Artificial Intelligence Review (2025) 58:260
https://doi.org/10.1007/s10462-025-11274-y

Machine learning after a decade: is it still a missing keystone in genomic-based plant breeding?

Mohsen Yoosefzadeh-Najafabadi[1] · Alencar Xavier[2] · Milad Eskandari[1] · Mohsen Hesami[1]

Accepted: 20 May 2025

Abstract

Plant breeding plays a crucial role in addressing the pressing challenges of food insecurity and global hunger, issues that are expected to worsen in the coming years. The development of effective plant breeding pipelines relies on a deep understanding and proficiency in various disciplines, such as phenomics and genomics. Leveraging the five G's - germplasm characterization, genome assembly, genomic breeding, gene function identification, and gene editing - can significantly boost the pace of crop improvement initiatives. In the past decade, the integration of machine learning (ML) algorithms into the five G's has gathered increasing attention for their abilities to integrate diverse omics and biological datasets to create precise breeding predictive models. Despite the promise of ML in advancing genomic-assisted breeding, there remains a critical question regarding the extent to which ML can help genomic-assisted breeding and what are their true potentials and efficacies compared to conventional methods. This review evaluates ML's role in genomics-based plant breeding, highlighting its strengths in prediction accuracy and breeding efficiency but also addressing challenges such as data biases and implementation barriers. It explores applications for improving crop resilience and productivity through the integration of multi-omics data and addressing data biases. Ultimately, the review underscores ML's potential to transform genomics-based plant breeding while identifying gaps and future research opportunities.

Keywords Data-driven breeding · CRISPR optimization · Five G's · Genomics · Omics integration

✉ Mohsen Yoosefzadeh-Najafabadi
myoosefz@uoguelph.ca

[1] Department of Plant Agriculture, University of Guelph, Guelph, ON N1G 2W1, Canada

[2] Corteva Agrisciences, Johnston, IA, USA

Published online: 04 June 2025

1 Introduction

Plant breeding, the genetic improvement of plant performance, is an essential tool to address food insecurity and global hunger resulting from the projected shortage of food in the near future (Hong et al. 2022; Voss-Fels et al. 2019). Since plant improvement is influenced by various intrinsic and extrinsic factors, successful plant breeding pipelines require expertise and knowledge in diverse areas such as omics (e.g., genomics, phenomics, transcriptomics, metabolomics, etc.) and the so-called "breeder's eye" to make informed decisions (Lenaerts et al. 2019; Voss-Fels et al. 2019).

Varshney et al. (2020) proposed "five G's" that can create a significant impact on crop improvement in plant breeding: germplasm characterization, genome assembly, gene (marker) assisted breeding, genomic prediction, and gene editing (Fig. 1). Germplasm characterization is the process of identifying and cataloging genetic diversity in a breeding population (Emanuelli et al. 2013). Genome assembly is the process of using sequencing data to assemble the entire genome (Jiao and Schneeberger 2017). Mapping quantitative trait loci (QTL) within a population or marker-trait associations (MTAs) with a given trait using an association panel is also considered as a powerful tool for marker-assisted breeding using genetic information (Yoosefzadeh Najafabadi et al. 2023). Likewise, the use of whole genome information across populations to identify the genetic architecture of desirable traits and predict optimal lines/crosses for achieving an ideal genotype is another important

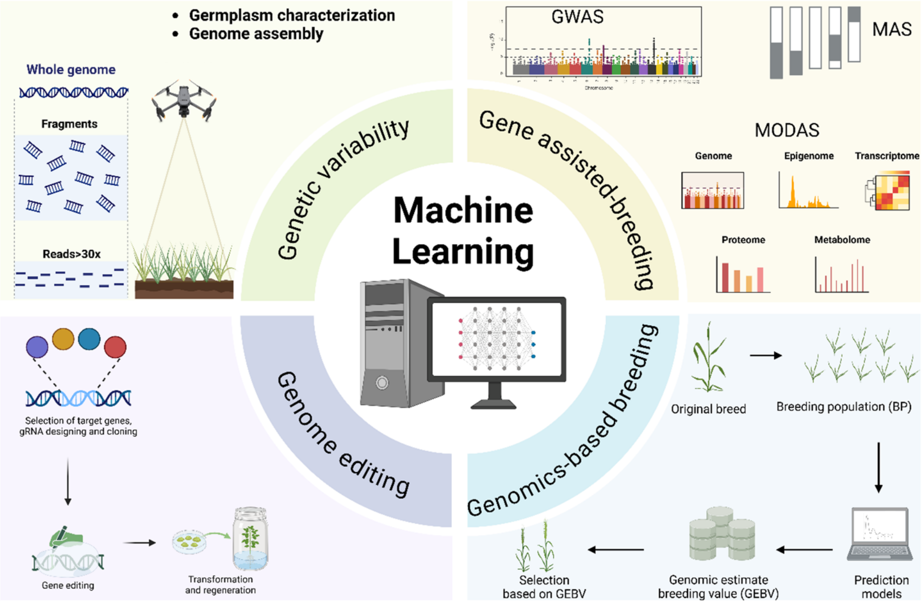

Fig. 1 Schematic illustration of the integration of ML algorithms across various "G's" in genetics, including genomics, genome editing, genetic variability, gene-assisted breeding, and genomics-based breeding. This figure highlights key processes such as whole genome sequencing, genome assembly, germplasm characterization, Genome-wide association studies (GWAS), MAS, MODAS, transcriptome analysis, epigenome, proteome, metabolome, selection of target genes, gene editing, transformation and regeneration, and genomic estimated breeding values (GEBV) with predictive modeling. This figure was created by BioRender.com

component in a successful breeding program (Varshney et al. 2020). Finally, the genetic information derived from different G's has enabled gene editing to become feasible for breeding at scale (Hickey et al. 2019).

There is no doubt that knowledge of five G's can greatly increase plant breeding efficiency through accelerating the development of new varieties with improved genetic performance. However, making the best use of the data obtained from genomics studies requires a broad range of sophisticated analytical tools and algorithms (Yoosefzadeh-Najafabadi et al. 2022). Machine learning (ML) algorithms, for example, have been increasingly deployed in plant breeding to handle various sources of data and understand the biological aspects of a phenomenon in greater detail (Y. Xu et al. 2022a, b). Modern algorithms hold great promise in plant breeding for enabling the integration of different omics and biological knowledge into accurate predictive models and, subsequently, implemented in different components of novel breeding schemes (Yan and Wang 2022). Conversely, the use of ML algorithms has also been subject to scrutiny, as these have shown questionable performance in a multitude of tasks and breeding scenarios (Y. Xu et al. 2022a, b; Yan and Wang 2022).

This review covers topics concerning the use of ML in the plant breeding community, particularly, "To what extent have ML algorithms been effective in assisting genomic-based plant breeding?" We focused on the following: 1) revisiting analytical methods involved in five important G's in plant breeding area, 2) providing an overview of efforts made to accelerate the rate of genetic gains through optimizing plant breeding, 3) assessing the strengths and limitations of ML algorithms in this context, and 4) identifying potential pathways for future research and studies. The ultimate goal of the current review is to assess the importance of ML algorithms in genomics-based plant breeding and to discuss the challenges and opportunities they could create in this area.

2 Revisiting genetic variability (germplasm characterization and genome assembly)

Exploration and exploitation of useful genetic variability are crucial for successful plant breeding program. Through germplasm characterization, genome assembly, and the incorporation of exotic haplotypes into elite pools, a wide array of potential characteristics can be unlocked and cultivated within a breeding population (McClung et al. 2020; Sinha et al. 2021). This technology breeders with adequate variation to enable the selection of varieties and hybrids with a combination of advantageous traits. A multitude of statistical methods have been developed to evaluate genetic variability in populations, envisioning to describe and explore aspects of population structure and the evolutionary history of a given population (Zhu et al. 2008).

