





## INVITED REVIEW

## Crop Breeding &amp; Genetics

# Prediction-based breeding: Modern tools to optimize and reshape programs

Roberto Fritsche-Neto<sup>1</sup>  | Rafael T. Resende<sup>2,3</sup>  | Tiago Olivoto<sup>4</sup>  |  
 Julian Garcia-Abadillo<sup>5</sup> | Moyses Nascimento<sup>5,6</sup> | Marco Antônio M. Bahia<sup>2</sup> |  
 Diego Jarquin<sup>5</sup> | Rafael Augusto Vieira<sup>7</sup> 

<sup>1</sup>Horticulture Sciences, College of Agriculture and Life Sciences, North Carolina State University, Raleigh, North Carolina, USA

<sup>2</sup>Plant Breeding Sector, School of Agronomy (EA), Federal University of Goiás (UFG), Samambaia Campus, Goiânia, Goiás, Brazil

<sup>3</sup>TheCROP, A Precision Breeding Project, FUNAPE, Samambaia Technological Park, Samambaia Campus—UFG, Goiânia, Goiás, Brazil

<sup>4</sup>Department of Crop Sciences, Federal University of Santa Catarina, Florianópolis, Santa Catarina, Brazil

<sup>5</sup>Agronomy Department, University of Florida, Gainesville, Florida, USA

<sup>6</sup>Laboratory of Intelligence Computational and Statistical Learning (LICAE), Department of Statistics, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

<sup>7</sup>Department of Research & Development, Bayer Crop Science–Breeding, Uberlândia, Minas Gerais, Brazil

## Correspondence

Roberto Fritsche-Neto, Horticulture Sciences, College of Agriculture and Life Sciences, North Carolina State University, Campus Box 7609, Raleigh, NC 27695-7609, USA.

Email: [roberto.neto@ncsu.edu](mailto:roberto.neto@ncsu.edu)

Assigned to Associate Editor Vivi Arief.

## Abstract

Prediction-based breeding reshapes plant genetic improvement by prioritizing the predictive ability of models over causal interpretation. This review examines recent advances in the use of tools such as genomic selection, high-throughput phenotyping, multi-omics integration, and enviromics to enhance genetic gain and improve the efficiency of breeding programs. Predictive models, while powerful, rely on validation within the genetic and environmental domains represented in the training set, with evident risks when extrapolated to unrelated scenarios. Traditional approaches such as marker-assisted selection and genome-wide association study remain limited for quantitative traits, reinforcing the need for prediction-oriented models. Moreover, the expansion of omics data sources, although capturing greater biological complexity, must be accompanied by rigorous validation practices to avoid fragile interpretations. Stochastic simulations are a strategic tool for testing selection schemes, optimizing training populations, anticipating overfitting risks, reducing costs, and guiding decisions based on prospective scenarios. This review also highlights the importance of ensuring independence between calibration and prediction, focusing on

**Abbreviations:** AI, artificial intelligence; CV, cross-validation; DL, deep learning;  $G \times E$ , genotype-by-environment; GBLUP, genomic best linear unbiased prediction; GIS, Geographic Information System; GS, genomic selection; GWAS, genome-wide association study; HTP, high-throughput phenotyping; LiDAR, light detection and ranging; MAS, marker-assisted selection; ML, machine learning; MTGS, multi-trait genomic selection; NDVI, normalized difference vegetation index; NIR, near-infrared; QTL, quantitative trait loci; RGB, red-green-blue; SEL, stacking ensemble learning; SNP, single nucleotide polymorphism; TPE, target population of environment; VI, vegetation index.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs License](https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *Crop Science* published by Wiley Periodicals LLC on behalf of *Crop Science Society of America*.

practical accuracy evaluation, and prioritizing operational utility over mechanistic explanation. In summary, prediction-based breeding is a core strategy for modernizing breeding programs, connecting computational tools, high-dimensional data, and pragmatic decision-making to deliver consistent results.

### Plain Language Summary

This review explains how modern plant breeding is changing. Instead of only trying to understand why plants grow a certain way, breeders now use tools that help predict which plants will do best in the future. These tools include DNA analysis (genomics), high-tech sensors to measure plant traits (phenotyping), and weather and soil data (enviromics). By using computer models and simulations, breeders can choose the best plants faster, cheaper, and with better results. This helps farmers grow crops that are more productive and better suited for different environments, even with climate change. The review also shows how combining many kinds of data can lead to smarter decisions in breeding programs.

## 1 | INTRODUCTION

Continued increases in genetic gain demonstrate the success of established public and private plant breeding programs (Crossa et al., 2021). This progress reflects advances in selection methodologies and a growing emphasis on anticipating future outcomes rather than merely explaining past ones. Much discussion has emerged around the contrast between explanatory and predictive models, notably framed by Breiman (2001) in his critique of conventional statistical modeling. In plant breeding, explanatory models—while useful for understanding biological mechanisms—are often limited in their generalizability, being at times closely linked to the specific structure of the datasets from which they are derived. When properly calibrated and validated through independent testing or cross-validation (CV), predictive models offer greater potential for supporting selection decisions in operational programs. Their value lies not in explaining why traits vary, but in reliably forecasting performance under diverse and future scenarios. This perspective is particularly important in the context of genomic selection (GS), where accuracy and applicability to untested genotypes are central objectives (Merrick et al., 2022; Resende, 2024).

At the same time, explanatory models remain essential for understanding the underlying mechanisms of biological systems. These models help provide structural and functional interpretations of genotype-by-environment ( $G \times E$ ) interactions, which are fundamental for guiding theoretical insights and applied decisions in breeding programs (Xu et al., 2022). However, relying exclusively on explanatory approaches can limit the integration of predictive frameworks,

especially in systems involving omics data or complex, multi-dimensional environments. For instance, transcriptomic or enviromic variables may improve model performance even without a clearly understood mechanistic basis (Guadagno et al., 2020; Resende et al., 2024).

These limitations pose a challenge in fields where scientific training prioritizes causal logic and biological interpretability, often at the expense of leveraging latent or abstract variables that offer high predictive value. Consequently, a balanced paradigm is needed—one that embraces both explanatory depth and predictive power—particularly as breeding increasingly depends on model-informed inference to manage uncertainty and support decisions across dynamic agro-environmental scenarios (G. Costa-Neto, Crossa, et al., 2021; Rincent et al., 2019).

Nevertheless, in the last two decades, a growing set of modern technologies has been developed and now awaits efficient integration into traditional breeding pipelines. This integration offers attractive benefits, yet comes with the challenges of modifying established and operational systems. Newly available technologies, genomics rapid cycling (Gaynor et al., 2017; Gorjanc et al., 2018), high-throughput phenotyping (HTP, phenomics) (Galli et al., 2020; Yassue, Galli, Borsato, et al., 2022), and historical descriptions of environmental relatedness (G. Costa-Neto, Galli, et al., 2021; Resende et al., 2021) are fundamental to improving conventional breeding schemes and increasing genetic gain. Furthermore, integrating these new technologies into routine breeding pipelines will support the delivery of cultivars with robust yields in the face of expected unfavorable future environmental conditions caused by climate change and the associated increased

occurrence of biotic and abiotic stresses (G. Costa-Neto & Fritsche-Neto, 2021; G. Costa-Neto, Fritsche-Neto, et al., 2021; Xu et al., 2022).

Therefore, to meet the well-documented challenges of crop production, there is a pressing need to use new technologies to accelerate the progress of plant breeding. These tools can be incorporated into conventional phenotypic breeding programs or help to gradually redesign established phenotypic breeding pipelines to shift toward an even more data-driven perspective. The benefits of phenomics and enviromics together in benchmark genomic pipelines offer the potential to deliver larger increases in the accuracy and efficiency of breeding pipelines by allowing for the selection of better-adapted genotypes in a shorter, more cost-effective manner (Crossa et al., 2021). Moreover, the layers of genomics, phenomics, and enviromics data are interconnected, and their use can be optimized based on available resources and program structure (G. Costa-Neto & Fritsche-Neto, 2021). Together, they can offer a pathway for conventional phenotypic breeding to envision diverse opportunities to accelerate genetic gains.

Finally, stochastic simulation is a powerful tool for optimizing breeding programs because it is a fast and inexpensive method for testing alternative breeding frameworks (DoVale et al., 2022). Hence, simulations have been used to improve plant breeding and address theoretical concepts in quantitative genetics and breeding, such as incorporating modern tools or optimizing breeding methods (Gaynor et al., 2021); however, they have rarely been used to improve the long-term performance of breeding programs (Sabadin et al., 2022). Therefore, there is a great opportunity to test several possible breeding scenarios, factors, tools, and ways to model data, then invest the resources in the most promising strategies.

This review aims to present how stochastic simulations and modern tools such as GS, enviromics, phenomics, and omics are key to optimizing breeding programs, relying on predictions to support decisions, reducing costs, increasing accuracy, and speeding up the process. Furthermore, we aim to present possibilities for integrating its breeding frameworks and perspectives into the future and how each tool can affect the breeder's equation components and, consequently, the response to selection.

## 2 | GS: THE NEW “SKELETON” OF BREEDING PROGRAMS

GS has taken root in breeding programs by providing a precise and efficient framework for genetic improvement (Crossa et al., 2017; Merrick et al., 2022). By leveraging genome-wide single nucleotide polymorphism (SNP) data, breeders can predict the genetic potential of individuals early in life, accelerating selection cycles, enabling work with a broader set of evaluated individuals, and enhancing genetic gains (Resende,

### Core Ideas

- Prediction-based tools are more efficient in early breeding stages, where the heritability is usually low.
- Explanatory models seek understanding; predictive models drive operational decisions, making them indispensable for effective plant breeding.
- Predictive models in breeding will only deliver reliable outputs within the genetic and environmental domains represented in the training data.
- For complex traits, the priority must be predictive reliability, not causal interpretation, especially when using high-dimensional omics data.
- Stochastic simulations optimize prediction pipelines by evaluating model robustness across architectures, heritabilities, and  $G \times E$  interaction, among other scenarios, before field application.

2024). This data-driven approach reduces reliance on traditional phenotypic selection, allowing for more accurate and cost-effective breeding decisions. As a result, GS has become a key component in breeding programs where its implementation is feasible, guiding selection strategies and supporting genetic progress across different crops.

In GS, training and testing sets should be genetically related to ensure that the training set captures the genetic diversity of the testing set (Fritsche-Neto et al., 2021; Simiqueli et al., 2023). Therefore, breeding programs must start forming foundational core parents (training populations) that represent the genetic diversity in the current progeny and conform to the testing population(s) to the greatest extent possible. These foundational parents should be extensively phenotyped in different target environments and genotyped with high-density marker systems. The resulting training sets of foundation parents will be able to produce a model with high accuracy for the current highly selected progenies (Crossa et al., 2017).

Beyond optimizing the training set composition, evaluating the predictive performance and reliability of GS models is fundamental for their effective application in breeding. Various complementary methods and strategies are available to assess different aspects of model quality. CV schemes such as  $k$ -fold, leave-one-out cross-validation, stratified sampling, and multi-environment strategy (e.g., CV0–CV2 from Jarquín et al., 2014) help quantify model generalization across genotypes and environments. However, accuracy alone is insufficient—uncertainty must also be addressed. Tools such as Bayesian credible intervals, posterior variance estimates, and bootstrapping allow breeders to quantify the confidence of predictions and understand the potential variability of

selection outcomes. Predictive ability metrics (e.g., Pearson's  $r$ ) remain essential for comparing models, while relative efficiency offers insight into the potential gain from GS relative to traditional phenotypic selection. These methods are not mutually exclusive and can be combined strategically to improve decision-making. A summary of key evaluation and uncertainty estimation approaches in GS is provided in Table 1.

The most important limitation to determining accurate phenotypes has been the time and cost required to measure traits in the field (Fritsche-Neto & Borém, 2015). Field phenomics aims to study all plant phenotypes under various environmental conditions. Modern phenomics methods can use hyperspectral/multispectral cameras to provide hundreds of reflectance data points at discrete narrow bands in many environments and stages of crop development (Yassue, Galli, Fritsche-Neto, et al., 2022). Phenotyping technology can now quickly and accurately obtain data on agronomic traits based on advancements in plant phenotyping technologies (Cossa et al., 2021). Therefore, the main goal of an HTP is to reduce the cost of data per plot and to increase the prediction accuracy early in the crop-growing season with the use of highly heritable secondary phenotypes, closely related to the selection phenotypes (Fritsche-Neto & Borém, 2015).

Obtaining large and interoperable phenomics datasets from field phenotyping is essential for effective model development. These datasets should be used to characterize the foundational core parents in different environments and to integrate visual data collected across conditions. The resulting information, combined with pedigree and genomic data, can be used to compute genotypic values of the training set materials (Cossa et al., 2021). In this context, HTP traits can replace the traditional phenotype, such as plant height or scores for disease resistance, used as secondary traits in multi-trait prediction (normalized difference vegetation index [NDVI]) (Rutkoski et al., 2016), or serve as latent traits to be explored. These HTP traits may also support genomic prediction models based on repeated measurements over time (Morales et al., 2024).

Phenotypic variation across environments arises from genetic and environmental effects and their interaction. In multi-environment contexts, modeling these components accurately is challenging. Enviromics has become a valuable approach for improving genomics-assisted prediction tools by enabling detailed characterization of environmental variation and supporting the modeling of  $G \times E$  interactions (G. Costa-Neto & Fritsche-Neto, 2021; Resende et al., 2021). Recent approaches have combined field trial data with genomic information using enviromic sources. Non-linear kernels improved predictive accuracy, especially for untested environments (G. Costa-Neto, Fritsche-Neto, et al., 2021). This strategy also allowed modeling genotype-specific reaction norms for key environmental factors, helping to

explain  $G \times E$  interaction and improve prediction over models without enviromic data. While many models still use a limited set of covariates, ongoing research emphasizes the need for large-scale environmental characterization to support prediction and genotype recommendation across multiple environments using remote sensing, envirotyping platforms, and integrative AI (artificial intelligence) frameworks. This approach has been conceptually and methodologically discussed and proposed in detail (Resende et al., 2024) and applied with real data in a country-wide study (Bahia et al., 2025). Tools such as EnvRtype (G. Costa-Neto, Galli, et al., 2021) and SoilType (Fritsche-Neto, 2024) help enable broader environmental assembly. The environment-enabled genomic prediction framework supports  $G \times E$  modeling and optimizes experimental networks (G. Costa-Neto, Galli, et al., 2021; Rincent et al., 2017). Beyond enviromics, a broader integrative view emerges through multi-omics approaches.

In plant breeding, multi-omics refers to the integration of different “-omics” layers—such as genomics, transcriptomics, proteomics, metabolomics, phenomics, and enviromics—to understand complex traits better and enhance crop improvement (Mahmood et al., 2022). This integration enables a more complete view of biological systems and the molecular basis of phenotypic expression. By capturing interactions across the molecular hierarchy, multi-omics contributes to trait dissection and phenotype prediction, especially when non-linear relationships are involved. Within GS-based pipelines, multi-omics can accelerate genetic gain, provided data collection, storage, and integration are efficiently managed. However, integrating genomics, phenomics, and enviromics data, the core of the modern plant breeding triangle (Cossa et al., 2021), poses substantial statistical challenges due to high dimensionality, multicollinearity, and heterogeneous data structures. Addressing this complexity requires statistical models tailored to large-scale multi-omics datasets, often implemented within smart breeding frameworks using machine learning (ML) algorithms to support data integration and improve prediction accuracy for complex traits (Xu et al., 2022).

