. 5FCV helps to assess the performance and generalization of a model across different data subsets.
Gaussian kernel (GK): a mathematical function used in ML for transforming data and capturing complex relationships, especially in non-linear tasks.
Generalized Poisson regression (GRP): allows overdispersion, thus accommodating situations where the variance exceeds the mean. It uses a broader class of distributions, such as

of cutting-edge technologies has given rise to the concept of 'the modern plant breeding triangle' [32]. This represents the synergistic combination of genomics, phenomics, and enviromics (Figure 2) in data analytics and predictive breeding. The power of genomic data, that allow exploration of genetic variation, is seamlessly combined with phenomic data (HTP) to capture complex trait information on a large scale. In addition, environmental data, encompassing diverse climatic and agronomic factors, contribute crucial contextual insights.

Applying genomic selection in crop breeding

Machine learning (ML) can be applied at various stages of a breeding program of a self- or cross-pollinated crop, but its impact is particularly significant in the early generations of breeding

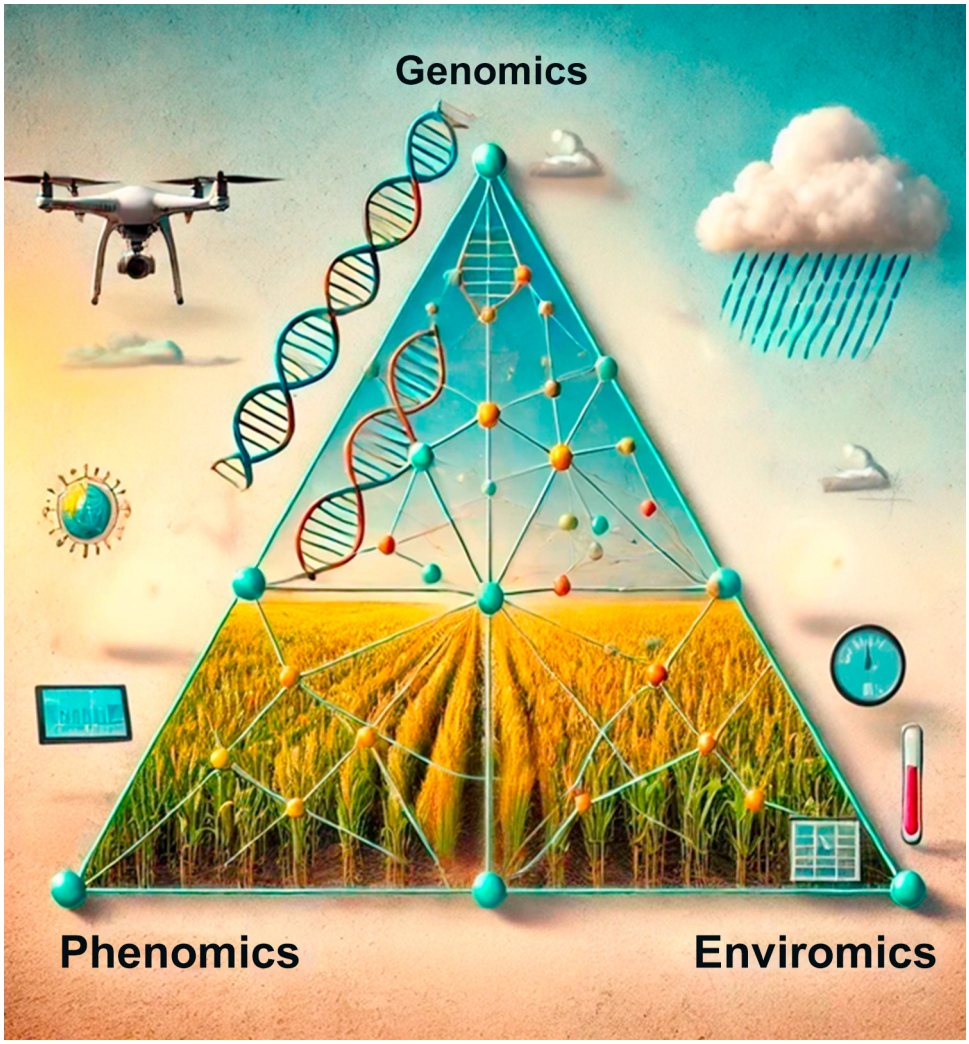

Figure 2. Summary of modern plant breeding study by interrelating genomics, phenomics, and enviromics. Machine learning (ML) statistical models and methods as well as deep learning models offer scientific solutions for efficient interrelations of these three components that can increase the prediction accuracy for cultivars that were not observed. Phenomics may include images from drones, airplanes, and satellites. Satellite images can come from optical sensors or radar sensors. Radar sensors use radio waves to create images that are known as radar or synthetic aperture radar (SAR) images. Radar can penetrate clouds and can be used day or night.

the negative binomial distribution, to better capture variability in count data.
Genomic estimated breeding value (GEBV): incorporates genomic information, such as DNA markers or genetic data, into the estimation of the breeding value of an individual. Genomic data provide a more direct assessment of the genetic makeup of an individual.
Genomic prediction (GP): estimates the genetic value of a cultivar without observing its phenotypic performance in the field.
Genomic selection (GS): selects individuals that were not observed but rather predicted.
Genotype × environment (G×E) interaction: the differential response of individuals when observing them in diverse environmental conditions.
High-throughput phenotyping (HTP): involves rapidly and systematically measuring and analyzing phenotypic traits in large-scale, often automated, ways.
Item-based collaborative filter (IBCF): a recommendation system technique based on previous positive interactions/relationships.
Leave one environment out cross-validation (LOEO): a technique in which each datapoint is tested against a model trained on all other environments, thereby ensuring robust performance evaluation across diverse settings.
Linear mixed model (LMM): combines fixed effects, representing systematic influences, with random effects to account for variability in data. LMM is used in statistics for complex, correlated data analysis.
Machine learning (ML): statistical models or methods that are used to make (genomic) predictions of unobserved cultivars.
Matrix factorization (MF): an ML technique that decomposes a matrix into the product of two or more matrices, revealing latent factors and patterns within the data.
Mega LMM: Bayesian latent factor-based LMM that integrates latent factors into a Bayesian framework for improved data modeling and analysis, particularly in the context of mixed-effects models.
Multi-environment trials (METs): field evaluation trials where cultivars are sown in different environments.
Multi-trait multi-environment (MTME) data: these involve studying multiple traits across various environments, thus providing a

when crosses are made (F_1) and segregating populations (F_2, F_3, F_4) are observed in a self-pollinated crop and the best cultivars are selected. A leading example of cutting-edge multi-omic selection in plant breeding, utilizing ML approaches, can be found in the bread wheat breeding program of the International Maize and Wheat Improvement Center (CIMMYT) (Figure 3).