For visualization purposes, principal component analysis (PCA) is the most common dimensionality reduction technique used in population structure analysis (Table 1). It is used to compress the SNP information into linearly orthogonal latent spaces referred to as principal components, which are ordered by the amount of variation recovered from the whole genomic dataset (Delicado 2011). Let M describe a n × p matrix of gene content, with n row (observations), p columns (parameters), and where the m_{ij} cell is filled with 0–1-2 representing three possible genotypes (aa-Aa-AA) for the i^{th} individual and j^{th} marker, in the diploid case. Principal components are extracted from the single-value decomposition of M, thus:

$$\mathbf{M} = \mathbf{UDV}\prime \tag{1}$$

Where U is n × n, the matrix of eigenvectors, D is a diagonal matrix with n eigenvalues, and V is the rotation matrix with dimension p × n. The principal components correspond to the set of eigenvectors with the largest eigenvalues. PCAs are orthogonal because $\mathrm{U}\prime\,\mathrm{U} = \mathrm{I}$, $\mathrm{V}\prime\,\mathrm{V} = \mathrm{I}$, and the amount of variance explained by the i^{th} components is inferred from the eigenvalues simply as $\frac{\mathrm{d}_\mathrm{i}^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{D}}$.

The use of PCs for modelling genomic relationships facilitates the computation of a large number of marker effects in linear models without the explicit modelling of single-nucleotide polymorphisms (SNPs), which is often unfeasible when datasets contain sequence data ($\mathrm{p} \geq \mathrm{n}$). Let:

$$\mathbf{Q} = \mathbf{UD} \tag{2}$$

where Q stands for a matrix representing transformed genomic data. So, a linear model fitting phenotypes (y) by an intercept ($\upmu$), the makers (M) with effect $\upbeta$, and residuals:

$$\mathbf{y} = \boldsymbol{\upmu} + \mathbf{M}\boldsymbol{\upbeta} + \mathbf{e} \tag{3}$$

can be reparametrized as:

$$\mathbf{y} = \boldsymbol{\upmu} + \mathbf{Q}\boldsymbol{\alpha} + \mathbf{e} \tag{4}$$

allowing for modelling a much smaller number of parameters (p) that will not exceed the number of observations (n). Subsequently, the marker effects can be recovered as $\upbeta = \mathrm{V}\alpha$. The PCA can also be directly computed from the genomic relationship matrix (G), defined according to VanRaden (2008), which uses the eigenvalue decomposition (EVD, $\mathrm{G} = \mathrm{UD}^2\mathrm{U}\prime$), which is, in this case, equivalent to the SVD of M, and any j^{th} eigenpairs (u_j, d_{jj}) have the property of $\left(\mathrm{G} - \mathrm{d}_{\mathrm{jj}}^2\,\mathrm{I}\right)\,\mathrm{u}_\mathrm{j} = 0$ renders G a ranking-deficient matrix under subtraction. The genomic relationship matrix (G) is defined as a standardized cross-product of the marker matrix (M), accounting for allele frequencies.

However, PCA is a linear technique and is not always suitable for non-linear relationships with the goal of identifying sub-populations or calculating marker effects (López-Cortés et al. 2020). Moreover, it is sensitive to the scale of the variables in the dataset and can be affected by outliers, meaning that its application to genomic information, where all markers display the same scale, is not extrapolatable to environmental covariates, expression data, and mixed sources of information unless the parameters (or features) are normalized beforehand (López-Cortés et al. 2020).

Multidimensional scaling (MDS) is another dimensionality reduction technique used to represent the relationships among a large set of similarity or dissimilarity data in a smaller set of dimensions while preserving the relative distances among data points (Hout et al. 2013). MDS involves the use of EVD of the row- and column-centered Euclidean function of m to calculate the Euclidean distance matrix (E) and, subsequently, the principal coordinates (Table 1). This approach takes into account the relative distances between each pair of individuals based on its similarities measured by their Euclidean distance (Hout et

Table 1 Comparison of dimensionality reduction techniques for genetic variability analysis

Method	Methodology	Advantages	Limitations	Computational Complexity	Handling Non-Linear Relationships	Sensitivity to Outliers	Suitability for Large Datasets
PCA	Linear dimensionality reduction using singular value decomposition (SVD) of the genetic data matrix or eigenvalue decomposition of the genomic relationship matrix. Compresses SNP data into orthogonal principal components.	- Computationally efficient for linear relationships. - Widely used for population structure analysis. -Facilitates modelling genomic relationships without explicit SNP modeming.	- Limited to linear relationships. - Sensitive to variable scaling and outliers. - Not suitable for environmental covariates or mixed data without normalization.	Moderate; scales well with moderate-sized datasets but can be intensive for very large p (markers).	Poor; assumes linear relationships.	High; outliers can skew principal components.	Suitable for moderate datasets but less effective for complex, non-linear structures.
MDS	Non-parametric dimensionality reduction preserving Euclidean distances between individuals. Uses eigenvalue decomposition of a cantered Euclidean distance matrix.	- Captures relative distances effectively. - Suitable for identifying homogeneous clusters. - Non-parametric, flexible for various data types.	- Computationally intensive with large datasets. - Struggles with missing data, requiring pairwise distance calculations. - Less effective for complex population structures.	High; pairwise distance calculations increase complexity, especially with missing data.	Moderate; can capture some non-linear relationships via distance metrics.	Moderate; depends on distance metric robustness.	Less suitable for very large datasets due to computational demands.
DAPC	Combines PCA with discriminant analysis to identify clusters and quantify variation. Reduces high-dimensional data to components capturing group differences.	- Effective for identifying population subgroups. - Handles missing data by using mean allele frequencies. - Computationally efficient compared to Bayesian methods.	- Limited to linear relationships. - Sensitive to outliers, which can skew results. - Not suitable for complex population structures reducible to few components.	Moderate; more intensive than PCA due to discriminant analysis but efficient for moderate datasets.	Poor; relies on linear combinations of variables.	High; outliers can affect discriminant functions.	Suitable for moderate datasets with clear group structures.
AE	Unsupervised neural network reducing data to a lower-dimensional latent space by minimizing reconstruction error. Uses multiple layers with weight matrices.	- Captures complex patterns in data. - Can handle multi-omics data effectively. - Flexible for various data types with proper tuning.	- Requires large datasets for training. - Computationally intensive, needing specialized hardware. - Prone to overfitting and difficult to interpret.	High; training neural networks is resource-intensive, especially for deep architectures.	Good; excels at capturing non-linear relationships.	Moderate; depends on training data and regularization.	Requires large datasets and significant computational resources.

Table 1 (continued)

Method	Methodology	Advantages	Limitations	Computational Complexity	Handling Non-Linear Relationships	Sensitivity to Outliers	Suitability for Large Datasets
VAE	Unsupervised neural network combining generative and discriminative models with variational inference. Optimizes a lower-dimensional representation of data.	- Captures underlying data structure effectively. - Suitable for multi-omics and complex datasets. - Robust to some extent due to probabilistic framework.	- Requires large datasets for effective training. - Computationally intensive, requiring specialized hardware. - Challenging to interpret and tune.	High; similar to AE but with additional complexity from variational inference.	Excellent; designed for complex, non-linear relationships.	Moderate; probabilistic nature mitigates some outlier effects.	Best for large datasets with sufficient computational power.

al. 2013) as $\mathrm{E_{ii\prime}} = (\mathrm{m_i} - \mathrm{m_{i\prime}})^2$. For using Euclidean distance, MDS is a non-parametric method that can be used to identify the presence of homogeneous clusters in a population (Borg et al. 2018; Hout et al. 2013). However, MDS can be computationally intensive and time-consuming under the presence of missing data, as that requires the explicit pairwise computation of Euclidean distance instead of the use of cross-products (Borg et al. 2018).