Empirical studies have shown that GS models can achieve reliable predictive ability for traits such as grain yield (GY), flowering time, and plant height in annual species, and for wood volume in perennial crops, under multi-environment testing designs with data collected over multiple years (Cui et al., 2020; Labroo et al., 2021; Simiqueli et al., 2023), supporting their implementation in breeding programs even when based on static or genetically unbalanced datasets, although model performance is still commonly assessed only by predictive accuracy. Hence, long-term breeding schemes should be analyzed concerning different genetic parameters, such as genetic gain, response to selection, genetic variance, and prediction accuracy, to determine the best long-term strategy for maximizing results (Sabadin et al., 2022). For that, stochastic

**TABLE 1** Summary of key methods for assessing predictive performance and uncertainty estimation in genomic selection pipelines. The table outlines commonly used cross-validation strategies (including  $k$ -fold, leave-one-out cross-validation (LOOCV), and environment-based cross-validation [CV]), statistical measures of predictive ability (e.g., Pearson's correlation), and Bayesian-based uncertainty quantification techniques (e.g., credible intervals and posterior variances). Additionally, it includes metrics for evaluating model utility in breeding, such as relative efficiency compared to traditional selection. These methods support model validation, decision-making, and long-term strategy optimization in genomic-assisted breeding programs.

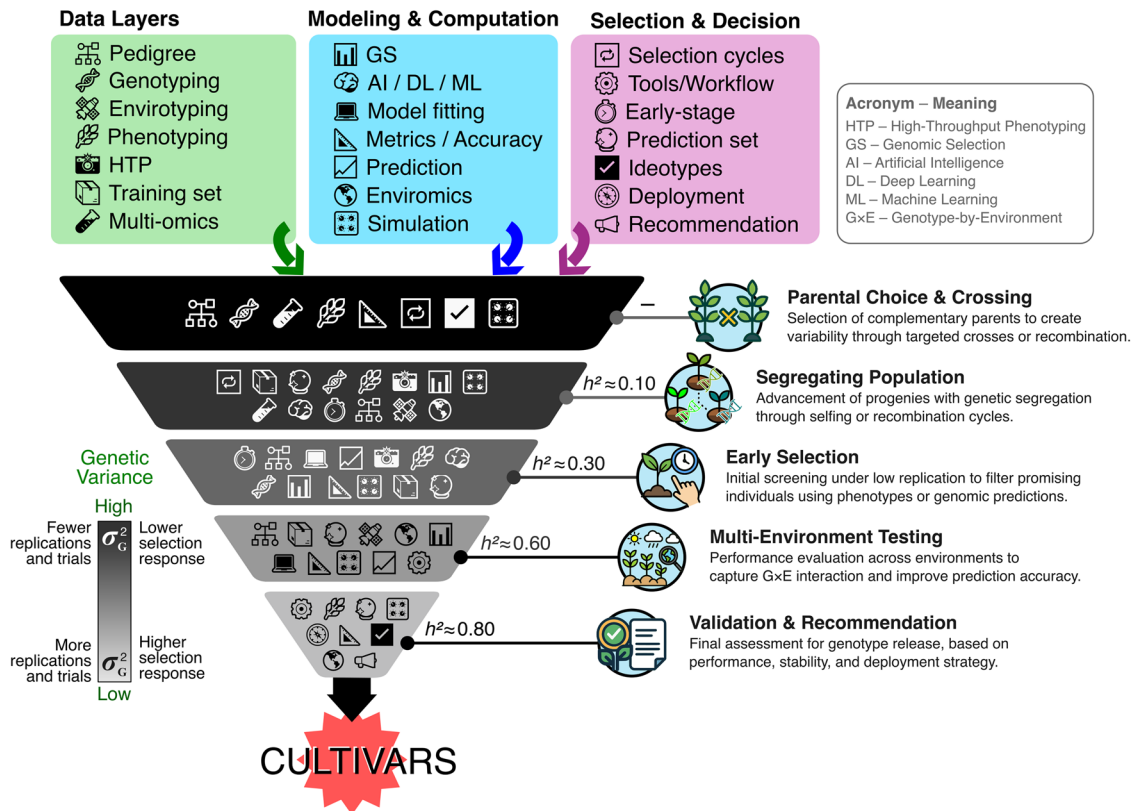
Method/strategy	What it measures/is used for	When to use
<b>Goodness of fit</b>	The suitability of a statistical model and its predictors for a given predictive task	Used in parametric models where parameters (e.g., fixed effects, variance components) are estimated
Heritability	The proportion of phenotypic variance explained by genetic factors	As a measure of response to selection and as the upper boundary for accuracy in genomic selection models
Lambda parameters	Regularization or penalization applied to random effects	To estimate the importance of a term in a mixed model
Variance components	The contribution of each term to the total variance in a mixed model	To quantify the proportion of phenotypic variance explained by a specific term
AIC/BIC	A balance of model fit and complexity	To choose the best model across a set of models with varying complexities
Confidence intervals	In frequentist statistics, a range quantifying the uncertainty around a parameter estimate	In likelihood-based frequentist models
Credible intervals	In Bayesian statistics, a range quantifying the uncertainty around a parameter estimate	In Bayesian models such as the Bayesian Alphabet
<b>Train-test split</b>	A general framework for obtaining predictions on observed values while preventing data leakage	To evaluate predictive models in supervised machine learning tasks
Cross-validation (CV)	A train-test split technique where a predicted value is obtained for each observation in the dataset	When predictions for every observation are required
$k$ -fold CV	A CV scheme where data are split into $k$ folds, and each fold is iteratively used as the test set	Default CV strategy for moderate to large datasets
Leave-one-out CV (LOOCV)	A specific $k$ -fold CV scheme where $k$ equals the number of observations, predicting one observation at a time	To maximize the training sample size in small datasets
Stratified CV	A specific $k$ -fold CV scheme where each fold is composed of observations belonging to the same group (e.g., family)	To evaluate the generalizability of models across different groups
MET CVs	A set of CV schemes developed for MET trials (genotype-in-environment observations)	To evaluate model performance across different predictive tasks
CV2	Sparse testing/unconstrained CV	Predict genotypes observed in other environments
CV1	Genotype-constrained CV	Predict genotypes that were never observed before in observed environments
CV0	Environment-constrained CV	Predict genotypes that were observed before but in completely new environments
CV00	Double-constrained CV	Predict genotypes that were never observed in completely new environments
Forward prediction	Time-constrained train-test split	Use training sets composed of observations that were collected before the ones that are predicted
<b>Performance metrics</b>	Functions that map observed and predicted values into scores	To compare the predictive performance of different models
bootstrapping	Measures variability in accuracy metrics across resamples	When dataset is small or needs uncertainty in metrics
Pearson's correlation	Parametric correlation between observed and predicted values	When interested in the linear relationship between values
Spearman's correlation	Nonparametric correlation between observed and predicted values	When interested in the relationship between rankings rather than values
RMSE	Squared deviations between observed and predicted values	When deviations from the actual values are more important than relationships
MAE	Absolute deviations between observed and predicted values	When all deviations, regardless of size, are weighted equally

(Continues)

TABLE 1 (Continued)

Method/strategy	What it measures/is used for	When to use
Relative efficiency	Gain from GS relative to traditional selection	When comparing predictive models vs. phenotypic selection
Conditional selection	Percentage of selected genotypes (positives) that are actually good (true positives)	When assessing if selection based on predictions is effective from a breeding operational standpoint

Abbreviation: AIC, Akaike information criterion; BIC, Bayesian information criterion; GS, genomic selection; MAE, mean absolute error; MET, multi-environment trials; RMSE, root mean square error.



**FIGURE 1** Funnel representation of the cultivar development pipeline and its main components (black and white flat icons inside each box). The five breeding stages are shown from top to bottom, illustrating the reduction in genetic variance ( $\sigma_G^2$ ) and the increase in heritability ( $h^2$ ) across the selection process. Each stage is linked to specific data layers (green), modeling tools (blue), and decision-making components (purple), all represented by flat icons. Icons are placed according to their most frequent or practical use, although most components could operate across multiple stages. The scale on the left highlights the decline in genetic variability and the increase in trial intensity. The pipeline converges in cultivar deployment, guided by data-driven decisions.

simulations are a suitable tool to assess the effect of a new tool or strategy in a long-term breeding program. Furthermore, several factors can be controlled in these studies to make inferences on given genetic parameters over several breeding cycles quickly, inexpensively, and consistently (DoVale et al., 2022). Consequently, the efforts and resources can be invested in the most promising scenarios.

After defining which elements will be used, guided by empirical analysis or stochastic simulations, the next step involves integrating diverse data layers and computational

approaches into a unified prediction framework. As illustrated in Figure 1, modern plant breeding can incorporate multiple components across distinct stages, from parental selection to cultivar deployment. These components include biological data layers, modeling tools, and decision-making elements, each contributing to predictions, evaluations, or selection criteria. While all components could be applied across stages, they are typically allocated where their use is most frequent or practical. This integration may follow a stepwise strategy—treating each stage as a framework—or adopt a more complex,

simultaneous model using AI or mixed models to process all inputs jointly (e.g., in multi-omics prediction). Regardless of strategy, aligning the right tools with the appropriate stage is critical for enhancing prediction accuracy and optimizing breeding decisions across the cultivar development pipeline.

It is important to highlight that modern tools are particularly helpful in the early stages of a breeding program (Figure 1) because they enable faster, more precise, and data-driven decision-making in those generations with low heritability (accuracy of selection). Advances in genomics, marker-assisted selection (MAS), and HTP allow breeders to identify desirable traits at the molecular level, reducing the need for time-consuming field trials or, at least, removing the garbage in a way that allocates our efforts and resources that are potentially superior. On the other hand, stochastic simulations are important before we start the crosses or selection. They may help us determine the best breeding framework and the stage to deploy modern tools or even define the crossing block and field trials, optimizing resource allocation and maximizing genetic gains per amount invested. Hence, in the next sections, we will detail how each tool can be integrated into breeding programs, focusing on genomic prediction, and discuss the associated challenges and perspectives.

### 3 | ENVIROMIC PREDICTION: MAPPING ENVIRONMENTAL DIVERSITY TO UNLOCK $G \times E$ INTERACTIONS

Genomics predicts genotypic performance (breeding values) using genetic and molecular data. However, when addressing  $G \times E$  interactions, the aim shifts to forecasting how genotypes respond to diverse environmental conditions. This enables more effective recommendation and selection strategies to optimize performance across various scenarios (Voss-Fels et al., 2018). Predicting genotypic performance in untested environments requires integrating relevant data to avoid “blind” predictions. This approach reduces reliance on recommendations for genotypes with good regional performance or broadly stable materials. Predictions become more informative when based on data from environments that are similar to those already evaluated, allowing genotype behavior to be inferred with greater confidence. Although not strictly required, environmental data help reveal the drivers behind  $G \times E$  interaction and can enhance predictive accuracy. Tools such as Geographic Information Systems (GISs) and large-scale environmental datasets support this process (Resende et al., 2024; Xu et al., 2022).

As far as records indicate, the use of GIS in plant breeding was first explored by Annicchiarico (2002) and Martins (2004). Later studies, such as those by Marcatti et al. (2017) and G. M. F. Costa-Neto et al. (2020), highlighted the potential of GIS in optimizing genotypic recommendations. On

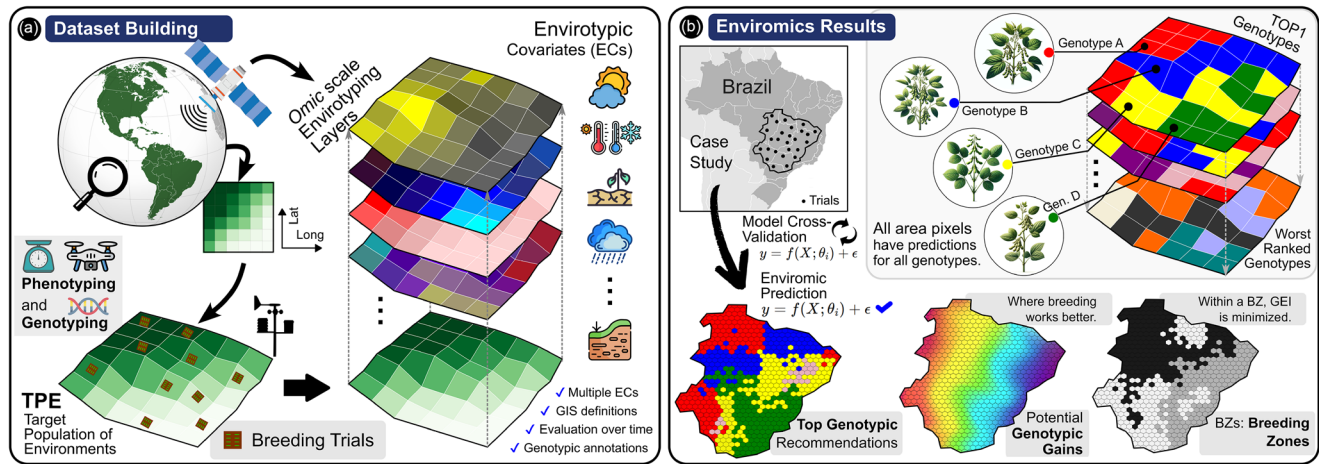
the other hand, the analysis of environmental covariates in plant breeding has a long history. Foundational studies by Kang et al. (1989), Piepho (1998), and Crossa et al. (1999) leveraged environmental data to enhance  $G \times E$  interpretation and predict genotypic performance across scenarios. These early efforts demonstrated the value of incorporating environmental information to refine plant breeding strategies and set the stage for more advanced approaches combining GIS and environmental covariates.

More recently, envirotyping and enviromics have emerged as central tools in plant breeding during the era of big data. Envirotyping, which collects environmental covariates across multiple scales, was introduced into plant breeding by Cooper et al. (2014) and Xu (2016). The term “enviromics,” an omics approach that uses the full spectrum of environmental variations to understand and predict  $G \times E$  interaction, was first proposed in a 2019 bioRxiv preprint later published by Resende et al. (2021) and further expanded by G. Costa-Neto, Crossa et al. (2021). Currently, models such as Kernels (G. Costa-Neto, Fritsche-Neto, et al., 2021; Jarquín et al., 2014), random regressions (Resende et al., 2021; Schaeffer, 2004), and factor analyses (Araújo et al., 2024) are being applied to enviromics studies. These advances have captured the attention of research groups, with applications reported by Cooper and Messina (2021), G. Costa-Neto et al. (2023), Crossa et al. (2021), Resende et al. (2021), and Resende et al. (2025).

#### 3.1 | Preliminary steps: GIS analyses and phenotype correction

##### 3.1.1 | GIS steps to prepare data for enviromics analyses

The first step in enviromics analyses using GIS is defining the target population of environments (TPEs), which involves delineating the geographic area of interest based on the study’s objectives and crop requirements (Cruz et al., 2025). This process requires selecting coordinate systems, projections, and metric units and determining spatial resolution according to data availability (Figure 2a). Phenotypic data, such as GY, are collected from field trials conducted across multiple locations and years, along with accompanying geographic metadata. Simultaneously, envirotypic covariates (ECs) are extracted from sources such as MODIS satellite imagery, WorldClim climate data, NASA POWER meteorological data, and SoilGrids soil maps (see Resende et al. [2024] for further details). These data are processed within a GIS environment, integrating and interpolating spatial data to generate continuous environmental layers essential for mapping patterns within the TPE and preparing data for subsequent analyses.



**FIGURE 2** Enviromic prediction for target populations of environments (TPEs). (a) Illustrates the dataset construction process, integrating phenotyping, genotyping, and environmental variables (ECs) derived from geoenvironmental layers. (b) Displays the outcomes, including validated genotypic predictions, performance heatmaps, and breeding zones (BZs) designed to minimize genotype-by-environment ( $G \times E$ ) interaction and optimize genotypic recommendations.

After data collection, standardization and quality control are conducted in the GIS environment (Resende et al., 2024). The goal is to optimize pixel size to enhance the precision of subsequent analyses. Data integration involves combining genotypic and envirotypic datasets based on their geographic coordinates. Additional data, such as genomic information, may be included to enrich the study. Finally, the integrated data are used for predictive analyses and modeling, facilitating the development of enviromic markers and  $G \times E \times M$  (genotype  $\times$  environment  $\times$  management) studies. Throughout the process, the extract, transform, load workflow ensures that data from diverse sources are extracted, transformed using GIS techniques, and loaded into analytical environments. This approach enables the identification of patterns and accurate predictions across environments, driving advancements in plant breeding.