At the beginning of the breeding cycle, the best parents and the best crosses are selected using different genomic-based approaches. The Bayesian decision theory applied via a multi-trait approach including the combined use of a GS index, as well as genetic diversity-related information, are currently used in the parental selection step [39,40]. The best crosses are then predicted using a simulation algorithm that incorporates both genome-wide markers and phenotypic records of the candidate parents [41]. Given that genotyping at the F_2 generation is cost-prohibitive, ML can be effectively applied at early stages in the breeding program. At the F_3 generation, initial phenotype screening assists in identifying promising individuals that should be selected, and a subset of selected individuals can be gathered. At the F_4 generation, family selection using pedigree (coefficient of parentage, matrix *A*) together with HTP for selecting lines within a family should assist breeders in achieving a rapid cycle advance. Genotyping is performed at F_5, and genomic selection is implemented for predicting the performance of new unobserved lines. In this generation, **genomic estimated breeding values (GEBVs)** are used to discard the worst lines. Note that the rapid cycle GP could include selecting parents from F_4 and/or F_5 to go directly for crossing. To manage costs, phenotypic screening can be prioritized in the early generations (F_2 and F_3), with more extensive genotyping and ML applications

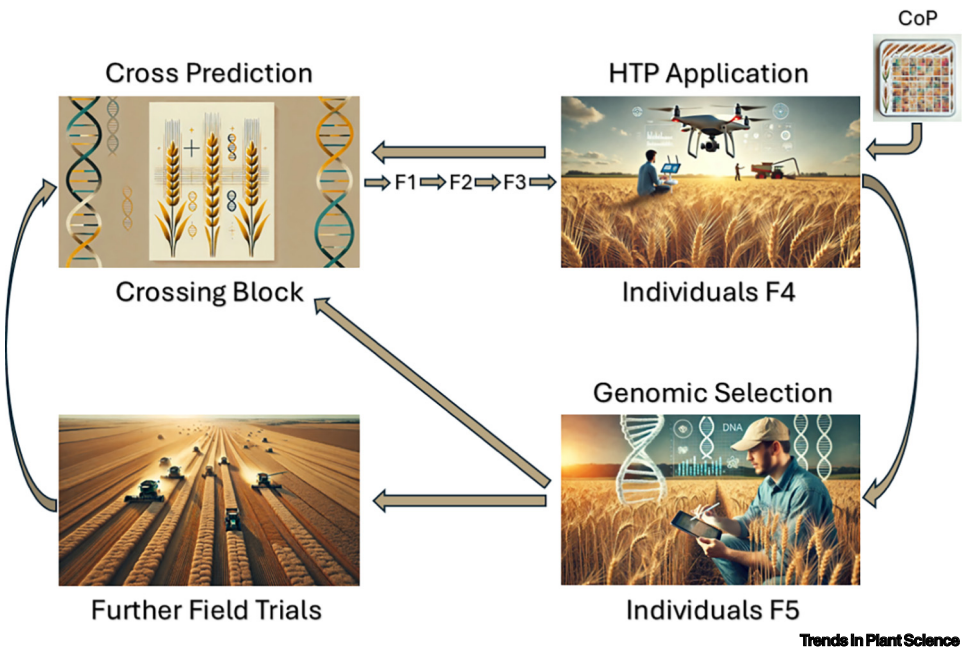

Figure 3. The sequence of potential events in a rapid-cycle genomic selection breeding program for a self-pollinated crop (wheat). Superior cultivars selected in the F_4 generation can be transferred for crossing and cross-prediction or advanced to the F_5 generation. In the F_5 generation, cultivars are genotyped, and the best performers are either moved directly to crossing or advanced to field trials. Abbreviations: CoP, coefficient of parentage; HTP, high-throughput phenotyping.

comprehensive analysis of genetic performance under diverse conditions.
Multi-trait partial least squares (MT-PLS): extends the PLS method to simultaneously analyze and model relationships between multiple sets of variables or traits. It is used for handling multivariate data and capturing complex interdependencies.
Neural networks (NNs): computational models inspired by the brain for processing data, enabling ML and pattern recognition in diverse applications.
Partial least squares (PLS): a statistical method for modeling relationships between independent and dependent variables, especially in situations with high dimensionality and collinearity. PLS finds latent factors that explain the variance in both the predictor and response variables.
Recurrent neural network (RNN): a type of neural network designed for sequential data, allowing information persistence for tasks such as language modeling and time-series analysis.
Target population of environments (TPE): sets of environments with particular climatic and environmental similarities.
Testing population (TST): the unobserved population to be predicted.
Training population (TRN): the observed population to be used as predictor of the individuals in the TST.

beginning in the F_4 or F_5 generation. This approach balances the costs and benefits of advanced data analysis.

Note that Figure 3 shows that shortening the cycle consists of (i) advancing lines quickly from F_1 to F_2 (obtained from each F_1) and rapidly deriving segregating populations, F_3, F_4, or F_5. (ii) Early sparse testing is then performed on lines at locations from the **target population of environments (TPE)** (Box 1) (sparse-tested lines consist of not planting all lines in all locations of the TPE and genomically predict those cultivars that were not observed in some locations). (iii) The selected parents may then be recycled based on GEBVs (genomic selection in wheat breeding is discussed in Box 2).

Genomic selection (GS) in wheat breeding shows promise for predicting genotypic values, considering both additive and non-additive effects. Although simulations highlight the efficiency of rapid-cycling GS for parental selection, its practical application in wheat and other crops is still limited (Box 2).

Linking genomics and phenomics

Interconnection in modern plant breeding implies the modernization of the statistical and quantitative genetic models for the analysis of plant breeding outcomes in METs. This has become clearer as the availability of genomics, phenomics, and environments information has increased [32]. Genetic gain refers to the improvement of desired traits through the selection and breeding of superior genotypes. Linking massive genomic and phenomic datasets can achieve further genetic gains but it has complexities that require a type of ML that deals with a very large number of correlated predictors (images) that might be introduced as covariables in the ML model. **Multi-trait multi-environment (MTME)** data take advantage of correlations between different traits evaluated across diverse environments to train accurate GS models. The use of GS in MTME data is a promising approach to reduce field phenotyping efforts.

Linking genomic and phenomic (e.g., HTP) data involves integrating genetic information with phenotypic traits to understand the genotype–phenotype relationship. Genomic data provide information about the genetic makeup of an individual, whereas phenomic data encompass observable traits or characteristics. By linking these two types of data, researchers can identify genetic markers that are associated with specific phenotypic traits and gain insights into the underlying biology. However, with the use of GS based on whole-genome marker data it is no longer necessary to precisely characterize the underlying genetic architecture of traits, particularly the

Box 1. Genomic sparse-testing approach

Sparse-testing methods have been proposed by researchers to improve the efficiency of GS in breeding programs. Montesinos *et al.* [83] evaluated four genomic sparse-testing methods for testing allocation of lines to environments under multi-environmental trials (METs) for genomic prediction (GP) of unobserved lines. The sparse-testing methods described in this study build the genomic training and testing sets in a strategy that allows each location or environment to evaluate only a subset of all genotypes rather than all of them. Several interesting findings were reported by the authors.

(i) The multi-trait model produced better GP accuracy than the single-trait model. Even under a scenario where we used a training–testing proportion of 15–85%, the prediction accuracy of the four methods barely decreased. This indicates that genomic sparse-testing methods for datasets under these scenarios can save considerable operational and financial resources with only a small loss in precision, which can be shown in our cost–benefit analysis.

(ii) Under the cost–benefit analysis, we observed the savings that breeders can obtain by using a genomic sparse-testing approach. For example, in a sparse-testing design with a training–testing scenario of 85–15%, under a fixed budget, breeders can increase the number of lines under evaluation by at least 17%. Under a sparse-testing scenario of 50–50% for training versus testing, the same fixed budget increased the lines under evaluation by at least 101%. Certainly, the larger the percentage of testing regarding the percentage of training, the larger the benefits of the sparse-testing method.

Box 2. Response to genomic selection for grain yield (GY) in CIMMYT spring bread wheat

Bonnett *et al.* [84] investigated increasing genetic gain for improving GY by using early-generation genomic selection. Part of this experiment compared the predictive ability of the different GEBV calculation methods in the F_2 generation by using a set of single plant-derived $F_{2:4}$ lines from randomly selected F_2 plants. GY results showed a significant positive correlation between observed yield of $F_{2:4}$ lines and predicted yield GEBVs of F_2 single plants from Gaussian kernel (predictive ability of 0.248, $P = <0.001$) and GBLUP (0.195, $P = <0.01$). Results demonstrate the potential for application of genomic selection in early generations of wheat breeding and the importance of using the appropriate statistical model for GEBV calculation, which may not be the same as the best model for inbreds.