To refine PCA and MDS drawbacks, discriminant analysis of principal components (DAPC) was introduced to identify and quantify the sources of variation in a dataset (Grünwald et al. 2010). DAPC is a multivariate analysis technique that combines PCA and discriminant analysis to explore the relationship between population structure and genetic variation (Grünwald et al. 2010). DAPC is employed to identify clusters of individuals that are more likely belong to a given population. As such, it can be used to assess population structure and genetic diversity (Table 1). Using DAPC, high-dimensional data can be reduced to a lower number of components that capture most of the variation in the data (Jombart and Collins 2015). It can be used to identify subgroups within a population, which can be useful in making decisions about how best to allocate resources, compare the populations of different groups, and identify significant differences between them (Grünwald et al. 2010). In cases of missing data, the mean frequency of the corresponding allele is used to replace the missing values in order to avoid skewed between-group differences.

Consider a genetic data matrix, X, with n individuals represented in rows and p columns representing the relative frequencies of alleles. Each column, denoted as X_j, represents a specific allele, such as B_1, B_2, or B_3. For instance, a homozygous genotype of B_1B_1 would be denoted as [1, 0, 0], while a heterozygous genotype of B_2B_3 would be denoted as [0, 0.5, 0.5]. To simplify the analysis, it should be assumed that each column in X is centred around a mean of zero (Jombart et al. 2010). The purpose of classical (linear) discriminant analysis is to identify linear combinations of alleles that can effectively distinguish between groups as follows:

$$\boldsymbol{f}\left(\boldsymbol{v}\right) = \sum\nolimits_{\boldsymbol{j}=1}^{\boldsymbol{p}} \boldsymbol{x}^{\boldsymbol{j}}\boldsymbol{v}_{\boldsymbol{j}} = \boldsymbol{x}\boldsymbol{v} \tag{5}$$

Where v is the vector of p alleles, considered as discriminant coefficients. The ultimate aim of DAPC is to select v in the case that f(xv) is maximum. DAPC is relatively fast and computationally efficient compared to other methods of genetic variability analysis (Table 1) (Jombart and Collins 2015). However, DAPC is limited to linear relationships between variables and does not take into account non-linear relationships (Grünwald et al. 2010). It can produce misleading results if there are a large number of outliers, as it is sensitive to them (Grünwald et al. 2010; Jombart and Collins 2015). Moreover, it is not suitable for situations where the population structure is complex and cannot be reduced to a few simple components (Jombart and Collins 2015).

Subsequently, the genetic distance matrix (GDM) analysis was introduced to quantify the divergence and similarity between different genotypes at a specific set of loci (Yang et al. 2021). The genetic distance between pairs of individuals is calculated by combining the differences between the alleles at each locus (Fitzpatrick and Keller 2015). The sum of these differences is then divided by the number of loci to give an overall measure of genetic distance (Yang et al. 2021). These measures can then be utilized to create a genetic distance matrix, which provides a visual representation of the genetic relatedness between individuals in the population (Yang et al. 2021). Considering A_i as the allele at locus i for individual A and B_i as the allele at locus i for individual B, the difference between the alleles can be denoted as $D_i =| A_i - B_i|$. Therefore, the sum of allele differences across all loci for a pair of individuals (A and B) can be calculated as $D_{(A \& B)} = \sum_{n}^{i=1} D_i$, where n is the total number of loci considered. The genetic distance between individual A and individual B can be calculated by dividing the sum of differences by the total number of loci as $GD_{(A \& B)} = \frac{D_{(A \& B)}}{n}$. GDM has the advantage of identifying population structure more accurately than other methods, such as PCA, detecting genetic relationships between different populations, which can be useful for conservation efforts, identifying and comparing genetic diversity among populations, and detecting outlier individuals that may represent new evolutionary lineages (Table 1) (Yang et al. 2021). However, it can be difficult to interpret the results of GDM analysis, as it does not provide an explicit definition of the population structure (Fitzpatrick and Keller 2015). It is computationally intensive, which can limit its use for large datasets, and it is only able to detect differences between populations and not the underlying reasons for those differences (Yang et al. 2021). Bayesian clustering methods calculate the probability of a data point belonging to a specific cluster based on prior knowledge about the data (Quintana and Iglesias 2003). In genetic variability, prior knowledge is often based on genetic data, environmental data, or other demographic information. The probability of a data point belonging to a specific cluster is then calculated based on prior knowledge, and the data point is assigned to the cluster with the highest probability (François and Durand 2010).

Bayesian clustering methods, such as Bayesian Gaussian Mixture Model (BGMM) or Dirichlet Process Mixture Model (DPMM), are known as other powerful methods for identifying population subgroups and analyzing their relationships (Table 1). These methods are particularly useful when dealing with complex genomics datasets that have varying sample sizes and a wide range of population genetic parameters (François and Durand 2010). As an example, in BGMM each data point (x_i) is assumed to be generated from one of (K) Gaussian distributions, each characterized by a mean (μ_k) and a covariance matrix (Σ_k), where (k = 1,…, K). The probability of data point (x_i) belonging to cluster (k) is given by the Gaussian density function:

$$\mathbf{p}_{(\mathbf{x}_i \mid \boldsymbol{\mu}_k,\ \boldsymbol{\Sigma}_k)} = 2\boldsymbol{\pi}^{-\frac{d}{2}} \times \boldsymbol{\Sigma}_{\mathbf{k}}^{\frac{1}{2}} \mathbf{e}^{\left[-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^t \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)\right]} \tag{6}$$

Where d stands for the dimensionality of the data and t stands for the transpose of a matrix or vector.

Prior distributions over the cluster parameters, such as the mean and covariance of each cluster can be defined, accordingly. These priors can capture any assumptions about the data distribution. By combining the likelihood function with the prior distributions using Bayes' theorem, we can compute the posterior distribution over the clustering parameters (Table 1). This allows to infer the most likely clustering of the data and estimate the cluster parameters simultaneously. Bayesian clustering methods are also able to account for the uncertainty when assigning individuals to clusters, making them more robust than other clustering methods (Onogi et al. 2011). However, Bayesian clustering methods require a significant computing power, making them difficult to implement on large datasets (Onogi et al. 2011).

In addition, Bayesian clustering methods are sensitive to the prior information used, which can lead to inaccurate results if the priors are not properly specified (Onogi et al. 2011). In recent decades, several new analyzing approaches, such as autoencoders (AE) and variation autoencoders (VAE), have been introduced to better analyze the available genetic variability in a population (D. Wang and Gu 2018). VAE is a type of unsupervised neural network that employs a variational inference, which uses a combination of generative and discriminative models to identify the underlying structure of the data by optimizing a lower-dimensional representation of the data (Table 1) (Falck et al. 2021). This lower-dimensional representation can then be used to identify patterns and trends in the data, allowing for more efficient analysis of multi-omics data than PCA (Falck et al. 2021). Autoencoder (AE), as another commonly used method (Geleta et al. 2023), is fit to reproduce the original dataset under a reduced number of dimensions (a.k.a. latent spaces):

$$\mathbf{X} = \boldsymbol{\alpha}\ (\boldsymbol{\alpha}\ (\boldsymbol{\alpha}\ (\mathbf{X})\,\mathbf{W}_1)\,\mathbf{W}_2)\,\mathbf{W}_3 \tag{7}$$

Weights (W_1, W_2, W_3) are commonly estimated under the minimization of the squared error:

$$\mathbf{argmin}\left(\mathbf{X} - \widehat{\mathbf{X}}\right)^2 \tag{8}$$

The single-value decomposition is a linear, single-layer counterpart of the autoencoder.