### 3.1.2 | Phenotype correction to experimental design factors

For enviromics analyses, computationally demanding models are often required. Therefore, adopting a two-stage modeling approach is recommended (Resende et al., 2025; Verbyla, 2023). The first step involves correcting experimental effects or unwanted factors, such as blocks, replicates, local spatial effects, and, in some cases, years or growing seasons, so each pixel contains only one value per genotype. If there is interest in evaluating additional factors, such as management practices or seasons, retaining one value per genotype for each factor per pixel is possible. By the end of the process, the focus should be on assessing  $G \times E$  (or  $G \times E \times M$ ) interactions at a geographic scale, between pixels, not within them, as environ-

mental dissimilarity is expected to be lower within a pixel than between pixels. Unless a specific non-genetic factor is critical for evaluation within the pixel, the data should be adjusted to eliminate internal redundancies, prioritizing analysis at the desired scale.

A traditional mixed model can be employed to perform the initial correction of experimental effects and unwanted variation, with the R package lme4 (Bates et al., 2024) being a sensible choice due to its computational efficiency. All experimental effects, whether nested or hierarchical, should be adjusted. Subsequently, only the components relevant to the next step are added to the genotypic effects, forming a “new” phenotypic vector,  $\mathbf{y}^*$ , which will serve as the effective data for the enviromics model. This approach isolates essential effects, ensuring the data are suitable for further analyses without compromising environmental or geographic integrity. At this stage, genotypes should preferably be considered as fixed effects (Holland & Piepho, 2024; Verbyla, 2023). If treated as random effects, the final  $\mathbf{y}^*$  vector must undergo a deregression step, analogous to what is done in two-stage genomic prediction (Garrick et al., 2009). This ensures that the resulting data are prepared for robust enviromics modeling while maintaining consistency across geographic scales.

### 3.1.3 | Genotype/genomic prediction for different environments using enviromics

Predicting the performance of genetic materials in untested environments is one of the core challenges in plant breeding, given the  $G \times E$  interactions. Modern approaches combine genomic, phenotypic, and environmental information in

sophisticated predictive models. This strategy overcomes the limitations of traditional phenotypic selection, which often evaluates genotypes in subsets of environments, failing to capture the full range of environmental variability. Reaction norm models stand out in these predictions (Schaeffer, 2004), interpreting genotypic performance as a continuous response to environmental gradients. Jarquín et al. (2014) highlighted this approach by incorporating genomic and environmental data to predict crop yields. These models combine fixed effects, such as year-specific intercepts, with random effects that account for genetic variations and reaction norms to capture  $G \times E$  (or  $G \times E \times M$ ) interaction. Validation techniques, such as CV and leave-one-out, assess predictive accuracy, while kriging interpolates results to generate heatmaps for practical applications (Marcatti et al., 2017). These techniques help identify genotypes adapted to specific environments, enhancing breeding efficiency.

Enviromics effectively integrates statistical methods and ML to develop predictive models using multiple ECs converted into markers, predicting how genotypes interact with different environmental conditions (Bahia et al., 2025; Resende et al., 2024). The predictive model can be described by the following equation:

$$\mathbf{y}^* = f(\mathbf{X}, \boldsymbol{\theta}) + \varepsilon$$

where  $\mathbf{y}^*$  represents the vector of phenotypic values corrected for experimental design effects;  $\mathbf{X}$  is the matrix of ECs;  $\boldsymbol{\theta}$  is the vector of model parameters estimated for each genotype-environment combination; and  $\varepsilon$  is the residual error.

A practical application of this approach is presented in Figure 2a. A case study demonstrates how predictive models validated using CV techniques estimate genotypic performance in untested environments. Trevisan et al. (2025) illustrate the construction of enviromic modeling matrices using block-diagonal structures and the Kronecker product, facilitating the integration of genetic and environmental data in mixed models. The results include maps visualizing specific genotypic recommendations, potential genetic gains, and breeding zones that minimize  $G \times E$  interactions. Each pixel on the map represents specific predictions for all evaluated genotypes, highlighting superior genotypes for specific environmental niches, while heatmaps reveal the potential for genetic gains in distinct regions. This approach informs breeding strategies tailored to local conditions.

Studies like G. Costa-Neto, Fritsche-Neto et al. (2021) demonstrated that methods such as Gaussian and Deep Kernels can model dominance and non-additive effects, while Xavier et al. (2025) introduced MegaLMM, which integrates genomic and environmental data into a unified predictive model. Additionally, Araújo et al. (2024) applied factor analytic (FA) models combined with partial least squares for such predictions. These advancements can enable the exploration

of complex  $G \times E$  patterns, optimizing hybrid recommendations for diverse environmental conditions (Malosetti et al., 2013). Comparisons with reference models, such as Kernel-based approaches, shed light on the advantages of innovative methods like enviromics. The combination of ML with geographic data provides a significant edge in exploring nonlinear interactions, improving predictions for yield and adaptability (Cooper & Messina, 2021; G. Costa-Neto et al., 2023; Resende et al., 2025; Xu et al., 2022).

These strategies represent a significant leap forward in plant breeding, transforming static predictions into dynamic ones. By incorporating high-dimensional data, advanced modeling techniques, and tools like GIS, the agricultural field is increasingly equipped to identify genotypes adapted to specific environmental niches, even under climate change scenarios. The formalization of methods, such as those proposed by Resende et al. (2024) and illustrated in Figure 2, exemplifies how modern tools can maximize agricultural productivity and adaptation at regional scales, fundamentally reshaping how genetic recommendations are made.

### 3.2 | Perspectives and challenges in enviromics

The ongoing expansion of enviromics is closely tied to the increasing availability of high-resolution environmental data and the development of envirotyping covariates with predictive relevance. A notable trend is the design of structured enviromic datasets—analogue to genotyping chips—allowing broad, customizable coverage of environmental factors in breeding trials and commercial fields. More researchers are embracing these covariate-driven models not for explanatory power, but for their ability to improve genotype selection and recommendation in variable environments. New applications include multi-trait frameworks where traits represent environmental sources (e.g., trial types, control checks, or commercial blocks). Additionally, genotype recommendation metrics are evolving to account for practical constraints such as plant propagule availability and regional demand. There is also growing integration with multi-omics platforms, enabling joint modeling of environmental, genomic, transcriptomic, and metabolic influence (Angione et al., 2016; González et al., 2025). Notably, the convergence of enviromics and epigenomics is expected to become increasingly relevant, as environmental cues can trigger heritable epigenetic modifications with potential predictive value (Meyer, 2015). These perspectives indicate a shift from static, location-based analyses to scalable, prediction-oriented frameworks capable of supporting precision breeding pipelines.

Despite these advances, enviromics faces several conceptual and operational challenges. One critical limitation is the low temporal resolution of many datasets, which

hinders modeling dynamic  $G \times E$  interactions (Resende et al., 2024). Addressing this requires integrating time-series climate data and simulation outputs from crop models, potentially linked to digital twin systems. Another challenge involves balancing spatial coverage with the risk of overfitting in prediction models—prompting using spatial CV as a best practice (Rogers & Holland, 2022; Bahia et al., 2025). Additionally, merging enviromics with socio-environmental layers (e.g., management practices or market trends) remains underexplored but offers potential for deployment decisions. Standardization is also a bottleneck; current studies use diverse metrics, scales, and modeling approaches, limiting interoperability and comparison. Finally, while ML and deep learning (DL) approaches are increasingly explored in breeding applications, their performance under low-replication field designs and sparse testing networks remains uncertain, suggesting the importance of careful benchmarking. Together, these challenges call for methodological refinement and greater harmonization of data pipelines across breeding programs.

#### 4 | HTP: SCALING TRAIT OBSERVATION FOR PRECISION BREEDING






A common misconception in literature is the interchangeable use of phenomics and HTP. While phenomics refers to the integrative and systematic study of phenotypic expression on a large scale, HTP specifically denotes the suite of technologies used to generate such data efficiently and non-destructively (Eberius & Lima-Guerra, 2009). Phenomics gains predictive capacity by integrating these high-resolution phenotypic datasets with genomic and environmental layers, supporting the modeling of complex traits. As an example, HTP-based pipelines have enabled the prediction of disease in crops (Thapa et al., 2024), the prediction of crop yields across environments (Adunola et al., 2024), and the genotype classification for drought tolerance. In this context, HTP serves as the technical foundation for predictive breeding pipelines by minimizing subjectivity and capturing repeated measurements across time and space. Sensors such as red-green-blue (RGB) cameras, hyperspectral and thermal imagers, LiDAR (light detection and ranging), and fluorescence detectors enable precise characterization of morphological, physiological, and compositional traits in large populations. By supporting trait quantification at scale and enhancing data quality, HTP provides the operational structure needed to develop predictive breeding models that refine selection decisions.

The foundation of HTP is built on three key pillars: sensor technology, acquisition platforms, and resolution in both space and time. Figure 3 illustrates the trait detection capac-

ity of the most used sensors in plant phenotyping. RGB and hyperspectral cameras broadly apply to morphological, physiological, and biochemical traits (D'Odorico et al., 2020; Gu et al., 2024; L. Zhang et al., 2019). Thermal imaging is effective for detecting plant water status and stress responses, while fluorescence sensors provide data related to photosynthetic efficiency and chlorophyll content (Raj et al., 2021; M. Zhang et al., 2023). LiDAR excels at structural measurements such as plant height and canopy coverage (Adak et al., 2024; Zang et al., 2023). When deployed on platforms like drones, ground vehicles, or conveyor-based systems, these technologies enable the large-scale, multi-temporal capture of complex phenotypic traits. The resulting datasets can then be used directly for phenotypic selection or incorporated into genomic models to enhance genomic prediction accuracy.

Sensors are deployed across various platforms, including drones, ground-based robots, satellites, and conveyor-based systems, enabling data collection at scales that span from individual leaves to entire crop fields (D. Li et al., 2021). The resulting datasets are inherently high-dimensional, capturing complex phenotypic traits with unprecedented spatial and temporal resolution. This multidimensionality opens new opportunities for integration with other “omics” technologies by, for example, linking drone-derived traits with genomic markers in a genome-wide association study (GWAS) (Adak et al., 2024; Dijkhuizen et al., 2025) or using thermal imagery for plant stress detection (Pineda et al., 2020). However, realizing this potential relies on the development and dissemination of data processing pipelines capable of translating the vast amounts of raw data into actionable insights for breeders (Crossa et al., 2025). Such pipelines may be developed using a variety of approaches, including, but not limited to, classical methods such as nonlinear models for plant maturity prediction (Volpato et al., 2021), ML techniques (O. Pérez et al., 2024), DL frameworks (Volpato et al., 2024), and models that incorporate spatiotemporal trends (Pérez-Valencia et al., 2022).

By utilizing multi-spectral sensors equipped with RGB and near-infrared (NIR) bands, researchers can compute a wide array of vegetation indices (VIs)—spectral reflectance metrics that quantify vegetation presence and condition based on photosynthetic responses to incident light, significantly enhancing the number of traits assessed in experimental plots. Indices such as the NDVI and tools like thermal imaging and canopy coverage analysis have proven effective in quantifying spatial variation and predicting critical traits for plant breeding programs. For example, these methods have been used to estimate GY in maize (*Zea mays*) (Gano et al., 2024) and soybean (*Glycine max*) (de Andrade Júnior et al., 2022), panicle number and shape in rice (*Oryza sativa* L.) (Lu et al., 2024), stem rust disease in wheat fields (Abdulridha et al., 2023), as well as to evaluate drought adaptation

Sensor	Traits	Morphological structural traits			Physiological functional traits			Component traits			
		Plant Height	Biomass	Canopy Coverage	Stay-green	Plant Maturity	Disease/Pests	Chlorophyll	Nitrogen	Water Content	Yield
RGB		✓	✓	✓	✓	✓	✓	✓			✓
Hyperspectral		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Thermal							✓			✓	
Chlorophyll Fluorescence					✓	✓	✓	✓	✓	✓	✓
LiDAR		✓	✓	✓							

**FIGURE 3** Overview of sensor capabilities in high-throughput plant phenotyping. Comparative matrix of five major sensor types—red-green-blue (RGB), multispectral/hyperspectral, thermal, chlorophyll fluorescence, and LiDAR—across morphological, physiological, and component traits. RGB and hyperspectral sensors cover a broad spectrum of traits; thermal imaging is suited for water status and stress; LiDAR for structural features; and chlorophyll fluorescence for photosynthetic activity.

in wheat (*Triticum aestivum*, winter wheat) (Reddy et al., 2024).

Plant growth and development are characterized by several phenotypic changes, which can only be studied by monitoring repeated phenotypes over time. Time-series data of VI allow researchers to track key time-dependent traits such as growth rate, flowering, and senescence dates. For example, a critical phenological trait, maturity date, is frequently predicted using RGB-based VIs, which have shown high correlations with ground-truth maturity dates (Volpato et al., 2024). Accurate predictions have been achieved using both smoothing methods, such as LOESS regression (Volpato et al., 2021), and more advanced approaches, including DL algorithms (Moeinizade et al., 2022) and ML techniques (O. Pérez et al., 2024). By capturing temporal dynamics, breeders can identify genotypes that exhibit resilience or superior performance under fluctuating conditions, providing valuable insights into trait stability and adaptability.

Integrating spatial and temporal data from HTP can offer a comprehensive understanding of how plants interact with their environments across space and time. In wheat disease monitoring, early detection using DL algorithms performed at 60% detection accuracy as early as 40 days after sowing, increasing to 71% and 77% at the later booting and flowering stages, respectively (Nguyen et al., 2023). Furthermore, spatiotemporal data can significantly improve the accuracy of selection models by incorporating dynamic environmental covariates and capturing phenotypic changes over time. Advanced statistical tools, such as random regression, Bayesian models, and recurrent neural networks, can lever-

age these datasets to model longitudinal traits more precisely (Moreira et al., 2020).

#### 4.1 | HTP enhances genomic prediction

GS accelerates early predictions of key traits, such as GY and tolerance to biotic and abiotic stresses, by training models on genome-wide markers and phenotypic data (Crossa, 1990), enabling prediction of genetic values for genotyped but untested lines (Crossa et al., 2025). While advancements like next-generation sequencing technologies (Behjati & Tarpey, 2013) have made genotypic data cheaper and more accessible, the accuracy of GS models heavily depends on the quality of phenotypic data (Crain et al., 2018; Resende, 2024). Due primarily to time and cost constraints, traditional phenotyping methods often struggle to collect large-scale, high-quality datasets. For example, Crain et al. (2018) reported manually measuring plant height and days to heading across all trial plots but collecting days to maturity data only in the first replicate. These inconsistencies can introduce biases, reducing the overall effectiveness of GS.

Historically, GS models have primarily focused on single-trait predictions, prioritizing key traits such as GY or disease resistance (Merrick et al., 2021). Advancements in computational power and ML algorithms have enabled the development of multi-trait genomic selection (MTGS) models, significantly improving prediction accuracy by leveraging correlations among traits. Within this context, HTP has become a cornerstone of GS, equipping breeders with tools

to evaluate large populations of plants or individuals with remarkable efficiency and enabling the collection of high-resolution, large-scale phenotypic data across multiple traits. This capability is valuable in early generations, such as F2 populations, where thousands of individual plants can be screened for desirable traits, such as yield and disease resistance.