Rapid-cycle recurrent genomic selection (GS) in CIMMYT spring bread wheat

GS in wheat breeding programs holds significant promise for predicting the genotypic values of individuals, where both additive and non-additive effects influence the final breeding value of lines. Although several simulations have underscored the efficiency of a rapid-cycling GS strategy for parental selection or population improvement, their practical application remains limited in wheat and other crops. In their study, Dreisigacker *et al.* [85] illustrated the potential of rapid-cycle recurrent GS (RCRGS) to enhance genetic gain for improved GY in wheat. Results consistently demonstrated realized genetic gains for GY after three cycles of recombination (C_1, C_2, and C_3) of biparental F_1s, when aggregated across 2 years of phenotyping. Over the combined evaluation years, genetic gain through RCRGS reached 12% from cycle C_0 to C_3, with a realized gain of 0.28 metric tons per hectare per cycle, elevating GY from C0 (6.88 metric tons per hectare) to C3 (7.73 metric tons per hectare). RCRGS also correlated with specific changes in essential agronomic traits (days to heading, days to maturity, and plant height) that were measured but not specifically selected for. To address these changes, we recommend implementing GS in conjunction with multi-trait prediction models.

complex agronomic and physiological characteristics under selection. Using genomic data it is possible to make predictions based on the additive or total genetic effects of many markers underlying complex traits. In addition, the integration of image data in the early stages of a breeding process when large segregating populations are already in the field can provide a similar resolution and definition of the complexity of the desired related sib phenotypes, such as plant morphology or disease symptoms, and this can further enhance genomic parental prediction.

Genomics with phenomics refers to the integration of wide-genomic markers and HTP data to improve prediction accuracy and gain a comprehensive understanding of the genotype–phenotype relationship. Instead of viewing them as separate entities, combining genomic and phenomic data allows a more holistic approach to understanding complex traits. It is not a matter of genomics versus phenomics but rather of leveraging and optimizing the use of both types of data to obtain a more complete picture and improve the predictive ability of the models. Genomics relies on HTP in plant breeding for several important reasons that are briefly described. HTP enables the rapid and accurate measurement of various plant traits on a larger scale than is possible using traditional manual approaches. This includes traits related to grain yield, disease resistance, drought tolerance, nutritional content, and other agronomically important characteristics. By quantifying these traits in a high-throughput manner, researchers can generate comprehensive phenotypic datasets that provide valuable information for understanding the genetic basis of traits and identifying desirable traits for breeding purposes. It can also use component or correlated traits for the purpose of predicting more complex phenotypes.

The goal of genomics in plant breeding is to understand how genetic variations influence phenotypic traits that are relevant to achieving breeding objectives (e.g., agronomic performance, yield, and product quality). HTP allows simultaneous evaluation of many genotypes, thereby enabling the identification of associations between specific genetic markers or variations and phenotypic traits. By combining genotyping data with HTP data, researchers can uncover the underlying genetic architecture of traits and identify marker–trait associations, which can inform breeding strategies and facilitate the selection of desirable genotypes. A recent study [42] showed the power of high-density phenomic information in predicting complex traits (including yield) in elite wheat breeding material. Because this only predicts phenotypes, it will only be reflected in an increasing

rate of genetic gains if combined with genomic-enabled prediction of genetic values (or additive genetic values). Furthermore, HTP provides breeders with the necessary tools to evaluate many plants or individuals efficiently. This allows the identification of individuals with desirable phenotypic traits such as high yield, disease resistance, or improved quality. By integrating HTP with genomic data, breeders can select genotypes based on both their genetic profiles and their phenotypic performance. This integrated approach of genomics and phenomics allows measurement of more plants, increased selection intensity, improves the efficiency of the breeding program, and thus accelerates the development of improved cultivars.

HTP allows rapid evaluation of large breeding populations or gene bank accessions, thus significantly reducing the time required for breeding cycles. Most phenotyping methods are often time-consuming, labor-intensive, and limited in their capacity to handle large numbers of individuals. HTP technologies such as automated imaging, sensor-based measurements, and robotics enable the efficient collection of phenotypic data from thousands of plants in a short time.

In summary, HTP is essential at the early stages of population improvement because (i) it enhances the identification of different individuals within a family that cannot be identified by the only use of the *A* matrix, and (ii) trait quantification facilitates the association between genotypes and phenotypes, enhances selection and breeding strategies, enables estimation of genetic gain, and accelerates breeding cycles. By combining genomics with HTP, plant breeders can make more informed decisions and achieve faster progress in developing improved cultivars with desired traits.

A novel statistical ML approach that combines functional regression for modeling the HTP images and genomic information or **coefficient of parentage (CoP)** (matrix *A*) to enhance GP in plants has been developed, and has potential to improve the breeding strategy and enhance crop productivity [43,44]. The authors use the conventional GBLUP ML model for GP but include the interaction between the HTP images and environments. This approach aims to improve the accuracy of predicting complex traits in plants by incorporating genomics and phenomics. This integration allows more comprehensive understanding of how genetic and environmental factors interact to influence plant traits. The proposed model considers both marker-by-marker and marker-by-environment interactions, and it uses genomic information from multiple markers simultaneously. The authors demonstrate the effectiveness of their approach through extensive real-world plant genomic data. A previous article indicated that using all bands (wavelengths) from HTP data produced better prediction accuracy than using vegetation indices [45].

Jointly modeling all bands and yields improved genetic value prediction accuracy when using **MegaLMM** [46], which is a Bayesian latent factor-based LMM approach for partitioning phenotypic correlations into genetic and environment components, thereby fully capturing the shared genetic information from HTP data and yield. Modeling non-linear genotype–phenotype functions by using RKHS improved the prediction of grain-yield genetic values from genotype plus HTP [46]. Breeders traditionally evaluate the performance of different cultivars by measuring traits related to yield, such as plant height, leaf area, and flowering time. However, this process can be time-consuming and resource-intensive, especially when large populations need to be evaluated. Aerial HTP offers a solution by using remote sensing techniques, such as drones or satellites, to capture high-resolution images of crops from above. These images can provide valuable information about plant health, growth patterns, and other phenotypic characteristics. Advanced image analysis algorithms are then applied to extract quantitative data from the images, thereby enabling the identification of key traits associated with grain yield.

Krause *et al.* [47] indicated that plant breeders can assess many individual plants or plots quickly and non-destructively, even when seed availability is limited. This enables them to make early selections based on predicted grain-yield potential, thereby accelerating the breeding process. The indirect selection approach reduces the time and resources required for field-based evaluations and enables breeders to prioritize promising genotypes for further development. Interestingly, Montesinos-Lopez *et al.* [48] evaluated the GP of the item-based collaborative filtering approach for predicting cultivars using MTME modeling. This approach proved to be the best strategy for cross-nursery (small plot) predictions.

In plant breeding, phenotypic values are very noisy, and new models must be able to integrate not only genotypic and environmental data but also HTP collected by breeders with advanced image technology [49]. These can be explored using **generalized Poisson regression (GPR)** for genome-enabled prediction of count phenotypes using genomic and HTP data. The GP model GPR allows the integration of input information from many sources including environments, genomic data, high-resolution data, and interaction terms between these three sources. The authors found that the best prediction performance ML was obtained when all information was considered in the predictor.

Integrating HTP and genomic information into prediction models significantly enhances the prediction performance of genomic plus phenotypic values compared to using only genomic information in soft winter wheat [50]. The results of Montesinos *et al.* [50] support the importance of incorporating phenotypic information to enhance prediction accuracy in GS and highlight the significant potential of phenomic data in improving GS by providing superior predictions compared to genomic information alone. We envisage substantial opportunities to improve the collection and processing of HTP as well as to refine the overall modeling process through optimal integration of genomics, phenomics, and other sources of information.