To date, supervised strategies such as Markov Clustering (MCL) have been used to identify haplotypes in multi-scale germplasm networks (Yang et al. 2022). This approach enabled the dissection of the germplasm resources to clarify aspects such as pedigree relationship, identify the founder lines and detect genetic flow. However, the need for a robust, cost-effective, and automated method to characterize genotypes has become increasingly apparent. Recent advances in ML, and specifically the development of deep learning algorithms, have enabled the automation and optimization of the genotype's characterization process. These algorithms can analyze and learn complex patterns in large datasets and identify genotypes with greater accuracy, speed, and cost efficiency than traditional methods (Bogard et al. 2021). López-Cortés et al. (2020) investigated how ML clustering methods (K-means and hierarchical clustering) combined with non-linear and linear data reduction techniques (deep autoencoder and principal component analysis) could be used to identify population structure and assign individuals to clusters in maize inbred lines. Results indicated that hierarchical clustering with deep autoencoder-based preprocessing (DeepAE-HC) had the highest accuracy for individual assignments (96%). It also had higher accuracy than a Bayesian clustering method (López-Cortés et al. 2020). Deep learning-based dimension reduction combined with ML clustering methods can be used to identify genetically differentiated groups and assign individuals to subpopulations in genome-wide studies without prior genetic assumptions (López-Cortés et al. 2020). In another study, three convolutional neural networks (CNN) models, VGG16, Inception-V3, and NASNet were developed to accurately identify three species of *Cinnamomum osmophloeum* using leaf images as inputs (H.-W. Yang et al. 2019). Score fusion was then applied to further improve the performance of the CNN classifiers, resulting in a test accuracy of up to 96.7%, which indicated the potential of using deep learning in germplasm characterization (H.-W. Yang et al. 2019). However, the use of deep learning in germplasm characterization requires large amounts of data to be well trained. As a result, a large amount of data must be collected in order to develop a reliable deep learning model. In addition, deep learning models are computationally intensive and require specialized hardware to run. Additionally, these models can be challenging to interpret and may lead to incorrect conclusions about the data. Overfitting is also a concern with deep learning models, as it can lead to inaccurate results (Roelofs et al. 2019). As such, the selection of proper method totally depends on the type, the volume of datasets, prior knowledge existed for the dataset, and the amount of computational power available to analyse the genetic variability.

There are available packages in Python and R programming languages that can handle analyses related to ML in this context. For example, in Python, TensorFlow (Pang et al. 2020) is a powerful library that excels in managing deep learning algorithms, providing robust support for building complex models. In R, the Caret (Kuhn et al. 2020) and tidymodels (Kuhn and Silge 2022) packages offer comprehensive frameworks and tools for developing and implementing a wide variety of ML models, making them invaluable resources for data analysis and model training. In terms of preprocessing datasets, the AllInOne preprocessing package (Najafabadi et al. 2023) in R specifically addresses the needs of plant breeding programs by offering specialized tools for preprocessing phenomics data. This package streamlines data preparation, enabling researchers to focus on extracting meaningful insights and accelerating the breeding process. By leveraging these advanced packages, researchers and breeders can enhance their analytical capabilities, improve model accuracy, and foster innovation in the development of resilient and high-yielding crops.

3 Revisiting gene assisted breeding

The advent of molecular markers in the 1980 s and subsequent advancements in their types and abundance have provided plant breeders with powerful tools for identifying genomic fragments associated with target traits and selecting genotypes with improved genetic backgrounds using marker-assisted selection (MAS). For instance, in the development of maize hybrids, molecular markers have reduced the number of generations required from ten to three (Holland 2004; Ribaut et al. 2010; Stuber et al. 1992). Molecular markers are identified from the successful phenotype-to-genotype (P2G) association that can be accelerated using next-generation genomics and phenomics (Brown et al. 2021). Therefore, the discovery of associated genes with a trait of interest is essential for utilizing natural genetic variations through approaches such as marker-assisted selection MAS or genome editing to induce targeted mutations. Qualitative traits, to which MAS suits best, are defined through the association between one or a few markers closely linked to the causative QTL. Let the phenotype be a presence or absence of the trait, so that $\mathrm{y} \in \{0{,}1\}$. MAS' premise is $\mathrm{y} = \mathrm{f}(\mathrm{x}) + \mathrm{e}$, where y is the expect phenotype, $\mathrm{x} = \{0{,}1, 2\}$, representing $\{\ \mathrm{AA}, \mathrm{Aa}, \mathrm{aa}\}$, respectively, e is a residual or misclassification due to imperfect LD between marker and QTL, leading to decision function:

$$\boldsymbol{f}(\mathbf{x}) = \begin{cases} x = 0 \rightarrow \ \textit{Do not select} \\ x = 1 \rightarrow \ \textit{Do not select} \\ x = 2 \rightarrow \ \textit{Select} \end{cases} \tag{9}$$

The marker (x) used as proxy for the QTL of interest is chosen from a set of markers (X) such that the misclassifications are minimized $(\mathrm{x}|\mathrm{X} = \mathrm{argmin}(\mathrm{e}))$ or, equivalently $(\mathrm{x} \in \mathrm{X} = \mathrm{argmax}(\mathrm{LD}(\mathrm{x}, \mathrm{QTL})))$.

While the successful application of molecular markers has greatly enhanced genetic improvement for simple traits, the improvement of complex traits presents challenges, including the difficulty of finding durable marker-trait associations across diverse environments and genetic backgrounds (Bernardo 2016). Genome-wide association studies (GWAS), as a common genetic study, have faced limitations in dissecting complex, polygenic traits as it undermines genes with minor effect and neglects the interaction among associated genes (Harfouche et al. 2019). Miao et al. (2024) presented significant concerns about the validity of such associations and introduces a new statistical framework called Post-Prediction GWAS (POP-GWAS). POP-GWAS is designed to enhance the reliability of GWAS analyses on ML-imputed outcomes, providing accurate and powerful statistical inferences independent of the imputation quality or the ML algorithm used. Furthermore, As the phenotype in the P2G association is significantly influenced by different environmental and climate factors, it would be necessary to include the climate information in the selection process in order to improve the selection accuracy (Tao et al. 2022). Therefore, in a breeding program, we are facing with a multi dimensional, heterogeneous, noisy, and complex dataset which require sophisticated statistical analyses to overcome these challenges.

To address the challenge of high-dimensional data in agricultural genomics, a method called dimensionality reduction (DR) is employed to avoid the drawbacks associated with having a large number of variables. The Multi-Omics Data Association Studies (MODAS) toolbox utilizes various DR algorithms to analyze genomics and other omics datasets for complex traits in plants (Fig. 2). By leveraging DR methodologies, the analysis of multi-omics datasets (e.g., transcriptomics, epigenomics, metabolomics, and proteomics) becomes more feasible and impactful. These multi-omics approaches provide deeper insights into molecular variation within breeding lines, extending beyond the observable genetic variation (Fig. 2). They generate data that are closer to phenotype, bridging the genome-to-phenome gap and offering complementary marker systems to genetic markers (Fig. 2).

For genomics, MODAS employs a combination of Jaccard similarity coefficient, density-based clustering non-parametric algorithm (DBSCAN), and PCA algorithms to create a simplified representation called a “pseudo-genotype index” from tens of thousands of genomic blocks, effectively summarizing thousands of SNPs for more efficient analysis and mapping (Liu et al. 2022). Similarly, the dimensionality of complex traits is reduced using the nonnegative matrix factorization (NMF) algorithm, which identifies patterns in gene-gene interactions and removes redundancy by breaking down input variables into meta-data and meta-sample dimensions.