By integrating both directly measured traits (e.g., plant height) and indirectly predicted traits (e.g., stress tolerance estimated through thermal imaging), HTP has enhanced the precision in several GS studies (Adak et al., 2024; Crossa et al., 2021; Kaushal et al., 2024). In wheat, the prediction accuracy of GS increased from 0.23 using a single-trait model to 0.37 using an HTP-based multi-trait model (Crain et al., 2018). While the overall accuracy remained modest, this nearly 50% improvement underscores the transformative potential of HTP in refining prediction breeding. In these studies, including HTP-based traits has been shown to significantly enhance prediction accuracy in GS, mainly because it leverages correlations between these traits and the target traits, such as GY, which can be effectively identified using ML algorithms.

Another important aspect of using HTP is that most traits reported using scale notes, such as disease scores or plot uniformity, can be transformed into quantitative traits, easily integrating them into a statistical model. In wheat, for example, digital surface models derived from drone imagery were successfully used to assess the underlying genetic architecture of stalk lodging. When combined with genome prediction models, they identified a key genomic region on chromosome 2A associated with this trait (Singh et al., 2019).

HTP also offers an efficient strategy to capture linkage disequilibrium (LD) phenotypically, without identifying specific genes associated with the target trait. Variables derived from HTP, such as NIR spectra or hyperspectral images, reflect expressed phenotypic characteristics and are thus associated with transcribed or structural polymorphisms in the DNA. Analogous to the use of microsatellites (but of course, in a modern way)—non-coding repetitive regions that can capture LD indirectly—HTP-derived variables do not need to be causally linked to the target trait to provide predictive ability. In an omic context, these variables function as phenomic markers, which, although they do not mechanistically explain trait variation, can offer high predictive power due to their structural correlation with the genomic background (Resende, 2024). Robert et al. (2022), in their protocol and review work, emphasize that variables obtained through NIR spectroscopy, even without direct phenotypic relation to the target trait, can capture sufficient genetic similarities to enable robust prediction, establishing high-density phenotyping as a functionally analogous approach to genomic prediction. In this framework, omic-based phenotyping integrates into the set of predictive

strategies, expanding the possibilities for selection in plant breeding programs.

## 4.2 | Perspectives and challenges in HTP

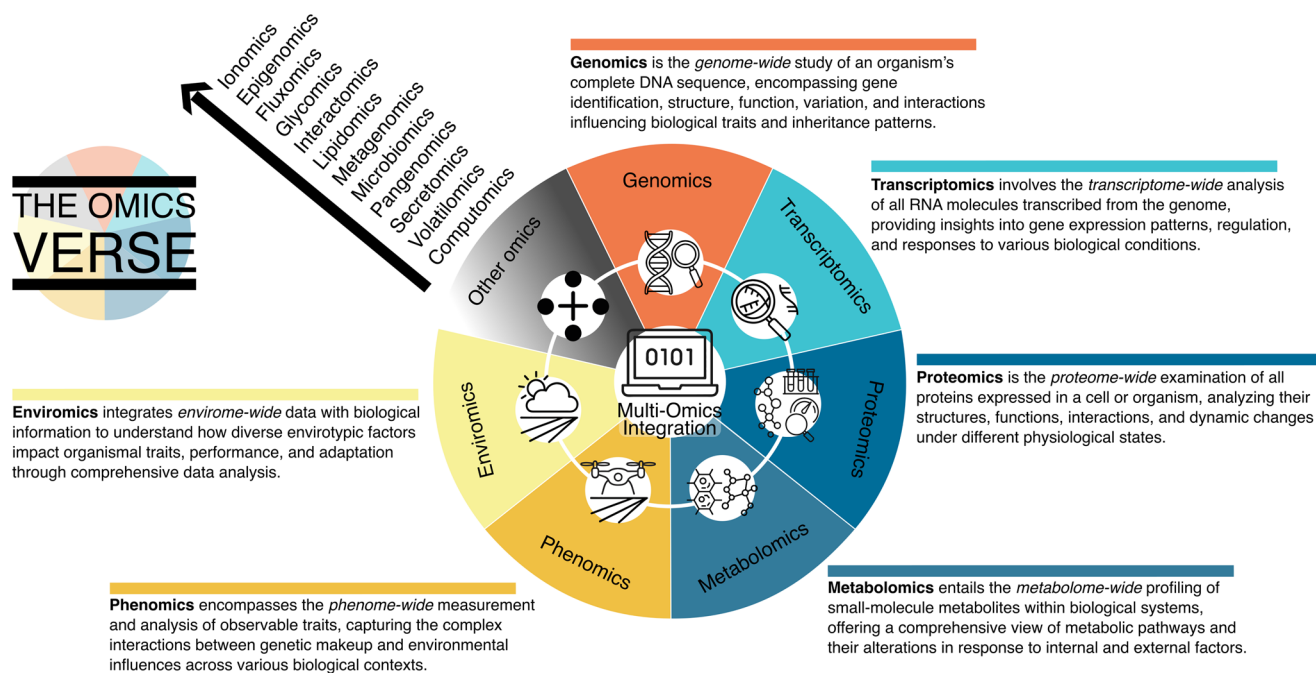
As cropping environments become increasingly unpredictable under climate change, characterizing variability across spatial and temporal scales is critical. HTP addresses this need by non-destructively capturing large-scale, multi-trait data that can be aligned with genome-wide information to strengthen predictive breeding pipelines, accelerating the development of resilient, high-performing crop varieties tailored to evolving environmental conditions.

The future of HTP is intertwined with integrating genomics, phenomics, and enviromics, creating a holistic and innovative approach to predicting genotypic performance. While advances in sensors, platforms, and analytical capabilities have enabled in situ measurements, efficiently integrating these data remains challenging. Enhancing communication among researchers and promoting interdisciplinary training, as encouraged by initiatives like Phenome Force (<https://phenome-force.github.io/PhenomeForce/>), is vital to advancing global phenotyping research. Open-source software like FieldImageR (Matias et al., 2020) and pliman (Olivoto, 2022), along with new platforms such as plimanshiny (<https://github.com/NEPEM-UFSC/plimanshiny>), offers promising avenues.

However, practical issues—such as the need to observe genotypes in the field to collect data for training DL and ML algorithms, low correlations between Vis and target trait (e.g., plant maturity in satay-green genotypes), and the high costs of commercial platforms limiting access in developing countries—continue to hinder widespread adoption. A potential solution lies in using marker data to predict phenotypic traits derived from HTP platforms and incorporating these as covariates in predictive models, unlocking HTP's full potential for GS. Overcoming these challenges will require collaboration, innovation, and the detailed expertise of researchers from diverse fields, fostering a fertile ground for sustainable advancements in plant breeding to meet future needs.

## 5 | MULTI-OMICS PREDICTION: LAYERING BIOLOGICAL DATA TO ENHANCE TRAIT FORECASTING

An expanded view of omics data considers genomic layers and a wide range of biological factors that influence phenotypic outcomes in field experiments. In this context, an omic dataset can be understood as a structured, tabular representation of variables such as phenomics, enviromics, and management practices—alongside genetic information—each contributing



**FIGURE 4** The omics verse: This diagram shows the different layers of information from relevant sources that contribute to trait prediction and biological understanding. Each -omics provides complementary context for the final prediction, which can be genetic (genomics, pedigree), environmental (enviromics), or a combination of the two (transcriptomics, proteomics, metabolomics, and epigenomics).

to trait variation. These layers capture distinct but interrelated aspects of biological systems, enabling more comprehensive modeling of complex traits. As revisited in Figure 4, the concept of multi-omics integration emphasizes the synergy among these datasets, underscoring their potential to improve phenotypic prediction and deepen our understanding of trait expression across environmental and molecular contexts.

The term “omics verse” continues to expand, incorporating new layers of information that deepen our understanding of plant function and trait expression. Once centered on genomics, this multiverse now includes a broad range of “-omics” disciplines, each revealing distinct facets of the genotype–phenotype continuum. This complexity becomes valuable for plant breeders when it leads to more accurate predictions, better selection strategies, or lower evaluation costs. Among the recent additions is enviromics (Resende et al., 2021), the youngest sibling in this network, introducing environmental variability into the predictive framework. Genomics is at the core of the omics verse—the study of genes and genomes in terms of structure, function, evolution, and editing. Structural genomics links markers to traits, while functional genomics clarifies gene roles. Epigenomics addresses heritable changes beyond DNA sequence, mutagenomics examines variation from mutations, and pangenomics explores the full gene repertoire of a species (Yang et al., 2021)—each offering complementary layers to enhance breeding outcomes.

Transcriptomics studies the transcriptome, the complete set of RNA transcripts produced by an organism’s genes, revealing which genes are active and how their expression changes in response to stimuli like stress. This helps identify genes for stress tolerance and crop improvement. Similarly, proteomics analyzes the complete set of proteins, focusing on their structure, function, and expression. Proteomics provides insights into the molecular mechanisms underlying crop resilience by identifying proteins involved in stress responses and cellular pathways. Advanced techniques like RNA-seq and mass spectrometry enable detailed and high-throughput analyses in both fields, facilitating the development of more resilient and productive crop varieties.

Metabolomics examines the complete metabolites, revealing the biochemical phenotype and providing insights into plant responses to various stresses and nutritional qualities. Meanwhile, ionomics focuses on the ionome, which is defined as the total mineral and trace element composition, using techniques like Inductively Coupled Plasma Mass Spectrometry/Optical Emission Spectroscopy to elucidate genes and networks related to mineral uptake and stress tolerance. Through high-throughput imaging and proximal and remote sensing, phenomics characterizes phenotypes on a large scale, enabling precise gene mapping and accelerated crop breeding by analyzing traits influenced by  $G \times E \times M$  interactions. To complement this, enviromics studies the envirome, the totality of environmental factors affecting plant traits, providing critical context for phenotypic expression. Together, these omics,

alongside metabolomics and ionomics, offer a holistic view of plant functionality, from molecular interactions to observable traits, ultimately aiding in developing more resilient and productive crops.

Recent studies demonstrate the power of multi-omics integration for yield prediction in crops. C. Wu et al. (2024) achieved a yield prediction accuracy increase from 0.32 to 0.43 in maize by combining SNPs, image traits, and metabolites with Random Forest. Similarly, Knoch et al. (2021) found that integrating transcripts with SNPs improved hybrid performance prediction in oilseed rape, with genomic best linear unbiased prediction (GBLUP) models showing effectiveness. Furthermore, P. Y. Wu et al. (2022) observed enhanced genomic prediction accuracy for barley (*Hordeum vulgare*) yield by combining SNPs, sequence variants, transcriptome (gene expression, transcriptome expression, expression present/absent variation), and metabolome data. These studies collectively highlight the benefit of multi-omics approaches and ML for improved crop yield prediction.

The tabular structure of these omics datasets allows us to capture and relate the levels of these factors, enabling the prediction of phenotypic outcomes even when some aspects are not directly observed (González et al., 2025; Jarquín et al., 2014).

Specifically, integrating multiple omics layers into GS enhances prediction accuracy by providing a more comprehensive biological representation of an organism. Genomics alone reveals genetic potential, but complex interactions between genes, environment, and other biological factors explain actual traits. Therefore, incorporating transcriptomics, proteomics, metabolomics, and enviromics allows us to capture these dynamic interactions, offering a deeper understanding of complex traits and gene-environment interactions. This multi-layered approach allows for constructing more accurate predictive models, as it accounts for the dynamic nature of biological systems and reduces the “unknowns” present when relying solely on genomic data (Roychowdhury et al., 2023). Consequently, the increased resolution from multi-omics data leads to more reliable predictions of breeding values, ultimately accelerating genetic improvement.

## 5.1 | Integration of secondary trait data in genomic prediction

GS modeling leverages genomic data to predict the genetic merit of individuals. In general, genomic data are comprised of high-density SNP markers. Directly (covariance structure) or indirectly (marker effect), GS establishes relationships between genotypes and phenotypes in a reference population to accurately estimate the genetic merit of new individuals (VanRaden, 2008). GS is particularly efficient when considering traits with low heritability, expensive evaluation, or late

manifestation. Hence, GS accelerates genetic gain by enabling early identification of superior individuals, reducing costs (Bonnett et al., 2022; Meuwissen et al., 2001; Saavedra et al., 2023) and, more importantly, time for releasing improved cultivars.

MTGS models allow the integration of information from multiple traits, including omics information (e.g., phenomics), within a single modeling framework. Crain et al. (2018) integrates HTP data (NDVI, canopy temperature) into GS models, effectively increasing selection accuracy. Since superior genotypes often should exhibit optimal values for multiple traits (Bernardo, 2020), incorporating genetically correlated traits into the model allows for the exploration of pleiotropic effects or LD between genes, potentially improving model accuracy (Jia & Jannink, 2012).

MTGS becomes especially insightful when secondary traits are easier to obtain than the primary trait or traits of interest. By leveraging genetic correlations between traits, MTGS outperforms conventional single-trait genomic selection (STGS) when the traits have missing records, as shown in Guo et al. (2014). Additionally, MTGS can increase the accuracy of predicting target low-heritability or difficult-to-measure traits by leveraging information from strongly correlated traits with higher heritability (Jia & Jannink, 2012). However, significant differences can be obtained depending on how the information about the secondary traits is implemented to conduct plant forecasting. Under an MTGS framework, the genetic correlations between traits are modeled, allowing also the prediction of missing genotypes across traits. On the other hand, using secondary traits only as covariates via STGS does not allow for the prediction of traits of disappeared genotypes. Still, this approach can outperform the MTGS when predicting only the main trait of interest.

Calus and Veerkamp (2011) studied the impact of genetic architectures in the relative performance of MTGS over STGS, and Jia and Jannink (2012) found that when the traits were controlled by few major quantitative trait loci (QTL), MTGS approaches outperformed STGS. This advantage disappeared when the trait was assumed to be regulated under a polygenic, more quantitative architecture. Runcie and Cheng (2019) described the potential problems that could arise from using STGS benchmarks, such as CV, when applied to MTGS frameworks due to the non-genetic correlations acting as confounding factors. The main limitation of the MTGS framework is that in the absence of phenotypic records for all traits for a group of genotypes, the results are comparable to those obtained with the STGS.

Examples of software developed for MTGS predictions include the sommer package (Covarrubias-Pazarán, 2016), which employs a frequentist approach and restricted maximum likelihood optimization; the BGLR package (P. Pérez & De Los Campos, 2014), integrating Bayesian hierarchical models with Markov Chain Monte Carlo methods;

and MegaLMM (Runcie et al., 2021), which utilizes a FA approach and is well-suited for large datasets with hundreds or thousands of secondary traits. Recently, González et al. (2025) presented CHiDO (Characterization and Integration of Drive Omics), a no-code implementation to overcome the technical and financial barriers to adopting predictive analytics, considering the multi-omics integration and interactions between omics.

Conventional MTGS models can be limited by unbalanced data, where not all genotypes are evaluated in all environments, leading to incomplete correlation estimates. Furthermore, these models typically do not incorporate information from different omics (e.g., transcriptomics, proteomics, and metabolomics) beyond phenotypic observations, reducing the ability to account for  $G \times E$  interactions fully.

## 5.2 | Implicit integration of multi-omics: AI and ML approaches

Conventional predictive methods in agriculture often rely on linear and additive assumptions, which may limit their ability to capture complex relationships between variables fully. In contrast, non-parametric alternatives have gained attention for their flexibility in identifying nonlinear and interactive patterns. One key advantage of these methods is that they do not require prior knowledge of hidden patterns or relationships (Montesinos-López et al., 2021), making them promising tools for integrating multi-omics datasets, where interactions are often complex and not easily modeled with parametric approaches.

ML approaches can be broadly divided into two categories. The first category consists of global optimization machines, which include complex algorithms trained on large datasets to find patterns. A prominent example in this group is deep neural networks, widely applied in computer vision for phenotyping purposes (Jiang & Li, 2020). These models have solved problems in image recognition, text processing, or game strategy, but typically require substantial data and computational resources (Montesinos-López et al., 2021).