However, HTP combined with genomic data cannot be compared to genomic data alone in terms of genetic value prediction accuracy. HTP coupled with genomic data can be difficult to interpret because they capture non-genetic sources of variation through the phenomic data. Despite this challenge, integrating HTP with genomic data can provide a more comprehensive understanding of plant traits and their environmental interactions. Most of the additive genetic effects, which are crucial for achieving genetic gains, come from genomic data, with a smaller contribution from HTP. This underscores the importance of genomic markers in an effective plant breeding strategy while recognizing the complementary role of HTP.

Linking genomics and enviromics

Environmental signals drive gene regulation and transcription, post-translational modifications of proteins, and the production of hormones and other metabolites that drive plant response and plasticity in the field together with temperature, light, and water. Hence, it is crucial to consider those effects to expect accurate GP, especially if the breeder is interested in studying the performance of the genotypes under multiple growing conditions. However, for training the models it is essential to keep in mind that every observed trait phenotype is the result of an intrinsic environmental influence that can only be estimated by measuring the trait variations across multiple environments, such as a specific growing scenario, a combination of management, soil condition, weather condition, elevation, and microbiome, or simply for research purposes by combining planting date with location and agronomic management.

A 'short' way to establish the effect of the environment on the phenotypic variation is to fit any method that associates environmental features (e.g., weather conditions, soil conditions, abiotic

stresses) with the actual trait values measured from METs. The metric from this is called the 'reaction-norm', which is a snapshot of the potential 'phenotypic plasticity' of a given genotype – the main source driving the observable G×E phenomena [51]. The process of collecting, processing, and using environmental data for this is called 'envirotyping' [52], in which their high-throughput use across multiple omics and levels (cell > tissue > plant > plot > field trial > location > region > mega-environment) is called 'enviromics' – or the study of the plant envirome. Consequently, the "breeder's eye" only sees one of the multiple facets of the envirome that drive a wide number of biological processes that culminate in a particular trait level and quality in the field.

Since 2014, when the first investigations linking genomics and envirotyping data for prediction were conceived [53,54], diverse new methods have been introduced. Essentially, they can form three groups, as described in the following sections.

Group 1. GBLUP environmental expansion was introduced by Jarquin *et al.* [54] and expanded on by diverse authors [55–57]. These models are conceived as expansions of the conventional GBLUP, thus accommodating multiple genetic kernels for pedigree [55], nonadditive effects [56,58], and even unknown variance–covariance structures [58,59]. From the 'enviromic side', Costa-Neto *et al.* [60] expanded this model to accommodate multiple environmental kernels due to different development stages or features (e.g., temperature-related kernel plus soil conditions-related kernel). Multiple environmental similarity has been used to predict future environments [57,61], thus helping plant breeders to use GP to anticipate near-future climate changes and their impact based on current breeding strategies. One common concern about these methods was the lack of additivity among the environmental covariables. Costa-Neto *et al.* [62] suggested replacing the quantitative covariables with 'environmental types' that were built up by breaking the continuous distribution into classes (typologies) and then measuring the frequency of each across a particular time or developmental stage. The so-called T-matrix is then used to ensure the additivity (sum of frequencies = 100%) and can be used to feed GP models (by environmental similarity, group 1) through other approaches such as environmental characterization (Group 4).

Some researchers [56,63] introduced the use of nonlinear kernels (**Gaussian kernel, GK** and **Arc-cosine, AK**) with a hierarchical Bayesian distribution of joint genetic and environmental effects, which proved to outperform the conventional linear way in terms of model accuracy and resolution (accuracy for a specific genotype) in tropical maize. This approach was expanded and tested for a large wheat dataset [57]. This research demonstrated that the conventional GBLUP has a poor ability to predict a future season (new genotypes in a new year), and the inclusion of envirotyping data into a linear model was outperformed using nonlinear kernel (e.g., GK) in terms of accuracy, resolution, and stability of the prediction (lower dispersion, more reliable predictions).

Group 2. Integrating crop modeling includes researchers specialized in fields such as crop simulation models (also known as **crop growth models, CGMs**) which bridge diverse fields such as agronomy, irrigation and nutrient management, and policy making [64]. One advantage of this method is the ability to explore the cultivar-specific CGM parameters as 'meta phenotypes', which are trained from the phenotype and envirotyping data and then used as a response variable (trait) in GP. Technow *et al.* [65] developed the so-called CGP-WGP (crop growth model – whole GP) method, later expanded by Cooper *et al.* [66] and Messina *et al.* [67]).

Group 3. Environmental indices have their roots in the classical works on quantitative genetics and ecology in the 1960s where linear regressions were used to empirically associate phenotypes

and an index of the environmental gradient as a proxy to measure phenotypic plasticity. In the modern era of GP [53], and to some degree the authors of Group 2 as well, the environmental index used for its models was obtained from CGM outcomes. An advantage of using CGM is the ability to extract an environmental index that has a higher ecophysiological significance (e.g., drought or heat stress index derived from process-based mechanistic CGMs). Ly *et al.* [68] used a supervised approach to learn an empirical association between environmental gradient, genotype sensibility, and its impact on G×E. Millet *et al.* [69] applied a genomic version of the factorial regression not only to predict grain yield in maize but also to learn the environmental driver of G×E. For this reason, this approach is useful when the interest of the research is to learn the G×E patterns and use them to select the most adapted cultivars [57].

The methods from groups 2 and 3 have a better resolution in detailing genotype-specific sensitivities to environmental conditions, which could be used as traits to perform other approaches such as association mapping. This latter is useful in exploring the genetic architecture of the phenotypic plasticity, consequently expanding the "breeder's eye" to a level not investigated before. However, methods from group 1 also open the breeder's eye to the importance of the environmental diversity and not the number of environments to be tested, and interprets G×E as a mix of multiple similarities that can be explored and used to optimize the predictive breeding protocols.

An overall view of the process for linking genomics and environmental covariables is shown in Figure 4. This figure displays how to associate enviromics, DNA sequencing, HTP images from satellites, airplanes, and/or drones, and phenotypic data in R codes. All information in Figure 4A,B is combined to conduct supervised and unsupervised learning tools to predict phenotypes across single or multiple environments or multiple traits.

Linking genomics, phenomics, and enviromics

Integrating genomics, phenomics, and enviromics offers a holistic approach that should improve the accuracy and reliability of GP compared to using only one or two of these components. Genomics provides genetic information, phenomics offers insights into observable traits, and enviromics accounts for environmental influences. Together, they create a comprehensive dataset that captures the complex interactions between genes, traits, and the environment, leading to more precise and robust predictions. Montesinos *et al.* [70] predicted partially tested lines and untested environments to assess the benefits of adding environmental covariates to genomic and phenomic information. The authors evaluated the accuracy of predictions by randomly excluding one environment and using it as the training set while utilizing the remaining environment as the testing set. The authors found that including environmental data increased prediction accuracy by 60% on average, and including all five datasets showed improvement. The smallest gain was ~6%, and the largest was ~101%. These results suggest therefore that incorporating additional inputs can significantly enhance prediction accuracy, but caution is necessary. We recommend using feature-selection techniques, such as Pearson's correlation and Boruta, to ensure effective integration of environmental covariates into the overall ML method.

G×E interactions

In plant breeding, environmental influences and the interaction of genotypes with the environment are crucial factors. On the one hand, breeders aim at breeding for 'stability', namely to achieve stable high-yield cultivars under any conditions; however, on the other hand, such a breeding approach may hamper selection gain for specific environmental conditions. A crucial point in breeding is therefore to define the target population of environments or target region(s).