To demonstrate the concept of using the Jaccard similarity coefficient and PCA approaches, assume a dataset representing genotypes of two individuals for three different genomic blocks (A, B, C). Each individual’s genotype can be represented as a binary vector indicating the presence (1) or absence (0) of a particular variant in each block:

1. Individual 1: [1, 0, 1].
2. Individual 2: [0, 1, 1].

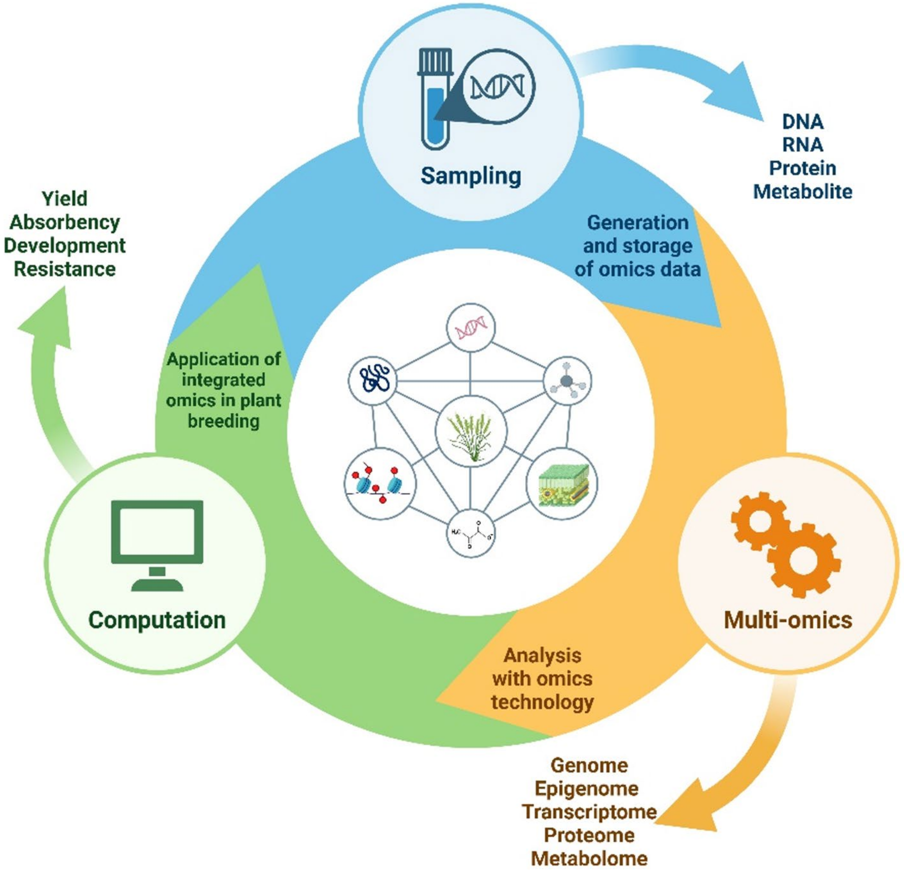

Fig. 2 The schematic diagram illustrates the integration of omics technologies in plant breeding. The process begins with sampling, leading to the generation and storage of omics data, including DNA, RNA, protein, and metabolite information. Multi-omics analysis encompasses the genome, epigenome, transcriptome, proteome, and metabolome, providing insights for plant development. Computational tools are then applied, focusing on yield, absorbency, development, and resistance, facilitating the application of integrated omics in enhancing plant breeding strategies. This figure was created by BioRender.com

For these individuals, the Jaccard similarity coefficient is defined as the size of the intersection (1) divided by the size of the union of two sets (0, 1). Therefore, the Jaccard similarity coefficient is equal to 0.5. afterward, PCA can be used to reduce the dimensionality of the dataset (as described in previous section). This approach helps in identifying key genetic factors and pathways associated with complex traits through GWAS analysis, leading to faster computation and resource savings while producing clear and interpretable outcomes.

4 Revisiting genomics - based breeding approaches

In recent years, the implementation of genomic selection (GS), which leverages all molecular markers simultaneously (Meuwissen et al. 2001) to identify the most promising genotypes, has been revolutionizing the field of cultivar development (Bernardo 2016).

4.1 Classical genomic models

Classical genomic prediction models, including ridge-regression best linear unbiased prediction (RR-BLUP), genomic best linear unbiased prediction (GBLUP), elastic net (EN), Bayes' theorem (Bayes A, B, and C) have long been regarded as a highly promising tool in the domains of plant and animal breeding. The primary objective of these models is to predict the phenotypic values of individuals within a breeding population, leveraging their genetic information. This enables breeders to effectively identify and select the most genetically enhanced individuals in earlier generations. In GBLUP, the equation for predicting the breeding value of a given genotype represent as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{10}$$

where y is the vector of phenotypes, X is the design matrix relating fixed effects (such as environment or other covariates) to the response variable, b is the vector of fixed effect coefficients, Z is genotype incidence matrix, and u is the vector of genomic estimated breeding values (GEBV) with $\mathrm{u} \sim \mathrm{N}(0, \mathrm{G}\sigma_{\mathrm{u}}^{2})$, where G is the genomic relationship matrix and $\sigma^{2}{}_{\mathrm{u}}$ is the additive genetic variance, e is the residual error term. The alternative parameterization, ridge regression (RR-BLUP), is as follows,

$$\mathbf{y} = \mathbf{Xb} + \mathbf{M}\boldsymbol{\beta} \ + \mathbf{e} \tag{11}$$

where M is the matrix of marker scores, β are the marker effects. Ridge regression solution for the marker effects is given by:

$$\boldsymbol{\beta} \ = (\mathbf{M}'\,\mathbf{M} + \boldsymbol{\lambda}\,\mathbf{I})^{-1}\mathbf{M}'\,(\mathbf{y} - \mathbf{Xb}) \tag{12}$$

where λ is the ridge regression parameter that controls the amount of shrinkage applied to the estimated genetic effects, estimated from cross-validation of variance components ($\lambda \ = \sigma_{\mathrm{e}}^{2}\sigma_{\beta}^{-2}$).

Elastic net regression constitutes a regularized regression method that combines the penalties of both Lasso, **L1**, and Ridge **L2**, regularization techniques, with squared regularization of residuals (Giglio and Brown 2018). Its primary purpose is to shrink or reduce the **β**-coefficients towards zero, while avoiding the risk of overfitting the data. The overall goal is to select a subset of features, molecular markers, that maximizes model accuracy (Zou and Hastie 2005). The EN algorithm attempts to solve the following optimization objective:

$$\boldsymbol{min}_{\boldsymbol{\beta}} \left[\frac{1}{2\boldsymbol{n}} (\boldsymbol{y} - \widehat{\boldsymbol{y}})' \ (\boldsymbol{y} - \widehat{\boldsymbol{y}}) + \boldsymbol{\lambda} \ \boldsymbol{\alpha} \ |\boldsymbol{\beta}\,| + \boldsymbol{\lambda} \ (1 - \boldsymbol{\alpha} \ ) \ \boldsymbol{\beta}\,'\,\boldsymbol{\beta} \right] \tag{13}$$

where **β** is the vector of marker coefficients, α balances ***L1*** and ***L2*** penalizations, and $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$ are the coefficients for the **L1** and **L2** regularization terms, respectively (Giglio and Brown 2018).

The objective function used in regularization includes a sum of squared errors, $(y-\widehat{y})'\,(y-\widehat{y})$, and coefficient regularization, $\lambda\ \alpha\ |\beta\,| + \lambda\ (1-\alpha\,)\ \beta\,'\,\beta$, that penalizes model complexity by decreasing the size of model coefficients (Giglio and Brown 2018; Zou and Hastie 2005). The first term encourages sparsity while the second term penalizes the size of coefficients (Zou and Hastie 2003). The elastic net algorithm can adjust the weights of these terms through the constants λ, λ1 and λ2, providing greater flexibility in the model compared to traditional **L1** and **L2** regularized regressions (Zou and Hastie 2003, 2005).