The second category includes ensemble algorithms, which combine the outputs of simpler models, often called “weaker learners,” to form a more powerful decision-making system. One common ensemble technique is bagging, which reduces overfitting by training multiple models on random subsets of observations and features. Popular examples of ensemble methods are random forests and gradient boosting machines, both of which have shown strong performance across various fields (Breiman, 2001) and are especially effective in GS (Barreto et al., 2024; Montesinos López et al., 2022; Ogotu et al., 2011).

As mentioned in the previously described methods, homogeneous ensembles utilize the same process for prediction.

In contrast, stacking ensemble learning (SEL) combines predictions from diverse methods, often overperforming homogeneous ensembles (Mendes-Moreira et al., 2012). SEL has demonstrated success in GS, enhancing prediction accuracy in various species, including pine (Liang et al., 2021) and arabica coffee (*Coffea arabica*) (Nascimento et al., 2024), frequently significantly outperforming GBLUP across evaluated traits.

## 5.3 | Explicit integration of genetic and non-genetic factors: Crop models

Crop growth models are computational simulations that leverage physiological prior knowledge to predict plant development and phenotypes. Once genotype-specific parameters are defined, these models incorporate environmental and management information to simulate growth dynamics (Cooper et al., 2014). Due to their explicit representation of physiological and environmental interactions, crop models are considered symbolic approaches, contrasting with statistical models' implicit and sub-symbolic nature (Negus et al., 2024).

For example, models such as CropSyst (Stöckle et al., 2003) integrate user-defined management practices and weather data with predefined physiological coefficients to compute internal parameters governing crop growth. These stochastic simulations generate predictions for key agronomic traits, such as yield and canopy cover, while accounting for non-genetic factors like environmental variability and management decisions (Gavasso-Rita et al., 2024). However, a major limitation is their limited capacity to explicitly model genetic variation, making them less suitable for plant breeders focused on optimizing additive genetic effects to enhance genetic gains. The combination of GS methods and crop growth models to address this limitation has been studied in Messina et al. (2018).

## 5.4 | Perspectives and challenges in multi-omics prediction

Although multi-omics data can often be structured similarly to well-established datasets, allowing for seamless integration into existing analytical frameworks, these frameworks were originally designed with different assumptions. For example, methods developed for SNP arrays account for LD and population structure in specific ways that may not directly translate to other omic layers. As a result, applying these models without proper adaptations may lead to misleading conclusions. Therefore, further research is needed to establish appropriate quality control procedures for each dataset before it can be reliably used for phenotype prediction. These procedures include variable transformation (e.g., standardization to  $z$ -scores, discretization into classes), criteria for outlier

removal, thresholds for missing data, and optimal imputation strategies.

On the modeling side, a key challenge is establishing fair benchmarking strategies to compare models effectively, as discussed by F. Wang et al. (2025). We also need to put on the table the trade-off between investing in multi-omics data and the effective gains in prediction accuracy. A dataset with extensive omic layers may slightly improve prediction performance, but its practical utility is limited if the computational and data collection costs outweigh the benefits.

Although multi-omics is appealing in terms of predictive ability, it does not provide an exact measure of the amount of variability explained by each one of the omics. Variance components help prioritize the information layers that most effectively contribute to prediction; however, in the absence of one of the omics, the remaining error variance will capture its signal. The orthogonalization between omics helps better represent the importance of the different layers.

Finally, further research is needed on the interpretability and real-world deployment of multi-omics models beyond CV. A critical question is which omic data can be reused across breeding cycles and which needs to be recollected for each prediction scenario. For example, if yield predictions rely on vegetative indices measurements, but obtaining those indices requires planting the genotypes, their predictive utility for early selection is constrained. Addressing these challenges will be essential for integrating multi-omics models into breeding programs.

## 6 | SIMULATIONS IN PLANT BREEDING: VIRTUAL BREEDING TO DESIGN REAL-WORLD STRATEGIES

When foundational elements such as clear breeding objectives, stringent selection criteria, a rich germplasm pool, and well-defined target environments are in place, the core determinant of a breeding program's effectiveness is the strategic use of resources to maximize genetic gain, driving continuous improvement in selected traits (Vieira et al., 2025). The path to significant genetic gains in breeding pipelines involves optimizing genomic and phenotypic selection approaches. Achieving this progress requires addressing multiple factors, from the number of breeding cycles and population size to parent selection, genomic prediction accuracy, genetic diversity, and effective cost management, significantly enhancing overall efficiency and output.

Computational simulations have emerged as essential tools in modern breeding strategies to support these advancements. In plant breeding and genetics, simulations employ mathematical models to replicate biological systems and address specific challenges. These models are categorized as deterministic or stochastic (Covarrubias-Pazarán et al.,

2022). Deterministic models, for instance, use quantitative genetics equations to predict selection responses, relying on key parameters such as selection intensity, heritability, and accuracy. Although deterministic models are useful, they fail to fully represent key breeding processes like crossing, generation advancement, and genetic introgression.

Stochastic simulations, by contrast, generate both genotypic and phenotypic data for each genetic entity, allowing for more accurate modeling of critical breeding stages such as recombination, evaluation, and selection. Furthermore, simulations deliver key insights into the factors that impact genetic gain, prediction accuracy, and cost-effectiveness across various conditions (Gorjanc et al., 2018; Hickey et al., 2014). By integrating critical variables like gene information, crossing schemes, population size, and selection intensity, simulations provide valuable insights that optimize breeding parameters, maximizing genetic gains and ensuring the efficient use of resources (Gorjanc et al., 2018; Hickey et al., 2014; X. Li et al., 2012).

However, the effective application of these simulation tools in breeding programs depends on how accurately practical strategies are translated into digital models. Translating practical breeding strategies into digital models requires a thoughtful balance of abstraction and simplification to capture biological and operational processes while considering the programming constraints. Breeding itself is a complex, iterative process of crossing, evaluation, and selection, which must be effectively encoded into simulations to reflect the real-world dynamics of breeding cycles. In this context, simulations typically operate under idealized conditions—such as error-free markers, no  $G \times E$  interactions, absence of epistasis, limited germplasm exchange, and flawless linkage maps—that may not fully capture the complexities of real-world breeding (Gaynor et al., 2017). As a result, interpreting the results requires critically assessing these assumptions and understanding how they influence the predicted outcomes.

Building on this, it is important to recognize both the strengths and limitations of simulations compared to real-world breeding practices. Besides being extremely helpful, simulations often achieve higher accuracy than real-world breeding programs, owing to idealized conditions (Gaynor et al., 2017). However, real-world breeding involves the consideration of multiple traits—such as agronomic performance, disease resistance, and end-use quality—which adds significant complexity. For simulations to provide meaningful insights, they must effectively model the complexities of multi-trait selection. Contemporary simulations are marked by their capacity to accurately model meiotic processes, including crossing over and crossover interference, by applying coalescent and gene-drop methods. These methods are integral to both backward-in-time and forward-in-time simulations, enabling a more detailed exploration of genetic inheritance and evolution (Hickey & Gorjanc, 2012).

Despite their advantages, simulations are not without limitations and challenges. One limitation of breeding scheme simulations is introducing random error, akin to real-world processes such as meiosis and genotype evaluation, where random deviations from the genotypic value are often observed (Jannink et al., 2023). Moreover, genomic predictions require large-scale linear mixed models, which are computationally demanding and can be time-intensive for real and simulated data. Despite inherent challenges, simulations bring significant advantages by enabling the investigation of various breeding scenarios, genetic models, and strategies. This flexibility aids in identifying the most efficient pathways to achieving target cultivars (X. Li et al., 2012). Moreover, genetic gain from simulations is often seen as the theoretical peak, but real-world programs are still influenced by attrition factors that reduce their practical realization (Gaynor et al., 2017). Integrating simulation insights into practice involves validating predictions with actual data, incorporating feedback loops from field trials, and refining models over time to align theoretical predictions with real-world applicability (Vieira et al., 2025).

Various computational tools have been developed to support these efforts to address different simulation objectives and contexts. Several R packages currently offer tailored solutions to meet specific simulation needs. AlphaSimR is highly regarded for its adaptability in modeling complex genetic architectures, selection schemes, and long-term breeding strategies. AlphaDrop supports studies of structured populations by enabling precise tracking of genetic inheritance across crossing programs. Synbreed is a standout tool for genomic prediction and analysis, utilizing SNP data to optimize breeding outcomes.

## 6.1 | Stochastic simulation applications

Conducting real-life experiments across various scenarios is often prohibitively expensive, time-consuming, and sometimes impossible due to the complexity and the need for repeated trials. In practical breeding programs, each experiment represents a single iteration, which limits the ability to observe and analyze consistent and reliable trends. Conversely, simulations enable the repetition of thousands of scenarios, allowing for the identification of central tendencies and the extraction of valuable insights on expected outcomes. Thus, simulations bridge the gap between theory and application, allowing breeders to computationally explore strategies and determine the most effective pathways to success (Figure 5). By computing and exploring diverse genetic strategies and scenarios, simulations offer a more efficient, cost-effective, and comprehensive approach to optimizing breeding strategies, making them essential for informed decision-making and advancing breeding pro-

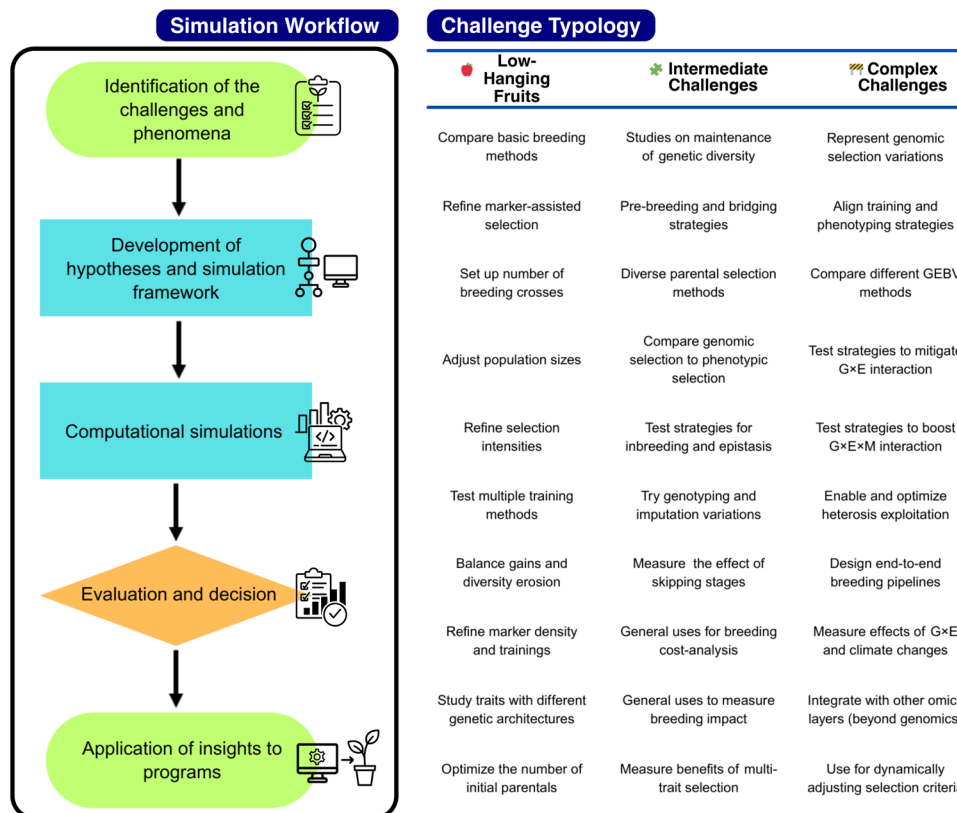
gram objectives. They are useful for maximizing genetic gain, minimizing the loss of genetic variance, and ensuring that resources are utilized cost-effectively, thereby driving breeding program optimization.

These applications facilitate the comparison of various breeding strategies, encompassing phenotypic selection, MAS, and GS (Fritsche-Neto et al., 2024; X. Li et al., 2012; Muleta et al., 2019; Sabadin et al., 2022). They provide a comprehensive view incorporating early- and late-stage processes, enabling informed decisions throughout the breeding pipeline. As breeding programs evolve and adopt new methodologies, the demand for simulations to assess their benefits and explore alternatives remains critical. Simulations offer significant advantages by allowing researchers to explore diverse scenarios, genetic models, and methodologies, helping them identify the most effective strategies for developing target cultivars, as well as different breeding strategies by integrating genetic data, crossing schemes, propagation methods, population sizes, selection intensities, and the number of generations involved.

Consequently, simulations are unparalleled tools for modern breeding programs, addressing key challenges and advancing innovative methodologies across various objectives. These include (a) balancing genetic diversity and genetic gains, navigating the delicate trade-off between maintaining genetic diversity and achieving substantial genetic gains; (b) enabling pre-breeding and parental selection strategies on how to integrate elite and non-elite genetic materials into advanced breeding pipelines; (c) advancing GS for sustainable gains by developing insights into configuring training sets and aligning genotyping with phenotyping efforts; (d) comparing breeding methods by comprehensive comparisons between phenotypic, marker-assisted, and GS; (e) optimizing breeding program parameters, such as population sizes, selection intensities, genomic training sets, marker densities, and so on; (f) refining breeding value estimation by exploring various simulations, such as Bayesian and BLUP models, under a variety of genetic architecture and complex traits; (g) addressing complexities, devising strategies to circumvent inbreeding and epistasis, managing  $G \times E$  interactions, and exploiting heterosis effectively; (h) designing cost-effective breeding programs, choosing roadmaps that balance resource efficiency with genetic progress, and providing actionable insights across multiple dimensions.

## 6.2 | Perspectives and challenges for stochastic simulations

In plant breeding and genetics, simulations replicate real-world biological conditions to address specific challenges and phenomena. Their growing importance stems from the need to optimize resource use concerning program scale and genetic



**FIGURE 5** A schematic representation of simulation in plant breeding, demonstrating how frameworks mimic biological conditions to connect theoretical concepts with real-world breeding practices.  $G \times E$ , genotype-by-environment; GEBV, genomic estimated breeding value.

gains. The complexities of genetic selection make simulations crucial, as breeding pipelines are dynamic, influenced by business decisions and advancements in genotyping technologies, and fluctuate in resource allocation from year to year.

Building upon this foundational role, simulations will continue to play a role in shaping the future of plant breeding by providing insights that help refine breeding pipelines and maximize genetic gains. Future research in this area is likely to focus increasingly on genetic diversity, pre-breeding efforts, and advances in GS strategies. Also, simulations are poised to explore several avenues to enhance models and their applications. One key focus area involves refining simulations to capture better  $G \times E$  interactions, particularly in the context of climate challenges. By integrating more sophisticated environmental variables and stress factors, researchers could improve the accuracy of breeding strategies, ensuring that new cultivars are resilient and sustainable under diverse conditions.

In parallel with environmental modeling, another potential direction for research is expanding simulations to multi-omic selection frameworks. As omics technologies advance, future studies may aim to integrate multiple data layers, such as epigenetics and transcriptomics, into simulations. This approach would establish a connection with the previously mentioned

genotype–environment enhancement opportunity while also contributing to a more comprehensive understanding of the mechanisms through which genetic networks influence trait expression. Such insights are essential for advancing our knowledge of complex traits and developing more refined and targeted breeding strategies.

Moreover, integrating simulation models with AI and ML represents a major frontier. Combining simulation models with AI and ML algorithms has significant potential. These technologies could be harnessed to optimize breeding pipelines by automatically identifying patterns, predicting outcomes, and dynamically adjusting selection criteria in real time. This approach would allow breeding programs to become increasingly adaptive and data-informed, facilitating faster genetic gains and greater breeding efficiency, primarily focusing on delivering value to farmers and consumers.

Complementing these advancements, another critical focus is expanding simulations to support multi-trait selection frameworks (Akdemir et al., 2018), while refining key breeding parameters such as parental number, population size, and selection timing in phenotypic and genomic programs. It is worth noting that exploring various GS strategies remains a relevant and continually evolving area of research as this approach becomes increasingly routine in breeding programs.