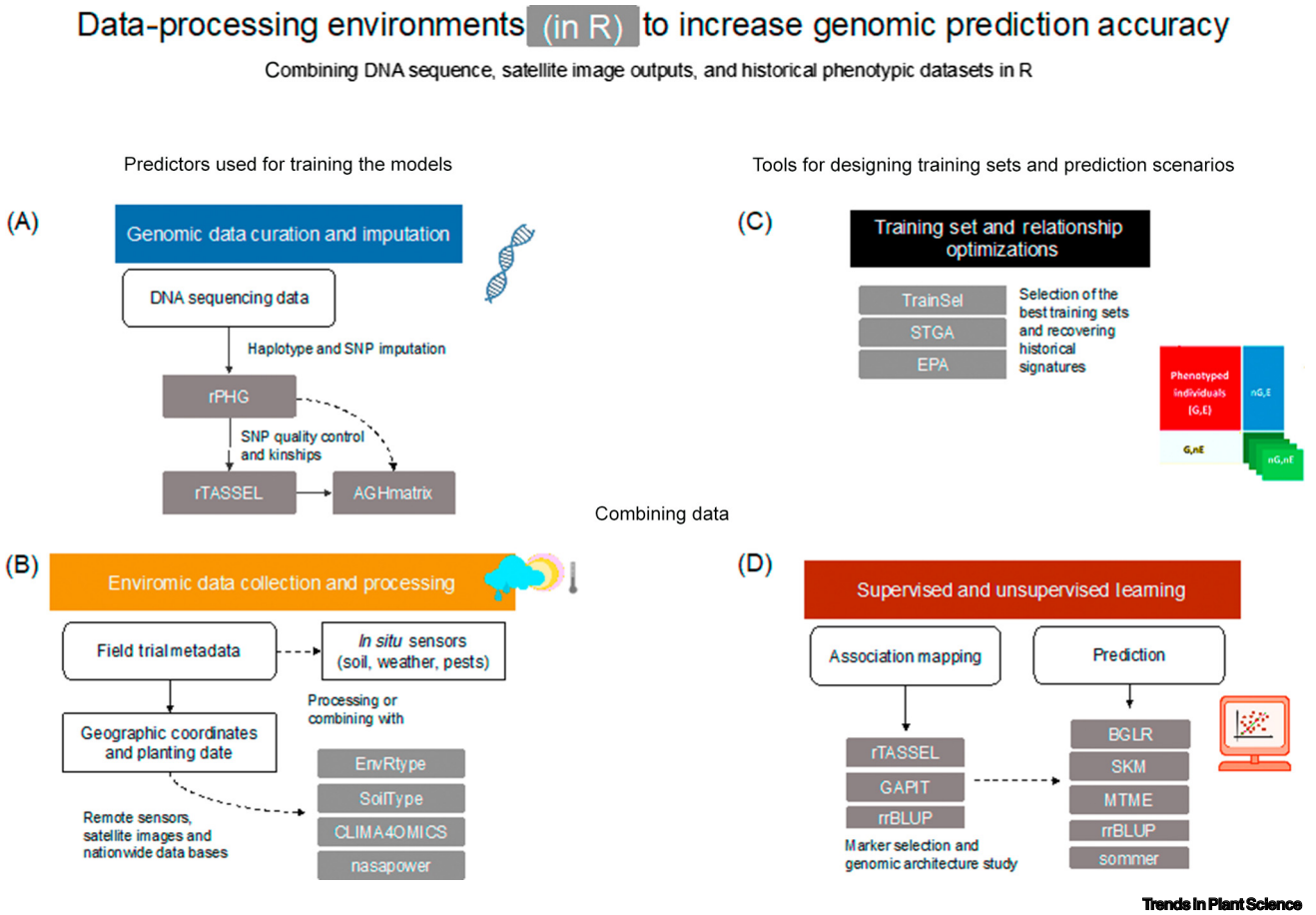

Figure 4. Open-source packages containing machine learning (ML) methods for an accurate genomic prediction (GP) pipeline. (A) Genomic data curation and imputation for calling SNPs and computing environmental relationship matrices to feed prediction models. (B) Enviromic data curation based on in field metadata (geographic coordinates, management, and *in situ* sensors) using (C) global-scale predictions of environmental features to be used either as predictors for multi-environment GP or to design training sets for (D) supervised or unsupervised learning. Abbreviations: E, environment; G, genotype; nE, number of environments; nG, number of genotypes; SNP, single-nucleotide polymorphism.

To account for G×E interactions, standard linear mixed effects models have been extended to incorporate multi-environment predictions. These models aim to predict the genetic value of a line or a hybrid across all environments as well as in specific environments, thereby allowing the breeder to select according to specific strength of the respective selection candidates. These models use marker and environmental classifications or detailed environmental covariates to define covariance structures of random effects [54,71]. The covariance structures are given by genomic markers, environmental covariates, and Kronecker (or Hadamard) products of both covariance matrices to model the covariance of the interactions [54]

Following Crossa *et al.* [9], we start from the baseline model for phenotypes evaluated in different environments (y_{ij}) given by:

$$y_{ij} = \mu + E_i + L_j + EL_{ij} + \varepsilon_{ij} \quad [1]$$

where μ is the overall mean, E_i ($i = 1,\ldots,I$) is the random effect of the i^{th} environment, L_j is the random effect of the j^{th} line ($j = 1,\ldots,J$), EL_{ij} is the interaction between the i^{th} environment

and the j^{th} line, and e_{ij} is the random error term. The assumptions are as follows: $E_i \overset{iid}{\sim} N(0, \sigma_E^2)$, $L_j \overset{iid}{\sim} N(0, \sigma_L^2)$, $EL_{ij} \overset{iid}{\sim} N(0, \sigma_{EL}^2)$, and $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$, where *N(.,.) denotes a normal distribution, and iid indicates independent and identically distributed.*

In this setup, environments are generally treated as random effects, indicated by the assumption of following a normal distribution, but they can also be modeled as a fixed effect. So far, the baseline model does not use information about genomic markers or environmental covariates.

Genomic markers can now be introduced in model (Equation 1) by introducing a covariance structure deviating from *iid* in the underlying assumptions. To indicate the difference, line L_j is replaced by g_j that represents the additive genetic value of the j^{th} line (often lines L_j and g_j are both used such that L_j can account for the remaining non-additive genetic covariance). The vector containing the genomic values is $\boldsymbol{g} \sim N\left(\mathbf{0}, \mathbf{G}\sigma_g^2\right)$, where σ_g^2 is the genomic variance component, and $\mathbf{G}$ is the genomic relationship matrix. This model change enables the genomic data to be included in a mixed model approach. Likewise, as an alternative, if a pedigree relationship matrix is available but genomic markers are not, the effect of line L_j can be replaced by a_j, where $\boldsymbol{a} \sim N(\mathbf{0}, \boldsymbol{A}\sigma_a^2)$, in which $\boldsymbol{A}$ is the numerical additive relationship matrix derived from pedigree and σ_a^2 is the additive variance. Analogously, we can replace the $E_i \overset{iid}{\sim} N(0, \sigma_E^2)$ by an environmental effect vector $\boldsymbol{e} \sim N(0, \sigma_e^2\, \mathbf{E})$ which incorporates environmental covariates in the covariance structure.

The G×E interaction covariance matrix is then the Kronecker product of the two covariance structures, the first describing relationships between lines based on genetic information (pedigree or genomic) and the other matrix relating environments by means of environmental covariates. When environmental covariables are used, the name 'reaction norm' is justified because the genotypic effect is a reaction to those environmental covariables.