Ridge regression has Bayesian variations with adaptative regularization of coefficients, that is, each $\beta_{\,j}$ has its own $\lambda_{\,j}$. Bayes A is known as a three-stage hierarchical model (Meuwissen et al. 2001) where in the first stage, a normal regression is used; the second stage assigns a normal conditional prior to each of the marker effects, all of which have a null mean but a variance specific to each marker; and the third stage assigns the same scaled inverted chi-square distribution with known scale ($S_2\beta$) and degrees of freedom (ν) parameters to each of the variances. This mechanistic argument aims to capture the idea that all markers can differentially contribute to genetic variance depending on their effects, allelic frequencies, and linkage disequilibrium with causal variants (Meuwissen et al. 2001). The formulation of Bayes B detailed in Gianola (2013) and Habier et al. (2011) has an hierarchical prior, which can be deceptive because in reality, it assigns each marker the same marginal prior. This occurs due to the formulas used to calculate the mean and variance of a mixture, as explained in Gianola et al. (2006). To clarify, the mean of a mixture is calculated by taking the component means and weighting them by probabilities π and $1-\pi$. The variance is calculated by taking the weighted average of the component variances and adding additional variance to account for discrepancies among the component means (Gianola et al. 2006). Bayes C requires a prior assumption regarding the distribution of the effects for a fraction of SNPs (π) (Pong-Wong and Woolliams 2014). Different values for π were tested and the optimal values were chosen based on the performance of the model. The GEBV for the validation was calculated taking into account the scaled genotypes of the bulls and the posterior mean of the SNP effects (Pong-Wong and Woolliams 2014).

4.2 ML-based approaches

It is important to recognize that classical models encounter certain limitations, particularly in their predictive accuracy and scalability when applied to diverse or large populations. Classical genomic prediction models are typically based on linear regression models that assume a linear relationship between genetic markers and phenotypic values (Meuwissen et al. 2001; VanRaden 2008). However, this assumption is often violated in hands-on scenarios, where the genetic architecture of complex traits is often characterized by non-linear interactions among genes as well as environmental factors. As a result, classical models may have limited predictive power, leading to inaccurate predictions of phenotypic values. Significant improvements in computational power have provided a remarkable opportunity to overcome these limitations and fully harness the potential of genomic selection, enabling breeders to make informed decisions and accelerate cultivar development at an unprecedented pace.

As an example, Support Vector Machines (SVMs) are a robust category of supervised ML algorithms employed for tasks such as classification and regression (Musa 2013). A study by Zhao et al. (2020) utilized SVMs with various kernel functions and hyperparameters on eight genomic datasets from pigs and maize. Their findings indicated that SVMs surpassed other methods in terms of time and memory efficiency. Similarly, Griffel et al. (2018) conducted research applying SVMs with specific light wavelengths to detect a potato virus, achieving an impressive accuracy of 89.8%, which was notably higher than other techniques. Random Forest (RF), as another important ML algorithm, is commonly used for both classification and regression tasks (Belgiu and Drăguţ 2016). It creates multiple decision trees from random subsets of features in the dataset, and combines their predictions to achieve more accurate and stable results (Qi 2012). In one study, Parmley et al. (2019) demonstrated the effectiveness of RF in predicting soybean seed yield using hyperspectral reflectance data. They analyzed 292 different soybean varieties across six environments to understand the relationships between various traits and yields. By using RF, they found important in-season traits that could reduce the need for extensive end-of-season yield assessments. Another study by Yoosefzadeh-Najafabadi et al. (2021) compared RF with other machine learning methods like multi-layer perceptrons (MLP) and SVMs for the same task, finding that RF provided the highest accuracy at 84%. This study suggested that soybean breeders could use RF to efficiently identify high-yield genotypes early on by analyzing spectral data. While the use of ML methods offers the potential to model nonlinear interactions more effectively, empirical results have shown that these methods have not consistently provided a significant improvement in predictive accuracy over classical models such as GBLUP. This suggests that while ML methods have not fully replaced classical approaches, they hold promise and necessitate further research to understand how and when they can be most beneficially applied (Yoosefzadeh Najafabadi 2021).

4.3 Integration of omics and ML in GS

ML- based genomic prediction approaches, specially deep learning algorithms, offer opportunities to address current limitations by identifying of non-linear relationships between genetic markers and phenotypic traits through different layers (Azodi et al. 2019; Libbrecht and Noble 2015). Zingaretti et al. (2020) developed a data-driven approach called Environment-Phenotype Association (EPA) to incorporate high-throughput environmental data into genomic selection using ML algorithms. The results showed that integrating EPA as a dimensionality reduction strategy and connecting phenotypic and environmental-wide variation significantly improved the prediction accuracy of genotype-by-environment interactions (G × E) in wheat breeding, compared to conventional models. EPA acted as a “reinforcement learner” algorithm, uncovering the effect of seasonality and enhancing the forecasting of similarities between past and future trial sites, ultimately increasing the resolution of GP at the genotype-specific level. Integrating multi-omics data (e.g., genomics, transcriptomics, proteomics, and metabolomics) into genomic prediction models can also significantly improve the accuracy of predictions, and ML algorithm can help with this integration. These techniques can analyze large multi-dimensional datasets and identify complex relationships between different types of biological data and phenotypic traits, allowing for more accurate and robust predictions. In a recent study by Wang et al. (2023), a novel deep learning method called Deep Neural Network Genomic Prediction (DNNGP) is reported being designed for integrating multi-omics data in plant genomics. It is reported that DNNGP outperformed traditional linear models, and other ML algorithms, in terms of prediction accuracy and computation time. It demonstrated versatility in handling different types of data, including omics data, and showed potential for practical implementation in existing genomic selection platforms.

Although various ML algorithms have been employed to predict phenotypic traits, previous studies have predominantly focused on single-trait models, overlooking the relationships between the target trait and other related traits. A comparison of classical and ML-based GS methods, including their predictive accuracy and ability to handle non-linear relationships, is provided in Table 2. To overcome this limitation, Liang et al. (2023) proposed a multi-trait prediction model named MAK, which leverages multi-target ensemble regressor chains to incorporate multiple traits and automatically extract valuable information. The performance of MAK was assessed using real animal and plant datasets and compared against established methods such as GBLUP, BayesRR, and BayesB. The results demonstrated that MAK surpassed these benchmark methods in terms of prediction accuracy while exhibiting favorable computational efficiency. This research signifies the importance of considering multiple traits in prediction models and highlights the potential of the MAK framework in improving accuracy in various fields. In another study, Costa-Neto et al. (2022) investigated the use of multi-trait GS algorithm for predicting seven different end-use quality traits in soft white wheat breeding. The results demonstrated that multi-trait GS models outperformed single-trait models, showing 5.5% and 7.9% improvement in prediction accuracies for within-environment and across-location predictions, respectively. ML and deep learning algorithms performed better than conventional models for across-location predictions, with the highest improvement (35%) observed for flour protein content using the multi-trait

Table 2 Comparison of classical and ML-based GS methods

Method	Model Type	Key Assumptions	Predictive Accuracy	Non-Linear Handling
GBLUP	Linear	Linear marker-phenotype relationship	Moderate	Limited
RR-BLUP	Linear	Linear effects, equal variance	Moderate	Limited
Bayes A/B/C	Bayesian	Marker-specific variances	Moderate-High	Limited
Elastic Net	Regularized Linear	Linear, sparse effects	Moderate	Limited
SVM	ML (Kernel-Based)	Non-linear possible with kernels	High	Strong
Random Forest	ML (Tree-Based)	Non-linear, feature interactions	High	Strong
DNNGP	ML (Deep Learning)	Complex, non-linear patterns	High	Very Strong
MAK	ML (multi-trait)	Trait correlations, non-linear	High	Strong

GBLUP, Genomic Best Linear Unbiased Prediction; RR-BLUP, Ridge-Regression Best Linear Unbiased Prediction; SVM, Support Vector Machine; DNNGP, Deep Neural Network Genomic Prediction; MAK, Multi-Trait Ensemble Regressor Chains; ML, Machine Learning.

MLP model. This study highlights the potential of multi-trait-based ML-derived GS models to enhance prediction accuracy, accelerate the breeding cycle, and reduce costs in wheat breeding programs. As research in this area progresses, it becomes important to explore how multiple omics data can be utilized to refine predictions and improve plant breeding strategies (Fig. 2). Consequently, the role of omics technologies in GS and their integration with ML models is gaining attention for their potential to be used in breeding approaches.