In our view, the future of plant breeding is undeniably integrated with the evolution of simulation approaches. As breeding programs grow and evolve to be more complex, simulations will be essential for navigating the multifaceted challenges ahead, from climate change to improved genetic gain in favorable conditions. The continued advancement of these tools can act as a game changer, allowing breeders to optimize genetic gains with unprecedented precision by combining AI, genomic insights, and environmental factors with still valid breeding concepts in quantitative genetics, field testing, and great training sets. As we embrace the power of innovations in plant breeding, simulations will no longer refine breeding strategies—they will redefine them for the benefit of farmers and consumers.

## 7 | FINAL CONSIDERATIONS

Although prediction-based breeding provides useful frameworks for genetic improvement, models are limited by the conditions under which they are developed and validated. Predictive models, similar to a linear regression, generate reliable outputs only within the scope of the data used for calibration. In GS, models are expected to perform for testing populations that maintain genetic relationships with the training set. Extending predictions to unrelated or distantly related groups without recalibration often leads to unreliable results. Similarly,  $G \times E$  models are conditioned by the environmental range captured in the training data. Unlike a simple regression, where boundary ranges are easily observed, in GS and  $G \times E$  modeling, the diversity and structure of the genetic and environmental samples define the limits. Extrapolation beyond these domains introduces uncertainty and error. Prediction models must therefore be used within their validated application space, and breeding decisions should rely on the specific conditions under which the models have demonstrated acceptable performance.

A similar challenge arises in MAS and GWAS, where practical application in breeding programs remains limited, particularly for quantitative traits. Two decades ago, Holland (2004) warned that MAS faced major obstacles: low precision in QTL mapping and the difficulty of extrapolating such findings to distinct breeding populations. Nevertheless, many current efforts remain focused on searching for and interpreting QTLs based on candidate genes, often anchored in functional descriptions available in repositories such as NCBI (National Center for Biotechnology Information) or GenBank. Cui et al. (2015) reinforced these limitations, noting that QTLs identified in experimental populations tend to exhibit unpredictable effects when tested in new genetic or environmental contexts. While this approach may be valid for

monogenic or qualitative traits, where loci of large effect have been consistently mapped and successfully applied, the same does not hold for complex traits. In such cases, the emphasis should shift toward predictive models with generalization capacity, even at the expense of mechanistic interpretation. The current challenge is not to explain everything, but to predict reliably.

This expectation for causal explanations—previously concentrated in genomics—has now extended to other layers of omics data, including phenomics, enviromics, and integrated multi-omics approaches. Although these sources enhance the representation of biological complexity, attempts to isolate individual effects, interpret specific spectral bands, or assign fixed meanings to environmental variables risk reproducing the same pitfalls long recognized in QTL analysis: relying on associations that appear biologically plausible but fail under validation. P. Wang et al. (2024) demonstrated that models built with different types of omics data can achieve similar accuracies, even when the sets of genes identified as “important” do not coincide—or even include the genes traditionally associated with the trait, such as those involved in flowering. Similarly, Robert et al. (2022) showed that NIR spectra, even without direct phenotypic connection to the target trait, capture sufficient genetic similarity to enable robust prediction—suggesting that useful signals may be distributed in a non-interpretable manner, akin to invisible LD. The lesson remains: the focus must be on the predictive ability of the model, validated across diverse scenarios, and ensuring independence between calibration and testing sets. In high-dimensional data, forcing causal interpretations can lead to overfitting or fragile conclusions. Ultimately, we are breeders—our commitment is to efficient selection, not constructing explanatory narratives.

An additional dimension in prediction-based breeding is the strategic use of stochastic simulations. By virtually modeling breeding cycles, environmental scenarios, and integrating omic data, simulations enable the optimization of selection schemes, training populations, and predictive pipelines before their real-world implementation. They provide a controlled setting to evaluate model robustness across various genetic architectures, heritabilities, and  $G \times E$  structures, thereby minimizing the risks associated with overfitting and extrapolation. Especially when working with high-dimensional data, simulations efficiently anticipate pitfalls that empirical calibration alone may not reveal. Simulation-based assessments strengthen the predictive strategy by ensuring that observed data and prospective, stress-tested outcomes inform decisions. In an era of expanding data sources and modeling tools, prediction-based breeding must stay anchored to validation principles: independence between calibration and testing, emphasis on real predictive performance, and pragmatic decision-making. Ultimately, breeders advance by prioritizing reliable selection outcomes, rather than exhaustive causal

explanations, as the guiding criterion for innovation and impact.


## AUTHOR CONTRIBUTIONS

**Roberto Fritsche-Neto:** Conceptualization; supervision; writing—original draft; writing—review and editing. **Rafael T. Resende:** Conceptualization; writing—original draft; writing—review and editing. **Tiago Olivoto:** Conceptualization; writing—original draft; writing—review and editing. **Julian Garcia-Abadillo:** Writing—original draft; writing—review and editing. **Moyses Nascimento:** Writing—original draft; writing—review and editing. **Marco Antônio M. Bahia:** Writing—original draft; Writing—review and editing. **Diego Jarquin:** Conceptualization; writing—original draft; writing—review and editing. **Rafael Augusto Vieira:** Conceptualization; writing—original draft; writing—review and editing.

## CONFLICT OF INTEREST STATEMENT


The authors declare no conflicts of interest.

## ORCID

Roberto Fritsche-Neto  <https://orcid.org/0000-0003-4310-0047>

Rafael T. Resende  <https://orcid.org/0000-0002-7139-708X>

Tiago Olivoto  <https://orcid.org/0000-0002-0241-9636>

Rafael Augusto Vieira  <https://orcid.org/0009-0008-1862-4352>

## REFERENCES

- Abdulridha, J., Min, A., Rouse, M. N., Kianian, S., Isler, V., & Yang, C. (2023). Evaluation of stem rust disease in wheat fields by drone hyperspectral imaging. *Sensors*, 23(8), 4154. <https://doi.org/10.3390/S23084154>
- Adak, A., DeSalvio, A. J., Arik, M. A., & Murray, S. C. (2024). Field-based high-throughput phenotyping enhances phenomic and genomic predictions for grain yield and plant height across years in maize. *G3 Genes|Genomes|Genetics*, 14(7), jkae092. <https://doi.org/10.1093/G3JOURNAL/JKAE092>
- Adunola, P., Tavares Flores, E., Riva-Souza, E. M., Ferrão, M. A. G., Senra, J. F. B., Comério, M., Espindula, M. C., Verdin Filho, A. C., Volpi, P. S., Fonseca, A. F. A., Ferrão, R. G., Munoz, P. R., & Ferrão, L. F. V. (2024). A comparison of genomic and phenomic selection methods for yield prediction in *Coffea canephora*. *The Plant Phenome Journal*, 7(1), e20109. <https://doi.org/10.1002/ppj.20109>
- Akdemir, D., Beavis, W., Fritsche-Neto, R., Singh, A. K., & Isidro-Sánchez, J. (2018). Multi-objective optimized genomic breeding strategies for sustainable food improvement. *Heredity*, 122(5), 672–683. <https://doi.org/10.1038/s41437-018-0147-1>
- de Andrade Júnior, A. S., da Silva, S. P., Setúbal, I. S., de Souza, H. A., de M J Vieira, P. F., & das C N Casari, R. A. (2022). Predicting soybean grain yield using aerial drone images. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 26(6), 466–476. <https://doi.org/10.1590/1807-1929/agriambi.v26n6p466-476>
- Angione, C., Conway, M., & Lió, P. (2016). Multiplex methods provide effective integration of multi-omic data in genome-scale models. *BMC Bioinformatics*, 17, Article 83. <https://doi.org/10.1186/s12859-016-0912-1>
- Annicchiarico, P. (2002). *Genotype × environment interactions: Challenges and opportunities for plant breeding and cultivar recommendations* (FAO plant production and protection Paper no. 174). FAO. [https://www.researchgate.net/publication/285754548\\_Genotype\\_x\\_environment\\_interactions\\_Challenges\\_and\\_opportunities\\_for\\_plant\\_breeding\\_and\\_cultivar\\_recommendations](https://www.researchgate.net/publication/285754548_Genotype_x_environment_interactions_Challenges_and_opportunities_for_plant_breeding_and_cultivar_recommendations)
- Araújo, M. S., Chaves, S. F. S., Dias, L. A. S., Ferreira, F. M., Pereira, G. R., Bezerra, A. R. G., Alves, R. S., Heinemann, A. B., Breseghello, F., Carneiro, P. C. S., Krause, M. D., Costa-Neto, G., & Dias, K. O. G. (2024). GIS-FA: An approach to integrating thematic maps, factor-analytic, and envirotyping for cultivar targeting. *Theoretical and Applied Genetics*, 137(4), Article 80. <https://doi.org/10.1007/S00122-024-04579-Z>
- Bahia, M. A. M., Marcatti, G. E., Breseghello, F., Melo, P. G. S., Dias, K. O. G., Xu, Y., & Resende, R. T. (2025). Predictive ability of enviromic modeling in G×E interactions for upland rice site recommendations. *bioRxiv*. <https://doi.org/10.1101/2025.04.05.647387>
- Barreto, C. A. V., das Graças Dias, K. O., de Sousa, I. C., Azevedo, C. F., Nascimento, A. C. C., Guimarães, L. J. M., Guimarães, C. T., Pastina, M. M., & Nascimento, M. (2024). Genomic prediction in multi-environment trials in maize using statistical and machine learning methods. *Scientific Reports*, 14(1), Article 1062. <https://doi.org/10.1038/s41598-024-51792-3>
- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, P. N., Tanaka, E., & Jagan, M. (2024). Package “lme4” Convergence (Version 1.1-37) [Computer software]. CRAN. <https://cran.r-project.org/web/packages/lme4/lme4.pdf>
- Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? Archives of disease in childhood. *Education and Practice Edition*, 98(6), 236–238. <https://doi.org/10.1136/ARCHDISCHILD-2013-304340>
- Bernardo, R. (2020). *Breeding for quantitative traits in plants* (3rd ed.). Stemma Press. <http://stemmapress.com>
- Bonnett, D., Li, Y., Crossa, J., Dreisigacker, S., Basnet, B., Pérez-Rodríguez, P., Alvarado, G., Jannink, J. L., Poland, J., & Sorrells, M. (2022). Response to early generation genomic selection for yield in wheat. *Frontiers in Plant Science*, 12, 718611. <https://doi.org/10.3389/FPLS.2021.718611>
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231. <https://doi.org/10.1214/SS/1009213726>
- Calus, M. P. L., & Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution*, 43(1), Article 26. <https://doi.org/10.1186/1297-9686-43-26>
- Cooper, M., & Messina, C. D. (2021). Can we harness “enviromics” to accelerate crop improvement by integrating breeding and agronomy? *Frontiers in Plant Science*, 12, 735143. <https://doi.org/10.3389/FPLS.2021.735143>
- Cooper, M., Messina, C. D., Podlich, D., Totir, L. R., Baumgarten, A., Hausmann, N. J., Wright, D., & Graham, G. (2014). Predicting the future of plant breeding: Complementing empirical evaluation with genetic prediction. *Crop and Pasture Science*, 65(4), 311–336. <https://doi.org/10.1071/CP14007>

- Costa-Neto, G., Crespo-Herrera, L., Fradgley, N., Gardner, K., Bentley, A. R., Dreisigacker, S., Fritsche-Neto, R., Montesinos-López, O. A., & Crossa, J. (2023). Envirome-wide associations enhance multi-year genome-based prediction of historical wheat breeding data. *G3 Genes|Genomes|Genetics*, *13*(2), jkac313. <https://doi.org/10.1093/G3JOURNAL/JKAC313>
- Costa-Neto, G., Crossa, J., & Fritsche-Neto, R. (2021). Enviromic assembly increases accuracy and reduces costs of the genomic prediction for yield plasticity in maize. *Frontiers in Plant Science*, *12*, 717552. <https://doi.org/10.3389/FPLS.2021.717552>
- Costa-Neto, G., & Fritsche-Neto, R. (2021). Enviromics: Bridging different sources of data, building one framework. *Crop Breeding and Applied Biotechnology*, *21*(spe), e393521S12. <https://doi.org/10.1590/1984-70332021v21sa25>
- Costa-Neto, G., Fritsche-Neto, R., & Crossa, J. (2021). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity*, *126*(1), 92–106. <https://doi.org/10.1038/S41437-020-00353-1>
- Costa-Neto, G., Galli, G., Carvalho, H. F., Crossa, J., & Fritsche-Neto, R. (2021). EnvRtype : A software to interplay enviromics and quantitative genomics in agriculture. *G3 Genes|Genomes|Genetics*, *11*(4), jkab040. <https://doi.org/10.1093/g3journal/jkab040>
- Costa-Neto, G. M. F., Morais Júnior, O. P., Heinemann, A. B., de Castro, A. P., & Duarte, J. B. (2020). A novel GIS-based tool to reveal spatial trends in reaction norm: Upland rice case study. *Euphytica*, *216*(3), Article 37. <https://doi.org/10.1007/S10681-020-2573-4>
- Covarrubias-Pazarán, G. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS One*, *11*(6), e0156744. <https://doi.org/10.1371/JOURNAL.PONE.0156744>
- Covarrubias-Pazarán, G., Gebeyehu, Z., Gemenet, D., Werner, C., Labroo, M., Sirak, S., Coaldrake, P., Rabbi, I., Kayondo, S. I., Parkes, E., Kanju, E., Mbanjo, E. G. N., Agbona, A., Kulakow, P., Quinn, M., & Debaene, J. (2022). Breeding schemes: What are they, how to formalize them, and how to improve them? *Frontiers in Plant Science*, *12*, 791859. <https://doi.org/10.3389/FPLS.2021.791859>
- Crain, J., Mondal, S., Rutkoski, J., Singh, R. P., & Poland, J. (2018). Combining high-throughput phenotyping and genomic information to increase prediction and selection accuracy in wheat breeding. *The Plant Genome*, *11*(1), 170043. <https://doi.org/10.3835/PLANTGENOME2017.05.0043>
- Crossa, J. (1990). Statistical analyses of multilocation trials. *Advances in Agronomy*, *44*(C), 55–85. [https://doi.org/10.1016/S0065-2113\(08\)60818-4](https://doi.org/10.1016/S0065-2113(08)60818-4)
- Crossa, J., Fritsche-Neto, R., Montesinos-Lopez, O. A., Costa-Neto, G., Dreisigacker, S., Montesinos-Lopez, A., & Bentley, A. R. (2021). The modern plant breeding triangle: Optimizing the use of genomics, phenomics, and enviromics data. *Frontiers in Plant Science*, *12*, 651480. <https://doi.org/10.3389/fpls.2021.651480>
- Crossa, J., Montesinos-Lopez, O. A., Costa-Neto, G., Vitale, P., Martini, J. W. R., Runcie, D., Fritsche-Neto, R., Montesinos-Lopez, A., Pérez-Rodríguez, P., Gerard, G., Dreisigacker, S., Crespo-Herrera, L., Pierre, C. S., Lillemo, M., Cuevas, J., Bentley, A., & Ortiz, R. (2025). Machine learning algorithms translate big data into predictive breeding accuracy. *Trends in Plant Science*, *30*(2), 167–184. <https://doi.org/10.1016/J.TPLANTS.2024.09.011>
- Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, *22*(11), 961–975. <https://doi.org/10.1016/j.tplants.2017.08.011>
- Crossa, J., Vargas, M., Van Eeuwijk, F. A., Jiang, C., Edmeades, G. O., & Hoisington, D. (1999). Interpreting genotype x environment interaction in tropical maize using linked molecular markers and environmental covariables. *Theoretical and Applied Genetics*, *99*(3–4), 611–625. <https://doi.org/10.1007/S001220051276>
- Cruz, D. D. M., Heinemann, A. B., Marcatti, G. E., & Resende, R. T. (2025). Defining the target population of environments (TPE) for enviromics studies using R-based GIS tools defining the target population of environments (TPE) for enviromics studies using R-based GIS tools. *Crop Breeding and Applied Biotechnology*, *25*(1), 50822519. <https://doi.org/10.1590/1984-70332024v25n1s09>
- Cui, Y., Li, R., Li, G., Zhang, F., Zhu, T., Zhang, Q., Ali, J., Li, Z., & Xu, S. (2020). Hybrid breeding of rice via genomic selection. *Plant Biotechnology Journal*, *18*(1), 57–67. <https://doi.org/10.1111/PBI.13170>
- Cui, Y., Zhang, F., Xu, J., Li, Z., & Xu, S. (2015). Mapping quantitative trait loci in selected breeding populations: A segregation distortion approach. *Heredity*, *115*(6), 538–546. <https://doi.org/10.1038/hdy.2015.56>
- Dijkhuizen, R., van Eijnatten, A. L., Mehrem, S. L., van den Bergh, E., van Lieshout, J., Spaninks, K., Kaandorp, S., Offringa, R., Proveniers, M., van den Ackerveken, G., & Snoek, B. L. (2025). From aerial drone to quantitative trait locus: Leveraging next-generation phenotyping to reveal the genetics of color and height in field-grown *Lactuca sativa*. *The Plant Journal*, *123*(3), e70405. <https://doi.org/10.1111/tj.70405>
- D’Odorico, P., Besik, A., Wong, C. Y. S., Isabel, N., & Ensminger, I. (2020). High-throughput drone-based remote sensing reliably tracks phenology in thousands of conifer seedlings. *New Phytologist*, *226*(6), 1667–1681. <https://doi.org/10.1111/nph.16488>
- DoVale, J. C., Carvalho, H. F., Sabadin, F., & Fritsche-Neto, R. (2022). Genotyping marker density and prediction models effects in long-term breeding schemes of cross-pollinated crops. *Theoretical and Applied Genetics*, *135*(12), 4523–4539. <https://doi.org/10.1007/S00122-022-04236-3>
- Eberius, M., & Lima-Guerra, J. (2009). High-throughput plant phenotyping—Data acquisition, transformation, and analysis. In D. Edwards, J. Stajich, & D. Hansen (Eds.), *Bioinformatics: Tools and applications* (pp. 259–278). Springer. [https://doi.org/10.1007/978-0-387-92738-1\\_13](https://doi.org/10.1007/978-0-387-92738-1_13)
- Fritsche-Neto, R. (2024). SoilType: An R package to interplay soil characterization in plant science. *Agronomy Journal*, *116*(3), 848–854. <https://doi.org/10.1002/AGJ2.21383>
- Fritsche-Neto, R., Ali, J., De Asis, E. J., Allahgholipour, M., & Labroo, M. R. (2024). Improving hybrid rice breeding programs via stochastic simulations: Number of parents, number of hybrids, tester update, and genomic prediction of hybrid performance. *Theoretical and Applied Genetics*, *137*(1), Article 3. <https://doi.org/10.1007/S00122-023-04508-6>
- Fritsche-Neto, R., & Borém, A. (2015). *Phenomics: How next-generation phenotyping is revolutionizing plant breeding*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-13677-6>
- Fritsche-Neto, R., Galli, G., Borges, K. L. R., Costa-Neto, G., Alves, F. C., Sabadin, F., Lyra, D. H., Morais, P. P. P., Braatz de Andrade, L. R., Granato, I., & Crossa, J. (2021). Optimizing genomic-enabled prediction in small-scale maize hybrid breeding programs: A roadmap