As a comment on the use of the Kronecker or the Hadamard product: there is an equivalence between the required mathematical formulation including a Hadamard product as used by Jarquín *et al.* [54] and an alternative formulation using the Kronecker product [72,73]. Both can be used, but – as a conceptual remark – Kronecker products describe the interaction between two different groups of variables, whereas Hadamard products describe interactions within a group. Kronecker products increase the dimension from a J × J covariance matrix for the genotypes and an I x I covariance matrix for the environments to a (J × I) × (J × I) covariance matrix of G×E. An example of the use of the Hadamard products is epistasis, namely G×G interactions, in which we model the interaction within the genome of a genotype. Each individual genotype has therefore one interaction term. The dimension of the corresponding covariance matrix therefore remains J × J. A mathematical description of how Hadamard and Kronecker products can be used for modeling G×E is given in Martini *et al.* [73].

An additional fine-tuning of the original reaction norm model uses a non-linear Gaussian kernel in marker × environment interactions [63,64]. Assessing G×E with a non-linear Gaussian kernel often outperforms linear kernel GBLUP for modeling G×E [63]. In summary, incorporating G×E commonly improves the prediction accuracy of unobserved cultivars in years or location–year combinations even when some cultivars were not included in all the environments. Further non-linear kernels often increase the contribution of G×E to GP accuracy.

Genomic prediction of multiple traits and environments

In GP, improving the accuracy of the prediction of cultivars in environments is difficult because the available information is generally sparse and usually has low correlations between traits across environments. Thus, it is necessary to improve the accuracy of the prediction models used in GP. Montesinos *et al.* [48] investigated two recommended systems techniques, **item-based collaborative filter (IBCF)** and **matrix factorization (MF)**, which are popular in the context of online marketing to recommend products or items. The IBCF works based on the similarity between items and it is calculated using people's ratings of those items and uses the items most similar to a user's already rated items to generate a list of predictions (or recommendations). The prediction of the IBCF is a weighted sum or linear regression. The authors obtained empirical evidence that both methods, IBCF and MF, work well for predicting phenotypes that are missing in some traits and environments, but the IBCF was the best if, and only if, the correlation between traits and between environments was moderately high.

The **Bayesian multi-output regressor stacking (BMORS)** model consists of two stages. In the first stage, a univariate genomic best linear unbiased prediction (GBLUP) model, including G×E, is implemented for each of the traits under study. Then, in the second stage, the predictions of all traits are included as covariates by implementing a ridge regression model. Montesinos *et al.* [74] compared the existing Bayesian MTME (BMTME) model versus the BMORS in terms of (i) genomic-enabled prediction accuracy, and (ii) potential advantages in computing resources and implementation. The authors compared the predictions of the BMORS model to those of the univariate GBLUP model, and their findings indicate that the proposed BMORS model produced similar predictions to the univariate GBLUP model and the BMTME model in terms of prediction accuracy. The proposed BMORS model serves as an alternative for predicting MTME data which are commonly encountered in genomic-enabled prediction in plant and animal breeding.

Recent research extended the study of GP to tetrasomic polyploid potato with the main objective of investigating the GP of single-trait (ST), multi-trait (MT), and multi-environment (ME) models using field trial data [75,76]. Furthermore, Cuevas *et al.* [76] investigated GP of four genome-based prediction models with GE: (i) ST reaction norm model (M1), (ii) ST model considering covariances between environments (M2), (iii) ST M2 extended. to include a random vector that utilizes the environmental covariances (M3), and (iv) MT model with G×E (M4). The best model method for predicting many of the traits was MT because it allows the exchange of information between traits and environments followed by M3 and M2, which efficiently used information between environments.

What and how to predict?

In the prediction of new environments (years or location–year combinations), ML struggles to produce reasonable predictions. In multi-environmental plant breeding trials information on environments (year) enhances the information contained in the G×E. The principal component regression relates environments with the principal components scores of the G×E matrix and it was the original idea that gave rise to the **partial least squares (PLS)** method. The PLS regression method has been proposed to describe G×E by considering the differential sensitivity of cultivars to environmental variables [77]. The single unit-trait (ST) PLS has shown empirical evidence of its effectiveness in predicting future seasons or new environments compared to ST-GBLUP [57,78]. In addition, an improved BMTME model has been proposed to capture correlations between lines, traits, and environments [79]. The GP ability of the MT PSL was explored by Montesinos *et al.* [80] who found that the MT PLS outperformed the BMTME. They concluded that the MT PLS methodology should be tried for the prediction of future seasons or new environments.

The use of MT models is not as common as ST models owing to the higher computational demands and complex G×E associated with MT models. MT models face challenges with convergence and the implementation of GP because of the large and intricate datasets involved [74,81]. However, a recent study [76] compared ST and MT models in predicting potato traits across different environments, and the MT model outperformed the ST model. One effective method for modeling complex biological events is **multi-trait partial least squares (MT-PLS)** regression. Research indicates that MT-PLS is a valuable approach for modeling high-dimensional biological data because it can handle multiple responses and address multicollinearity efficiently. Compared to ST-PLS, MT-PLS utilizes the correlation structure among traits, leading to greater statistical power and improved prediction accuracy. A recent study [80] demonstrated that MT-PLS achieved higher prediction accuracy than MT-GBLUP.

In plant breeding it is important to have methods that can handle large numbers of predictor variables and a limited number of sample observations, as well as efficient methods for dealing with high correlations among predictors and measured traits. Ortiz *et al.* [82] recently investigated the prediction performance of the PLS methods using ST and MT modeling of potato traits. The first prediction was conducted for tested lines in tested environments using a **fivefold cross-validation (5FCV)** plan, and the second prediction was for tested lines in untested environments (referred to as **leave one environment out cross-validation, LOEO**). The results show good prediction performance and the accuracy mostly exceeded a correlation of 0.5. The accuracy of different models was tested using 5FCV and LOEO. 5FCV was found to be better than LOEO. Empirical results show evidence that ST and the MTP-PLS framework is a valuable tool for predicting the context of potato breeding data. Another important use of PLS is to discover environmental signatures and recycle G×E information from historical datasets [57]. This approach would increase the ability to use past envirotyping data to predict a yet-to-be-seen year (i.e., a year without phenotypic and environmental data).

Why do ML models not work equally well for different datasets?

ML models often exhibit varying performance depending on the characteristics of the datasets they are applied to. There are several factors that influence how well a model works on a particular dataset. Some models, such as deep learning networks, perform better with large datasets, whereas simpler models such as linear regression work well with smaller datasets. Complex datasets with non-linear relationships may require advanced models such as NNs or ensemble methods (e.g., random forests, gradient boosting) to capture the underlying patterns. Simpler datasets may be adequately modeled with linear or polynomial models. Different models have different biases and variances, and these influence their ability to generalize from training data. A model with high bias might oversimplify the data, whereas a model with high variance might overfit the training data but fail to perform well on new data. Understanding the balance between bias and variance for each model and dataset is complex and not always straightforward.

Datasets with high-dimensional features or mixed data types (numerical, categorical) might benefit from models that can handle such complexity, such as support vector machines (SVMs) and tree-based models. High-dimensional datasets may also require dimensionality reduction techniques before applying an ML model. Some models are more robust to noise and outliers (e.g., decision trees, random forests), whereas others (e.g., linear regression) may be more sensitive and perform poorly in their presence. Imbalance data has an imbalanced class distribution, and models such as logistic regression might struggle without proper adjustments. Complex models such as deep NNs are prone to overfitting, especially on small or noisy datasets. Simpler models, although less powerful, may generalize better in such cases. Overfitting occurs when a model captures noise or random fluctuations in the training data, whereas underfitting happens when

a model is too simple to capture the underlying structure of the data. Both issues can lead to poor performance on new datasets, and understanding when and why they occur is not always easy. Decision trees or linear models are more interpretable, making them preferable for datasets where understanding the decision-making process of the model is important. Large models such as deep learning networks require significant computational power and memory. Simpler models might be preferred when resources are limited or when real-time predictions are needed.