As great complementary to GS, omics layers such as gene expression and metabolite concentrations integrate signals from multiple genetic loci. These layers often provide superior, predictive power for phenotypes compared to SNPs due to their more direct relationship with the phenotype. Genomic signatures, such as alleles and haplotypes, often become lost in the complexity of interacting omics layers, and the effect on a phenotype is diminished. Next-generation ML has the potential to revolutionize GS if the acquired omics data are appropriately utilized to leverage the capabilities of ML and if explainable ML approaches are pursued that build upon existing human knowledge to continually refine the models based on biologically confirmed results (Harfouche et al. 2019). The acquisition of extensive phenomics and genomics data, as well as the molecular layers between them, such as transcriptomics, proteomics, transcriptomics, and metabolomics, will create an environment where ML models can discover and explain intricate relationships; for instance, predicting how changes in water availability affect the expressions of genes involved in plant growth while influencing resistance to pests.

5 Optimizing genomic-based plant breeding

Optimization procedures are largely applicable to a wide variety of breeding decisions and operations beyond regression and classification problems. The availability of genomic information provides the necessary information to enable a more wholistic optimization of breeding schemes aiming higher genetic gains in the short and long term (Henryon et al. 2014). The design of a breeding program's blueprint is contingent to its breeding objective, factoring the economic and the genetic resources available (Goddard 1998). Therefore, the objective function consumed by the optimizer must factor in genetic gains per unit of time and cost (Heffner et al. 2010; Lorenz and Nice 2017). Evaluations and comparison of different breeding scenarios often involve (1) retrospective studies using cross-validation with previous year of data, (2) deterministic calculations to infer trends using parameters estimated from data, and (3) simulations to replicate breeding processes with data created in silico (Chapman 2008; Jahufer et al. 2021; Li et al. 2012). These methods are crucial in understanding and optimizing the genomic-based plant breeding pipeline as outlined in the objectives, specifically in revisiting analytical methods within the five important G's and setting a course for enhancing genetic gains.

5.1 Deterministic optimization

Deterministic accuracy is a metric suitable for the optimization of breeding trials, field designs, and training populations for genomic prediction (Morota 2017; Rincent et al. 2017; Schopp et al. 2017; Wientjes et al. 2020). Deterministic accuracy is based on the known properties of linear stochastic models, making it possible to infer the correlation between estimated and true values under the assumption that the true model and variance components are known, even though the true values are unknown. Predictive ability (PA) is defined as the correlation between estimated and true values as $\mathbf{PA} = \mathbf{cor}\,(\mathbf{u}, \widehat{\mathbf{u}})$ and accuracy (a) can be calculated as:

$$\mathbf{a} = \frac{\mathbf{PA}}{\sqrt{\mathbf{h}^2}} \tag{14}$$

Except for simulation studies, true values are generally unknown, and the correlation cannot be computed directly. However, the correlation can be estimated under the assumption that the parametrization of the statistical model is the same as the true model, and the variance components are known. Considered the following mixed model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{15}$$

with variances var(u) =G =Kσ_u^2, var(e) =R =Iσ_e^2, var(y) =V, where y is the vector of phenotypes, X is the design matrix of fixed effects, b is the vector of fixed effect coefficients corresponding to the *best linear unbiased estimators* (BLUEs), Z corresponds to the incidence matrix of genotypes, u is vector of random effect coefficients corresponding to the breeding values estimated as the *best linear unbiased predictors* (BLUPs), V is the variance-covariance matrix of the phenotypes, G and R are the variance-covariance matrices of the genetic values and residuals, respectively, the matrix K represents the additive genetic relationship among individuals as depicted by pedigree, genomics, or a combination of both. If genomic information is available, the additive genomic relationship is computed as $\mathrm{K} = \alpha^{-1}\mathrm{MM}'$, where M is the genotyping matrix of centralized scores, so that $\{AA, Aa, aa\}$ are coded as $\{2q,\ 1-2p,\ -2p\}$ for diploid organisms, and the scale α that normalizes the trace, $\alpha = \mathrm{tr}\,(\mathrm{MM}')\,\mathrm{m}^{-1}$, such that the mean diagonal of K is one. The variance components, σ_{u}^2 and σ_{e}^2, are herein scalars of the additive genetics and random term variances.

The accuracy ($\mathrm{a}_{\mathrm{A|B}}$) of population A predicted by population B can be described as follows:

$$\mathbf{a}_{\mathbf{A|B}} = \mathbf{cor}\left(\mathbf{u}_{\mathbf{A}}, \widehat{\mathbf{u}}_{\mathbf{A|B}}\right) = \frac{\mathbf{cov}\left(\mathbf{u}_{\mathbf{A}}, \widehat{\mathbf{u}}_{\mathbf{A|B}}\right)}{\sqrt{\mathbf{var}\,(\mathbf{u}_{\mathbf{A}})\,\mathbf{var}\left(\widehat{\mathbf{u}}_{\mathbf{A|B}}\right)}} \tag{16}$$

Under the assumption the statistical model is the true model,

$$\mathbf{cov}\left(\mathbf{u}_{\mathbf{A}}, \widehat{\mathbf{u}}_{\mathbf{A|B}}\right) = \mathbf{cov}\left(\widehat{\mathbf{u}}_{\mathbf{A|B}}, \widehat{\mathbf{u}}_{\mathbf{A|B}}\right) = \mathbf{var}\left(\widehat{\mathbf{u}}_{\mathbf{A|B}}\right) \tag{17}$$

the equation simplified to.

$$\mathbf{a}_{\mathbf{A|B}} = \sqrt{\frac{\mathbf{var}\left(\widehat{\mathbf{u}}_{\mathbf{A|B}}\right)^2}{\mathbf{var}\,(\mathbf{u}_{\mathbf{A}})\,\mathbf{var}\left(\widehat{\mathbf{u}}_{\mathbf{A|B}}\right)}} = \sqrt{\frac{\mathbf{var}\left(\widehat{\mathbf{u}}_{\mathbf{A|B}}\right)}{\mathbf{var}\,(\mathbf{u}_{\mathbf{A}})}} \tag{18}$$

In matrix notation, $\mathrm{var}\left(\widehat{\mathrm{u}}_{\mathrm{A|B}}\right)=\mathrm{G}_{\mathrm{A,B}}\mathrm{Z}'\,\mathrm{V}_{\mathrm{B}}^{-1}\mathrm{Z}\mathrm{G}_{\mathrm{B,A}}$ and $\mathrm{var}\left(\mathrm{u}_{\mathrm{A}}\right)=\mathrm{G}_{\mathrm{A}}^{-1}$, that translates accuracy into:

$$\mathbf{A}_{\mathbf{A|B}}=\sqrt{\mathbf{G}_{\mathbf{A}}^{-1}\,\mathbf{G}_{\mathbf{A,B}}\mathbf{Z}'\,\mathbf{V}_{\mathbf{B}}^{-1}\mathbf{Z}\mathbf{G}_{\mathbf{B,A}}} \tag{19}$$

Where G_{AB} herein represents the genetic covariance between individuals from populations A and B. Such matrix can be computed as $\mathrm{G}_{\mathrm{AB}}=\mathrm{K}_{\mathrm{A,B}}\sigma_{\mathrm{u(A,B)}}=\alpha\,\mathrm{M}_{\mathrm{A}}\mathrm{M}_{\mathrm{B}}'\sigma_{\mathrm{u(A,B)}}$. Note that when the data fitting the model is the same as the population being predicted (A =B), then the accuracy becomes the squared root of the heritability:

$$\mathbf{A}_{\mathbf{A}}=\sqrt{\mathbf{G}_{\mathbf{A}}^{-1}\mathbf{G}_{\mathbf{A}}\mathbf{Z}'\,\mathbf{V}_{\mathbf{A}}^{-1}\mathbf{Z}\mathbf{G}_{\mathbf{A}}}=\sqrt{\mathbf{Z}'\,\mathbf{V}_{\mathbf{A}}^{-1}\mathbf{Z}\mathbf{G}_{\mathbf{A}}}=\sqrt{\mathbf{H}_{\mathbf{A}}^{2}} \tag{20}$$