- review. *Frontiers in Plant Science*, 12, 658267. <https://doi.org/10.3389/FPLS.2021.658267>
- Galli, G., Horne, D. W., Collins, S. D., Jung, J., Chang, A., Fritsche-Neto, R., & Rooney, W. L. (2020). Optimization of UAS-based high-throughput phenotyping to estimate plant health and grain yield in sorghum. *The Plant Phenome Journal*, 3(1), e20010. <https://doi.org/10.1002/PPJ2.20010>
- Gano, B., Bhadra, S., Vilbig, J. M., Ahmed, N., Sagan, V., & Shakoob, N. (2024). Drone-based imaging sensors, techniques, and applications in plant phenotyping for crop breeding: A comprehensive review. *Plant Phenome Journal*, 7(1), e20100. <https://doi.org/10.1002/PPJ2.20100>
- Garrick, D. J., Taylor, J. F., & Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution*, 41(1), Article 55. <https://doi.org/10.1186/1297-9686-41-55>
- Gavasso-Rita, Y. L., Papalexiou, S. M., Li, Y., Elshorbagy, A., Li, Z., & Schuster-Wallace, C. (2024). Crop models and their use in assessing crop production and food security: A review. *Food and Energy Security*, 13(1), e503. <https://doi.org/10.1002/FES3.503>
- Gaynor, R. C., Gorjanc, G., Bentley, A. R., Ober, E. S., Howell, P., Jackson, R., Mackay, I. J., Hickey, J. M., Gaynor, R. C., Gorjanc, G., Hickey, J. M., Institute, R., Dick, R., Bent-Ley, A. R., Ober, E. S., Howell, P., Jackson, R., Mackay, I. J., & Bingham, J. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Science*, 57(5), 2372–2386. <https://doi.org/10.2135/CROPSCI2016.09.0742>
- Gaynor, R. C., Gorjanc, G., & Hickey, J. M. (2021). AlphaSimR: An R package for breeding program simulations. *G3 Genes/Genomes/Genetics*, 11(2), jkaa017. <https://doi.org/10.1093/g3journal/jkaa017>
- González, F., García-Abadillo, J., & Jarquín, D. (2025). Introducing CHiDO—A No Code Genomic Prediction software implementation for the characterization and integration of driven omics. *The Plant Genome*, 18(1), e20519. <https://doi.org/10.1002/TPG2.20519>
- Gorjanc, G., Gaynor, R. C., & Hickey, J. M. (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics*, 131(9), 1953–1966. <https://doi.org/10.1007/S00122-018-3125-3>
- Gu, Y., Wang, Y., Wu, Y., Warner, T. A., Guo, T., Ai, H., Zheng, H., Cheng, T., Zhu, Y., Cao, W., & Yao, X. (2024). Novel 3D photosynthetic traits derived from the fusion of UAV LiDAR point cloud and multispectral imagery in wheat. *Remote Sensing of Environment*, 311, 114244. <https://doi.org/10.1016/J.RSE.2024.114244>
- Guadagno, C. R., Millar, D., Lai, R., Mackay, D. S., Pleban, J. R., McClung, C. R., Weinig, C., Wang, D. R., & Ewers, B. E. (2020). Use of transcriptomic data to inform biophysical models via Bayesian networks. *Ecological Modelling*, 429, 109086. <https://doi.org/10.1016/j.ecolmodel.2020.109086>
- Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., & Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genetics*, 15(1), Article 30. <https://doi.org/10.1186/1471-2156-15-30>
- Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., Grondona, M., Zambelli, A., Windhausen, V. S., Mathews, K., & Gorjanc, G. (2014). Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Science*, 54(4), 1476–1488. <https://doi.org/10.2135/CROPSCI2013.03.0195>
- Hickey, J. M., & Gorjanc, G. (2012). Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. *G3 Genes/Genomes/Genetics*, 2(4), 425–427. <https://doi.org/10.1534/G3.111.001297>
- Holland, J. B. (2004). Implementation of molecular markers for quantitative traits in breeding programs—challenges and opportunities. In *Proceedings of the 4th International Crop Science Congress*. Brisbane, Australia. <https://www.ars.usda.gov/ARSUserFiles/60701500/Publications/Holland/Holland4thInt%20CropSciCong04.pdf>
- Holland, J. B., & Piepho, H.-P. (2024). Don't BLUP twice. *G3 Genes/Genomes/Genetics*, 14(12), jkae250. <https://doi.org/10.1093/g3journal/jkae250>
- Jannink, J. L., Astudillo, R., & Frazier, P. (2023). Insight into a two-part plant breeding scheme through Bayesian optimization of budget allocations. *Crop Science*, 65(1), e21124. <https://doi.org/10.1002/CSC2.21124>
- Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J., & de los Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, 127(3), 595–607. <https://doi.org/10.1007/S00122-013-2243-1>
- Jia, Y., & Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, 192(4), 1513–1522. <https://doi.org/10.1534/GENETICS.112.144246>
- Jiang, Y., & Li, C. (2020). Convolutional neural networks for image-based high-throughput plant phenotyping: A review. *Plant Phenomics*, 2020, 4152816. <https://doi.org/10.34133/2020/4152816>
- Kang, M. S., Harville, B. G., & Gorman, D. P. (1989). Contribution of weather variables to genotype × environment interaction in soybean. *Field Crops Research*, 21(3–4), 297–300. [https://doi.org/10.1016/0378-4290\(89\)90011-7](https://doi.org/10.1016/0378-4290(89)90011-7)
- Kaushal, S., Gill, H. S., Billah, M. M., Khan, S. N., Halder, J., Bernardo, A., Amand, P. S., Bai, G., Glover, K., Maimaitijiang, M., & Sehgal, S. K. (2024). Enhancing the potential of phenomic and genomic prediction in winter wheat breeding using high-throughput phenotyping and deep learning. *Frontiers in Plant Science*, 15, 1410249. <https://doi.org/10.3389/FPLS.2024.1410249>
- Knoch, D., Werner, C. R., Meyer, R. C., Riewe, D., Abbadi, A., Lücke, S., Snowdon, R. J., & Altmann, T. (2021). Multi-omics-based prediction of hybrid performance in canola. *Theoretical and Applied Genetics [Theoretische Und Angewandte Genetik]*, 134(4), 1147–1165. <https://doi.org/10.1007/S00122-020-03759-X>
- Labroo, M. R., Ali, J., Aslam, M. U., de Asis, E. J., dela Paz, M. A., Sevilla, M. A., Lipka, A. E., Studer, A. J., & Rutkoski, J. E. (2021). Genomic prediction of yield traits in single-cross hybrid rice (*Oryza sativa* L.). *Frontiers in Genetics*, 12, 692870. <https://doi.org/10.3389/FGENE.2021.692870>
- Li, D., Quan, C., Song, Z., Li, X., Yu, G., Li, C., & Muhammad, A. (2021). High-throughput plant phenotyping platform (HT3P) as a novel tool for estimating agronomic traits from the lab to the field. *Frontiers in Bioengineering and Biotechnology*, 8, 623705. <https://doi.org/10.3389/FBIOE.2020.623705>
- Li, X., Zhu, C., Wang, J., & Yu, J. (2012). Computer simulation in plant breeding. *Advances in Agronomy*, 116, 219–264. <https://doi.org/10.1016/B978-0-12-394277-7.00006-3>
- Liang, M., Chang, T., An, B., Duan, X., Du, L., Wang, X., Miao, J., Xu, L., Gao, X., Zhang, L., Li, J., & Gao, H. (2021). A stacking ensemble

- learning framework for genomic prediction. *Frontiers in Genetics*, *12*, 600040. <https://doi.org/10.3389/FGENE.2021.600040>
- Lu, X., Shen, Y., Xie, J., Yang, X., Shu, Q., Chen, S., Shen, Z., & Cen, H. (2024). Phenotyping of panicle number and shape in rice breeding materials based on unmanned aerial vehicle imagery. *Plant Phenomics*, *6*, 0265. <https://doi.org/10.34133/PLANTPHENOMICS.0265>
- Mahmood, U., Li, X., Fan, Y., Chang, W., Niu, Y., Li, J., Qu, C., & Lu, K. (2022). Multi-omics revolution to promote plant breeding efficiency. *Frontiers in Plant Science*, *13*, 1062952. <https://doi.org/10.3389/fpls.2022.1062952>
- Malosetti, M., Ribaut, J. M., & van Eeuwijk, F. A. (2013). The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology*, *4*, Article 44. <https://doi.org/10.3389/fphys.2013.00044>
- Marcatti, G. E., Resende, R. T., Resende, M. D. V., Ribeiro, C. A. A. S., dos Santos, A. R., da Cruz, J. P., & Leite, H. G. (2017). GIS-based approach applied to optimizing recommendations of *Eucalyptus* genotypes. *Forest Ecology and Management*, *392*, 144–153. <https://doi.org/10.1016/j.foreco.2017.03.006>
- Martins, A. S. (2004). *Aplicação de Sistema de Informações Geográficas no Estudo da Interação de Genótipos com Ambientes*. Universidade Federal de Goiás. [https://files.cercomp.ufg.br/weby/up/396/o/SIG-GE\\_soja\\_%28Martins\\_2004\\_Ms\\_PPGA%29.pdf](https://files.cercomp.ufg.br/weby/up/396/o/SIG-GE_soja_%28Martins_2004_Ms_PPGA%29.pdf)
- Matias, F. I., Caraza-Harter, M. V., & Endelman, J. B. (2020). FIELDimageR: An R package to analyze orthomosaic images from agricultural field trials. *Plant Phenome Journal*, *3*(1), e20005. <https://doi.org/10.1002/ppj2.20005>
- Mendes-Moreira, J., Soares, C., Jorge, A. M., & Sousa, J. F. D. (2012). Ensemble approaches for regression. *ACM Computing Surveys*, *45*(1), 1–40. <https://doi.org/10.1145/2379776.2379786>
- Merrick, L. F., Burke, A. B., Chen, X., & Carter, A. H. (2021). Breeding with major and minor genes: genomic selection for quantitative disease resistance. *Frontiers in Plant Science*, *12*, 713667. <https://doi.org/10.3389/fpls.2021.713667>
- Merrick, L. F., Herr, A. W., Sandhu, K. S., Lozada, D. N., & Carter, A. H. (2022). Optimizing plant breeding programs for genomic selection. *Agronomy*, *12*(3), 714. <https://doi.org/10.3390/agronomy12030714>
- Messina, C. D., Technow, F., Tang, T., Totir, R., Gho, C., & Cooper, M. (2018). Leveraging biological insight and environmental variation to improve phenotypic prediction: Integrating crop growth models (CGM) with whole genome prediction (WGP). *European Journal of Agronomy*, *100*, 151–162. <https://doi.org/10.1016/j.eja.2018.01.007>
- Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829. <https://doi.org/10.1093/GENETICS/157.4.1819>
- Meyer, P. (2015). Epigenetic variation and environmental change. *Journal of Experimental Botany*, *66*, 3541–3548. <https://doi.org/10.1093/jxb/eru502>
- Moheinizade, S., Pham, H., Han, Y., Dobbels, A., & Hu, G. (2022). An applied deep learning approach for estimating soybean relative maturity from UAV imagery to aid plant breeding decisions. *Machine Learning with Applications*, *7*, 100233. <https://doi.org/10.1016/j.mlwa.2021.100233>
- Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Random forest for genomic prediction. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 633–681). Springer. [https://doi.org/10.1007/978-3-030-89010-0\\_15](https://doi.org/10.1007/978-3-030-89010-0_15)
- Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Barrón-López, J. A., Martini, J. W. R., Fajardo-Flores, S. B., Gaytan-Lugo, L. S., Santana-Mancilla, P. C., & Crossa, J. (2021). A review of deep learning applications for genomic selection. *BMC Genomics*, *22*(1), Article 19. <https://doi.org/10.1186/S12864-020-07319-X/FIGURES/6>
- Morales, N., Anche, M. T., Kaczmar, N. S., Lepak, N., Ni, P., Romay, M. C., Santantonio, N., Buckler, E. S., Gore, M. A., Mueller, L. A., & Robbins, K. R. (2024). Spatio-temporal modeling of high-throughput multispectral aerial images improves agronomic trait genomic prediction in hybrid maize. *Genetics*, *227*(1), iyae037. <https://doi.org/10.1093/genetics/iyae037>
- Moreira, F. F., Oliveira, H. R., Volenec, J. J., Rainey, K. M., & Brito, L. F. (2020). Integrating high-throughput phenotyping and statistical genomic methods to genetically improve longitudinal traits in crops. *Frontiers in Plant Science*, *11*, 681. <https://doi.org/10.3389/fpls.2020.00681>
- Muleta, K. T., Pressoir, G., & Morris, G. P. (2019). Optimizing genomic selection for a sorghum breeding program in Haiti: A simulation study. *G3: Genes/Genomes/Genetics*, *9*(2), 391–401. <https://doi.org/10.1534/G3.118.200932>
- Nascimento, M., Nascimento, A. C. C., Azevedo, C. F., Oliveira, A. C. B. D., Caixeta, E. T., & Jarquin, D. (2024). Enhancing genomic prediction with stacking ensemble learning in Arabica coffee. *Frontiers in Plant Science*, *15*, 1373318. <https://doi.org/10.3389/fpls.2024.1373318>
- Negus, K. L., Li, X., Welch, S. M., & Yu, J. (2024). The role of artificial intelligence in crop improvement. *Advances in Agronomy*, *184*, 1–66. <https://doi.org/10.1016/BS.AGRON.2023.11.001>
- Nguyen, C., Sagan, V., Skobalski, J., & Severo, J. I. (2023). Early detection of wheat yellow rust disease and its impact on terminal yield with multi-spectral UAV-imagery. *Remote Sensing*, *15*(13), 3301. <https://doi.org/10.3390/RS15133301>
- Ogutu, J. O., Piepho, H. P., & Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proceedings*, *5*(Suppl 3), Article S11. <https://doi.org/10.1186/1753-6561-5-S3-S11>
- Olivoto, T. (2022). Lights, camera, pliman! An R package for plant image analysis. *Methods in Ecology and Evolution*, *13*(4), 789–798. <https://doi.org/10.1111/2041-210X.13803>
- Pérez, O., Diers, B., & Martin, N. (2024). Maturity prediction in soybean breeding using aerial images and the random forest machine learning algorithm. *Remote Sensing*, *16*(23), 4343. <https://doi.org/10.3390/RS16234343>
- Pérez, P., & De Los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, *198*(2), 483–495. <https://doi.org/10.1534/GENETICS.114.164442>
- Pérez-Valencia, D. M., Rodríguez-Álvarez, M. X., Boer, M. P., Kronenberg, L., Hund, A., Cabrera-Bosquet, L., Millet, E. J., & Eeuwijk, F. A. V. (2022). A two-stage approach for the spatio-temporal analysis of high-throughput phenotyping data. *Scientific Reports*, *12*(1), Article 3177. <https://doi.org/10.1038/s41598-022-06935-9>
- Piepho, H.-P. (1998). Methods for comparing the yield stability of cropping systems. *Journal of Agronomy and Crop Science*, *180*(4), 193–213. <https://doi.org/10.1111/j.1439-037X.1998.tb00526.x>
- Pineda, M., Barón, M., & Pérez-Bueno, M. L. (2020). Thermal imaging for plant stress detection and phenotyping. *Remote Sensing*, *13*(1), 68. <https://doi.org/10.3390/rs13010068>