ML models often require careful tuning of hyperparameters. The optimal hyperparameters can vary significantly between datasets, and finding the right configuration is often a trial-and-error process. Without proper tuning, a model that works well on one dataset may perform poorly on another. The quality of the data and the preprocessing steps applied can significantly affect model performance. Variations in data cleaning, normalization, or feature engineering across datasets can lead to inconsistent results. Poor data quality, such as missing values or noisy data, can also undermine the effectiveness of a model. Finally, the 'no free lunch' theorem in ML states that no single model is universally the best for all possible problems. This means that some models will naturally work better on some types of datasets, and there is no one-size-fits-all solution.

Concluding remarks and future perspectives

In agriculture and biological systems, GP using G×E as a bilinear (product operator) relationship between G and E is usually more appropriate in terms of prediction accuracy than a linear relationship. Multiplicative operators for studying GP including G×E have shown an increase in accuracy as compared with a linear term. The use of MTME information for GP is challenging because traits might have different degrees of correlations between themselves and the environment. In breeding, the main task is the GP of unobserved individuals in future years and/or environments or the combination of both. In METs, the principal component regression scores of G×E might improve the information of the environment to be genomically predicted. This method for enhancing the genomic-enabled selection of cultivars that were never tested in particular environments has some interpretation problems that can be improved by using the PLS regression for predicting the performance of the cultivars across environments.

Environmental information plays a crucial role as a central bottleneck in the application of modern genomic-assisted prediction tools, particularly when dealing with multiple environments. The fundamental inclusion of environmental data in the modeling process contributes significantly to the accurate prediction of cultivars across diverse growing conditions.

The synergy between genomics, enviromics, and phenomics is essential for advancing modern plant breeding. Harnessing extensive datasets, encompassing both genomic and HTP information as well as environmental data, holds the key to unlocking further genetic gains. MTME data, coupled with the seamless integration of genomic, environmental, and HTP data, presents exciting avenues for precise GP, thus effectively streamlining field phenotyping efforts and ultimately improving breeding efficiency (see Outstanding questions).

Declaration of interests

The authors declare no competing interests.

References

1. Crossa, J. (1990) Statistical analyses of multilocation trials. *Adv. Agron.* 44, 55–85
2. Bernardo, R. (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Sci.* 48, 1649–1664
3. Bernardo, R. (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Sci.* 34, 20–25
4. Meuwissen, T.H.E. *et al.* (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829

Outstanding questions

One of the key innovations in modern plant breeding compared to conventional crossbreeding is the reduction of the cycle time. This consists of advancing lines quickly to obtain F_3 or F_4 germplasm, conducting sparse testing across multiple sites within a defined target population of environments (TPE), and recycling the lines based on their breeding values. How can rapid cycle genomic-assisted breeding be effectively implemented for different cultivar species, particularly when incorporating large MTME datasets?

In recent years science has undergone a data revolution that enables the collection of large volumes of data from different sources. In breeding, this shift is reflected in the transition from relying on a small number of gene-based markers to utilizing a combination of high-dimensional genotypic data. These data can be obtained from next-generation sequencing, epigenetic information, large numbers of cultivars in evaluation trials, multiple replicates, multiple locations, multiyear datasets, environmental data, and phenomic information on different traits, as well as environmental covariables collected across location–year combinations. In addition, HTP technologies, including multispectral images from remote sensing, contribute to the generation of 'big data' that can be used for predictive breeding. Because these collected data require appropriate statistical ML approaches to extract meaningful insights, fields such as enviromics and phenomics have gained importance in statistical agriculture. Given this complexity, how can big data be efficiently used to drive big discoveries?

The results for partially tested cultivars within specific environments cannot be considered to accurately predict genetic values. Therefore, HTP combined with genomic data cannot be compared to genomic data alone regarding genetic value prediction accuracy, instead, they can only be compared in terms of phenotypic prediction accuracy. However, genomic models that incorporate HTP data can be difficult to interpret because they capture non-genetic sources of variation present in the phenomic data.

5. Henderson, C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–447
6. Quaas, R.L. (1976) Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics* 32, 949–953
7. Jonas, E. and De Koning, D.-J. (2013) Does genomic selection have a future in plant breeding? *Trends Biotechnol.* 31, 497–504
8. Roorkiwal, M. *et al.* (2016) Genome-enabled prediction models for yield related traits in chickpea. *Front. Plant Sci.* 7, 1666
9. Crossa, J. *et al.* (2017) Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975
10. Wolfe, M.D. *et al.* (2017) Prospects for genomic selection in cassava breeding. *Plant Genome* 10, eplantgenome2017.03.0015
11. Huang, M. *et al.* (2019) Use of genomic selection in breeding rice (*Oryza sativa* L.) for resistance to rice blast (*Magnaporthe oryzae*). *Mol. Breed.* 39, 114
12. Hickey, J.M. *et al.* (2017) Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet.* 49, 1297–1303
13. Gholami, M. *et al.* (2021) A comparison of the adoption of genomic selection across different breeding institutions. *Front. Plant Sci.* 12, 728567
14. Gaynor, R.C. *et al.* (2017) A two-part strategy for using genomic selection to develop inbred lines. *Crop Sci.* 57, 2372–2386
15. Henderson, C.R. (1973) Sire evaluation and genetic trends. In *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush*, pp. 10–41, American Society of Animal Science
16. Gianola, D. *et al.* (2009) Additive genetic variability and the Bayesian alphabet. *Genetics* 183, 347–363
17. Gianola, D. (2013) Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 194, 573–596
18. De Los Campos, G. *et al.* (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182, 375–385
19. Gianola, D. and Van Kaam, J.B.C.H.M. (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178, 2289–2303
20. Crossa, J. *et al.* (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724
21. Jiang, Y. and Reif, J.C. (2020) Efficient algorithms for calculating epistatic genomic relationship matrices. *Genetics* 216, 651–669
22. Martini, J.W.R. *et al.* (2016) Epistasis and covariance: how gene interaction translates into genomic relationship. *Theor. Appl. Genet.* 129, 963–976
23. Montesinos-López, O.A. *et al.* (2023) Statistical machine-learning methods for genomic prediction using the SKM library. *Genes* 14, 1003
24. Montesinos-López, O.A. *et al.* (2022) Random forest for genomic prediction. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, pp. 633–681, Springer
25. Ogutu, J.O. *et al.* (2011) A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proc.* 5, S11
26. Gianola, D. *et al.* (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet.* 12, 87
27. Pérez-Rodríguez, P. *et al.* (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3 (Bethesda)* 2, 1595–1605
28. Abdollahi-Arpanahi, R. *et al.* (2020) Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genet. Sel. Evol.* 52, 12
29. Desta, Z.A. and Ortiz, R. (2014) Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci.* 19, 592–601
30. Heffner, E.L. *et al.* (2009) Genomic selection for crop improvement. *Crop Sci.* 49, 1–12
31. Sallam, A.H. *et al.* (2015) Assessing genomic selection prediction accuracy in a dynamic barley breeding population. *Plant Genome* 8, eplantgenome2014.05.0020
32. Crossa, J. *et al.* (2021) The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. *Front. Plant Sci.* 12, 651480
33. 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 (Bethesda)* 8, 3961–3972
34. Robert, P. *et al.* (2022) Phenomic selection: a new and efficient alternative to genomic selection. *Methods Mol. Biol.* 2467, 397–420
35. Tang, H. and Thomas, P.D. (2016) Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics* 203, 635–647
36. Barshai, M. *et al.* (2020) Identifying regulatory elements via deep learning. *Annu. Rev. Biomed. Data Sci.* 3, 315–338
37. Zhang, G. *et al.* (2020) C-RNNCrispr: prediction of CRISPR/Cas9 sgRNA activity using convolutional and recurrent neural networks. *Comput. Struct. Biotechnol. J.* 18, 344–354
38. Singh, R. *et al.* (2016) DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* 32, 639–648
39. Chung, P.Y. and Liao, C.T. (2020) Identification of superior parental lines for biparental crossing via genomic prediction. *PLoS One* 15, e0243159
40. Villar-Hernández, B.J. *et al.* (2018) A Bayesian decision theory approach for genomic selection. *G3 (Bethesda)* 8, 3019–3037
41. Mohammadi, M. *et al.* (2015) PopVar: a genome-wide procedure for predicting genetic variance and correlated response in bipa-rental breeding populations. *Crop Sci.* 55, 2068–2077
42. Jackson, R. *et al.* (2023) Phenomic and genomic prediction of yield on multiple locations in winter wheat. *Front. Genet.* 14, 1164935
43. Montesinos-López, A. *et al.* (2017) Genomic Bayesian functional regression models with interactions for predicting wheat grain yield using hyper-spectral image data. *Plant Methods* 13, 62
44. Montesinos-López, A. *et al.* (2018) Bayesian functional regression as an alternative statistical analysis of high-throughput phenotyping data of modern agriculture. *Plant Methods* 14, 46
45. Montesinos-López, O.A. *et al.* (2017) Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. *Plant Methods* 13, 4
46. Runcie, D.E. *et al.* (2021) MegaLMM: mega-scale linear mixed models for genomic predictions with thousands of traits. *Genome Biol.* 22, 213
47. Krause, M.R. *et al.* (2020) Aerial high-throughput phenotyping enables indirect selection for grain yield at the early generation, seed-limited stages in breeding programs. *Crop Sci.* 60, 3096–3114
48. Montesinos-López, O.A. *et al.* (2018) Prediction of multiple-trait and multiple-environment genomic data using recommender systems. *G3 (Bethesda)* 8, 131–147
49. Kismiantini *et al.* (2021) Prediction of count phenotypes using high-resolution images and genomic data. *G3 (Bethesda)* 11, jkab035
50. Montesinos-López, O.A. *et al.* (2023) Genomics combined with UAS data enhances prediction of grain yield in winter wheat. *Front. Genet.* 14, 1124218
51. Costa-Neto, G. and Fritsche-Neto, R. (2021) Enviromics: bridging different sources of data, building one framework. *Crop Breed. Appl. Biotechnol.* 21, 393521–393533
52. Xu, Y. (2016) Envirotyping for deciphering environmental impacts on crop plants. *Theor. Appl. Genet.* 129, 653–673
53. Heslot, N. *et al.* (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* 127, 463–480
54. Jarquín, D. *et al.* (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127, 595–607
55. Morais Júnior, O.P. *et al.* (2018) Single-step reaction norm models for genomic prediction in multienvironment recurrent selection trials. *Crop Sci.* 58, 592–607
56. Costa-Neto, G. *et al.* (2021) Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity* 126, 92–106

57. Costa-Neto, G. *et al.* (2023) Envirome-wide associations enhance multi-year genome-based prediction of historical wheat breeding data. *G3 (Bethesda)* 13, jkac313
58. Rogers, A.R. *et al.* (2021) The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. *G3 (Bethesda)* 11, jkaa050
59. Tolhurst, D.J. *et al.* (2022) Genomic selection using random regressions on known and latent environmental covariates. *Theor. Appl. Genet.* 135, 3393–3415
60. Costa-Neto, G. *et al.* (2021) EnvRtype: a software to interplay enviromics and quantitative genomics in agriculture. *G3 (Bethesda)* 11, jkab040
61. Fradgley, N.S. *et al.* (2023) Prediction of near-term climate change impacts on UK wheat quality and the potential for adaptation through plant breeding. *Glob. Chang. Biol.* 29, 1296–1313
62. Costa-Neto, G. *et al.* (2021) Enviromic assembly increases accuracy and reduces costs of the genomic prediction for yield plasticity in maize. *Front. Plant Sci.* 12, 717552
63. Cuevas, J. *et al.* (2016) Genomic prediction of genotype × environment interaction kernel regression models. *Plant Genome* 9, eplantgenome2016.03.0024
64. Cooper, M. and Messina, C.D. (2021) Can we harness 'enviromics' to accelerate crop improvement by integrating breeding and agronomy? *Front. Plant Sci.* 12, 735143
65. Technow, F. *et al.* (2015) Integrating crop growth models with whole genome prediction through approximate Bayesian computation. *PLoS One* 10, e0130855
66. Cooper, M. *et al.* (2016) Use of crop growth models with whole-genome prediction: Application to a maize multienvironment trial. *Crop Sci.* 56, 2141–2156
67. Messina, C.D. *et al.* (2018) Leveraging biological insight and environmental variation to improve phenotypic 1 prediction: Integrating crop growth models (CGM) with whole genome prediction (WGP). *Eur. J. Agron.* 100, 151–162
68. Ly, D. *et al.* (2018) Whole-genome prediction of reaction norms to environmental stress in bread wheat (*Triticum aestivum* L.) by genomic random regression. *Field Crop Res.* 216, 32–41
69. Millet, E.J. *et al.* (2019) Genomic prediction of maize yield across European environmental conditions. *Nat. Genet.* 51, 952–956
70. Montesinos-López, O.A. *et al.* (2024) Enhancing winter wheat prediction with genomics, phenomics and environmental data. *BMC Genomics* 25, 544
71. Burgueño, J. *et al.* (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Sci.* 52, 707–719
72. Slyusar, V. (1999) A family of face products of matrices and its properties. *Cybern. Syst. Anal.* 35, 379–384
73. Martini, J.W.R. *et al.* (2020) On Hadamard and Kronecker products in covariance structures for genotype × environment interaction. *Plant Genome* 13, e20033
74. Montesinos-López, O.A. *et al.* (2019) A Bayesian genomic multi-output regressor stacking model for predicting multi-trait multi-environment plant breeding data. *G3 (Bethesda)* 9, 3381–3393
75. Enciso-Rodriguez, F. *et al.* (2018) Genomic selection for late blight and common scab resistance in tetraploid potato (*Solanum tuberosum*). *G3 (Bethesda)* 8, 2471–2481
76. Cuevas, J. *et al.* (2023) Modeling genotype × environment interaction for single and multitrait genomic prediction in potato (*Solanum tuberosum* L.). *G3 (Bethesda)* 13, jkac322
77. Aastveit, A.H. and Martens, H. (1986) ANOVA interactions interpreted by partial least squares regression. *Biometrics* 42, 829–844
78. Montesinos-López, O.A. *et al.* (2022) Partial least squares enhances genomic prediction of new environments. *Front. Genet.* 13, 920689
79. Montesinos-López, O.A. *et al.* (2019) An R package for Bayesian analysis of multi-environment and multi-trait multi-environment data for genome-based prediction. *G3 (Bethesda)* 9, 1355–1369
80. Montesinos-López, O.A. *et al.* (2022) Multi-trait genome prediction of new environments with partial least squares. *Front. Genet.* 13, 966775
81. Pérez-Rodríguez, P. and de Los Campos, G. (2022) Multitrait Bayesian shrinkage and variable selection models with the BGLR-R package. *Genetics* 222, iyac112
82. Ortiz, R. *et al.* (2023) Partial least squares enhance multi-trait genomic prediction of potato cultivars in new environments. *Sci. Rep.* 13, 9947
83. Montesinos-López, O.A. *et al.* (2023) Optimizing sparse testing for genomic prediction of plant breeding crops. *Genes* 14, 927
84. Bonnett, D. *et al.* (2022) Response to early generation genomic selection for yield in wheat. *Front. Plant Sci.* 12, 718611
85. Dreisigacker, S. *et al.* (2023) Results from rapid-cycle recurrent genomic selection in spring bread wheat. *G3 (Bethesda)* 13, jkad025