- Raj, R., Walker, J. P., Vinod, V., Pingale, R., Naik, B., & Jagarlapudi, A. (2021). Leaf water content estimation using top-of-canopy airborne hyperspectral data. *International Journal of Applied Earth Observation and Geoinformation*, *102*, 102393. <https://doi.org/10.1016/J.JAG.2021.102393>
- Reddy, S. S., Singh, G. M., Kumar, U., Bhati, P., Vishwakarma, M., Navathe, S., Yashavanthakumar, K. J., Mishra, V. K., Sharma, S., & Joshi, A. K. (2024). Spatio-temporal evaluation of drought adaptation in wheat revealed NDVI and MTSI as powerful tools for selecting tolerant genotypes. *Field Crops Research*, *311*, 109367. <https://doi.org/10.1016/J.FCR.2024.109367>
- Resende, R. T. (2024). Balancing genomic selection efforts for allogamous plant breeding programs. *Journal of Crop Science and Biotechnology*, *27*(3), 349–358. <https://doi.org/10.1007/S12892-024-00234-7>
- Resende, R. T., Hickey, L., Amaral, C. H., Peixoto, L. L., Marcatti, G. E., & Xu, Y. (2024). Satellite-enabled enviromics to enhance crop improvement. *Molecular Plant*, *17*(6), 848–866. <https://doi.org/10.1016/J.MOLP.2024.04.005>
- Resende, R. T., Piepho, H. P., Rosa, G. J. M., Silva-Junior, O. B., e Silva, F. F., de Resende, M. D. V., & Grattapaglia, D. (2021). Enviromics in breeding: Applications and perspectives on envirotypic-assisted selection. *Theoretical and Applied Genetics*, *134*(1), 95–112. <https://doi.org/10.1007/S00122-020-03684-Z>
- Resende, R. T., Xavier, A., Silva, P. I. T., Resende, M. P. M., Jarquin, D., & Marcatti, G. E. (2025). GIS-based G×E modeling of maize hybrids through enviromic markers engineering. *New Phytologist*, *245*(1), 102–116. <https://doi.org/10.1111/NPH.19951>
- Rincint, R., Kuhn, E., Monod, H., Oury, F. X., Rousset, M., Allard, V., & Le Gouis, J. (2017). Optimization of multi-environment trials for genomic selection based on crop models. *Theoretical and Applied Genetics*, *130*(8), 1735–1752. <https://doi.org/10.1007/S00122-017-2922-4>
- Rincint, R., Malosetti, M., Ababaei, B., Touzy, G., Mini, A., Bogard, M., Martre, P., Le Gouis, J., & van Eeuwijk, F. (2019). Using crop growth model stress covariates and AMMI decomposition to better predict genotype-by-environment interactions. *Theoretical and Applied Genetics*, *132*(12), 3399–3411. <https://doi.org/10.1007/S00122-019-03432-Y>
- Robert, P., Brault, C., Rincint, R., & Segura, V. (2022). Phenomic selection: A new and efficient alternative to genomic selection. In N. Ahmadi & J. Bartholomé (Eds.), *Genomic prediction of complex traits: Methods and protocols*, (Vol. 2467, pp. 397–420). Methods in molecular biology. Humana Press Inc. [https://doi.org/10.1007/978-1-0716-2205-6\\_14](https://doi.org/10.1007/978-1-0716-2205-6_14)
- Rogers, A. R., & Holland, J. B. (2022). Environment-specific genomic prediction ability in maize using environmental covariates depends on environmental similarity to training data. *G3 Genes/Genomes/Genetics*, *12*, jkab440. <https://doi.org/10.1093/g3journal/jkab440>
- Roychowdhury, R., Das, S. P., Gupta, A., Parihar, P., Chandrasekhar, K., Sarker, U., Kumar, A., Ramrao, D. P., & Sudhakar, C. (2023). Multi-omics pipeline and omics-integration approach to decipher plant's abiotic stress tolerance responses. *Genes*, *14*(6), 1281. <https://doi.org/10.3390/GENES14061281>
- Runcie, D., & Cheng, H. (2019). Pitfalls and remedies for cross validation with multi-trait genomic prediction methods. *G3 Genes/Genomes/Genetics*, *9*(11), 3727–3741. <https://doi.org/10.1534/G3.119.400598>
- Runcie, D., Qu, J., Cheng, H., & Crawford, L. (2021). MegaLMM: Mega-scale linear mixed models for genomic predictions with thousands of traits. *Genome Biology*, *22*(1), Article 213. <https://doi.org/10.1186/S13059-021-02416-W>
- Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., Reynolds, M., & Singh, R. (2016). Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3 Genes/Genomes/Genetics*, *6*(9), 2799–2808. <https://doi.org/10.1534/G3.116.032888>
- Saavedra, L. M., Caixeta, E. T., Barka, G. D., Borém, A., Zambolim, L., Nascimento, M., Cruz, C. D., Oliveira, A. C. B. D., & Pereira, A. A. (2023). Marker-assisted recurrent selection for pyramiding leaf rust and coffee berry disease resistance alleles in *Coffea arabica* L. *Genes*, *14*(1), 189. <https://doi.org/10.3390/genes14010189>
- Sabadin, F., DoVale, J. C., Platten, J. D., & Fritsche-Neto, R. (2022). Optimizing self-pollinated crop breeding employing genomic selection: From schemes to updating training sets. *Frontiers in Plant Science*, *13*, 935885. <https://doi.org/10.3389/FPLS.2022.935885>
- Schaeffer, L. R. (2004). Application of random regression models in animal breeding. *Livestock Production Science*, *86*(1–3), 35–45. [https://doi.org/10.1016/S0301-6226\(03\)00151-9](https://doi.org/10.1016/S0301-6226(03)00151-9)
- Simiqueli, G. F., Resende, R. T., Takahashi, E. K., de Sousa, J. E., & Grattapaglia, D. (2023). Realized genomic selection across generations in a reciprocal recurrent selection breeding program of *Eucalyptus* hybrids. *Frontiers in Plant Science*, *14*, 1252504. <https://doi.org/10.3389/FPLS.2023.1252504>
- Singh, D., Wang, X., Kumar, U., Gao, L., Noor, M., Imtiaz, M., Singh, R. P., & Poland, J. (2019). High-throughput phenotyping enabled genetic dissection of crop lodging in wheat. *Frontiers in Plant Science*, *10*, 394. <https://doi.org/10.3389/FPLS.2019.00394>
- Stöckle, C. O., Donatelli, M., & Nelson, R. (2003). CropSyst, a cropping systems simulation model. *European Journal of Agronomy*, *18*(3–4), 289–307. [https://doi.org/10.1016/S1161-0301\(02\)00109-0](https://doi.org/10.1016/S1161-0301(02)00109-0)
- Thapa, S., Gill, H. S., Halder, J., Rana, A., Ali, S., Maimaitijiang, M., Gill, U., Bernardo, A., St Amand, P., Bai, G., & Sehgal, S. K. (2024). Integrating genomics, phenomics, and deep learning improves the predictive ability for Fusarium head blight-related traits in winter wheat. *The Plant Genome*, *17*(3), e20470. <https://doi.org/10.1002/tpg2.20470>
- Trevisan, B. A., Junqueira, V. S., Florêncio, B. D. M., Coelho, A. S. G., Marcatti, G. E., & Resende, R. T. (2025). A framework for building enviromics matrices in mixed models. *arXiv*. <https://doi.org/10.48550/arXiv.2501.04147>
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. <https://doi.org/10.3168/JDS.2007-0980>
- Verbyla, A. (2023). On two-stage analysis of multi-environment trials. *Euphytica*, *219*(11), Article 121. <https://doi.org/10.1007/S10681-023-03248-4>
- Vieira, R. A., Nogueira, A. P. O., & Fritsche-Neto, R. (2025). Optimizing the selection of quantitative traits in plant breeding using simulation. *Frontiers in Plant Science*, *16*, 1495662. <https://doi.org/10.3389/FPLS.2025.1495662>
- Volpato, L., Dobbels, A., Borem, A., & Lorenz, A. J. (2021). Optimization of temporal UAS-based imagery analysis to estimate plant maturity date for soybean breeding. *The Plant Phenome Journal*, *4*(1), e20018. <https://doi.org/10.1002/PPJ2.20018>

- Volpato, L., Wright, E. M., & Gomez, F. E. (2024). Drone-based digital phenotyping to evaluate relative maturity, stand count, and plant height in dry beans (*Phaseolus vulgaris* L.). *Plant Phenomics*, 6, 0278. <https://doi.org/10.34133/plantphenomics.0278>
- Voss-Fels, K. P., Cooper, M., & Hayes, B. J. (2018). Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics*, 132(3), 669–686. <https://doi.org/10.1007/S00122-018-3270-8>
- Wang, F., Feldmann, M. J., & Runcie, D. E. (2025). Don't benchmark phenomic prediction against genomic prediction accuracy. *bioRxiv*. <https://doi.org/10.1101/2025.01.09.632209>
- Wang, P., Lehti-Shiu, M. D., Lotreck, S., Segura Abá, K., Krysan, P. J., & Shiu, S. H. (2024). Prediction of plant complex traits via integration of multi-omics data. *Nature Communications*, 15(1), Article 6856. <https://doi.org/10.1038/s41467-024-50701-6>
- Wu, C., Luo, J., & Xiao, Y. (2024). Multi-omics assists genomic prediction of maize yield with machine learning approaches. *Molecular Breeding*, 44(2), Article 14. <https://doi.org/10.1007/S11032-024-01454-Z>
- Wu, P. Y., Stich, B., Weisweiler, M., Shrestha, A., Erban, A., Westhoff, P., & Inghelant, D. V. (2022). Improvement of prediction ability by integrating multi-omic datasets in barley. *BMC Genomics*, 23(1), Article 200. <https://doi.org/10.1186/S12864-022-08337-7>
- Xavier, A., Runcie, D., & Habier, D. (2025). Megavariate methods capture complex genotype-by-environment interactions. *Genetics*, 229(4), 179. <https://doi.org/10.1093/GENETICS/IYAE179>
- Xu, Y. (2016). Envirotyping for deciphering environmental impacts on crop plants. *Theoretical and Applied Genetics*, 129(4), 653–673. <https://doi.org/10.1007/S00122-016-2691-5>
- Xu, Y., Zhang, X., Li, H., Zheng, H., Zhang, J., Olsen, M. S., Varshney, R. K., Prasanna, B. M., & Qian, Q. (2022). Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. *Molecular Plant*, 15(11), 1664–1695. <https://doi.org/10.1016/J.MOLP.2022.09.001>
- Yang, Y., Saand, M. A., Huang, L., Abdelaal, W. B., Zhang, J., Wu, Y., Li, J., Sirohi, M. H., & Wang, F. (2021). Applications of multi-omics technologies for crop improvement. *Frontiers in Plant Science*, 12, 563953. <https://doi.org/10.3389/FPLS.2021.563953>
- Yassue, R. M., Galli, G., Borsato, R., Cheng, H., Morota, G., & Fritsche-Neto, R. (2022). A low-cost greenhouse-based high-throughput phenotyping platform for genetic studies: A case study in maize under inoculation with plant growth-promoting bacteria. *The Plant Phenome Journal*, 5(1), e20043. <https://doi.org/10.1002/PPJ2.20043>
- Yassue, R. M., Galli, G., Fritsche-Neto, R., & Morota, G. (2022). Classification of plant growth-promoting bacteria inoculation status and prediction of growth-related traits in tropical maize using hyperspectral image and genomic data. *Crop Science*, 63, 88–100. <https://doi.org/10.1002/CSC2.20836>
- Zang, J., Jin, S., Zhang, S., Li, Q., Mu, Y., Li, Z., Li, S., Wang, X., Su, Y., & Jiang, D. (2023). Field-measured canopy height may not be as accurate and heritable as believed: Evidence from advanced 3D sensing. *Plant Methods*, 19(39), Article 39. <https://doi.org/10.1186/S13007-023-01012-2>
- Zhang, L., Niu, Y., Zhang, H., Han, W., Li, G., Tang, J., & Peng, X. (2019). Maize canopy temperature extracted from UAV thermal and RGB imagery and its application in water stress monitoring. *Frontiers in Plant Science*, 10, 1270. <https://doi.org/10.3389/FPLS.2019.01270>
- Zhang, M., Chen, T., Gu, X., Kuai, Y., Wang, C., Chen, D., & Zhao, C. (2023). UAV-borne hyperspectral estimation of nitrogen content in tobacco leaves based on ensemble learning methods. *Computers and Electronics in Agriculture*, 211, 108008. <https://doi.org/10.1016/J.COMPAG.2023.108008>

**How to cite this article:** Fritsche-Neto, R., Resende, R. T., Olivoto, T., Garcia-Abadillo, J., Nascimento, M., Bahia, M. A. M., Jarquin, D., & Vieira, R. A. (2025). Prediction-based breeding: Modern tools to optimize and reshape programs. *Crop Science*, 65, e70175. <https://doi.org/10.1002/csc2.70175>