For computational convenience, the accuracy is computed at the individual level. For the i^{th} individual of population A, the accuracy is defined by:

$$\mathbf{a}_{\mathbf{A(i)}}=\sqrt{\frac{\mathbf{g}_{\mathbf{A(i),\,B}}\mathbf{Z}'\left(\mathbf{Z}\mathbf{G}_{\mathbf{B}}\mathbf{Z}'\,+\mathbf{R}\right)^{-1}\mathbf{Z}\mathbf{g}_{\mathbf{B,A(i)}}}{\mathbf{g}_{\mathbf{ii}}}} \tag{21}$$

Based on the formula above, accuracy is a function of (1) the individuals that constitute population B, as represented by variance and covariance matrices G_{B} and $\mathrm{G}_{\mathrm{A,B}}$; (2) experimental design and number of replicates (Z); and (3) by the variance components. The genetic relationship between populations A and B is captured by the covariance $\mathrm{K}_{\mathrm{A,B}}$, such that a better representation of the genetics from A in B in key factor in improving accuracy (Habier et al. 2007, 2013). The genotype-by-environment correlation is also captured under $\mathrm{G}_{\mathrm{A,B}}$, as $\sigma_{\mathrm{u(A,B)}}=\rho_{\mathrm{A,B}}\sigma_{\mathrm{B}}^{2}$, since $\rho_{\mathrm{A,B}}$ corresponds to the additive genetic correlation between the environments where population B was observed and the environments where population A is being predicted upon (Fig. 3). Genotype-by-environment interactions within the environments of the calibration set (ρ_{GxE}) are depicted within G_{B}, an acceptable assumption when data from population B has been observed in multiple correlated environments, thus $\mathrm{G}_{\mathrm{B}}=\Sigma_{u}\otimes\mathrm{K}$. The experimental design information is captured by Z, as the residual variance linearly decrease with an increase in the number of replicates, hence increasing the proportion of genetic signal in the calibration population B (Fig. 3).

Changing any of the various components of accuracy (Z, $\mathrm{K}_{\mathrm{A,B}}$, $\rho_{\mathrm{A,B}}$, ρ_{GxE}) enables the optimization questions such as: (1) What is the best set of individuals of population B that optimize the prediction of population A? (2) What is the number of replicates necessary to achieve the desired accuracy? (3) Are there changes in model parametrization that can improve accuracy? (4) What are the environments that maximize the genotype-by-environment correlation where population B has been observed and the environments where population A is planned to appear? (5) Are there secondary correlated traits that may increase the accuracy? Such questions are pertinent to resource allocation and blueprinting the various stages of a breeding program.

A general optimization task regarding the population of environments considered in breeding zone, is set as follows: Which subset of environments (ϵ) constitute a homogeneous envirotype (E). Such task is necessary to stablish the number of envirotype as well

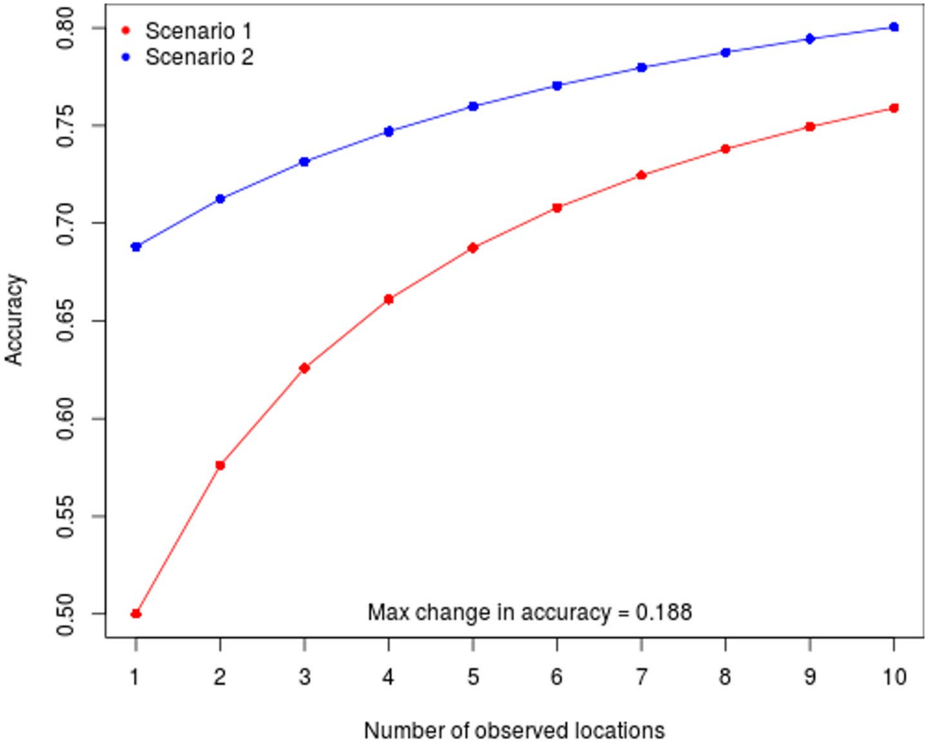

Fig. 3 Case study on prediction accuracy changes under a heritability (h^2) of 0.25 and a genotype-by-environment interaction correlation (ρ_{GxE}) of 0.75. The graph compares two scenarios in a population of families with 10 full siblings: Scenario 1 (red), where relationships among individuals are ignored, and Scenario 2 (blue), where relationships are considered. The x-axis represents the number of observed locations (1 to 10), and the y-axis shows prediction accuracy (0.55 to 0.80). The maximum difference in accuracy between the two scenarios is 0.188

as the proportion of envirotypes within a target population of environments (TPE). The grouping of environments that will define envirotypes is attained through the maximization of genotype-by-environment correlation (ρ_{GxE}) within environments. Thus

$$\underset{\boldsymbol{\epsilon}\ \mathbf{i} \in\ \mathbf{E_i} \forall\ \mathbf{E}}{\mathbf{argmax}}\ \boldsymbol{\rho}_{\ \mathbf{GxE}} \tag{22}$$

corresponding to a straightforward function that can be tackled through both supervised and unsupervised clustering methods. For instance, techniques such as hierarchical clustering are applicable as the correlations between pairs of environments "a" and "b" can be linearly transformed into distance metrics, $\mathrm{d}_{\mathrm{GxE(a,b)}} = 1 - \rho_{\ \mathrm{GxE(a,b)}}$, and the optimization is reframed into the minimizing the pairwise distance between all environments within the envirotype cluster E. Thus:

$$\mathbf{argmin}\left\{\mathbf{d}_{\mathbf{GxE(a,b)}} : \boldsymbol{a} \in\ \mathbf{E_i}, \boldsymbol{b} \in\ \mathbf{E_i}\right\} \tag{23}$$

The complexity of the problem increases as the genotype-by-environment correlation estimated from multiple traits are considered simultaneously, since ρ_{GxE} patterns vary across traits for a same population of genotypes and environments.

Another general optimization task regards the population of genotypes considered in a genomic prediction calibration set optimizations: Which subset of the calibration population ($\mathrm{b} \in \mathrm{B}$) containing a predetermined number of individuals that would provide the highest value of accuracy when predicting the target population ($\mathrm{A}_{\mathrm{A|b}}$).

$$\underset{\mathbf{b} \in \mathbf{B}}{\mathbf{argmax}} \mathbf{A}_{\mathbf{A|b}} \tag{24}$$

The search for a solution containing the best subset of individuals that would yield the highest possible accuracy can be tackled in different ways. The simplest solutions are not computationally feasible, such as the exhaustive search of all possible subsets or stepwise searches that add or drop one candidate at a time. More efficient approaches include genetic algorithms and the search of random spaces (Akdemir et al. 2015). At its simplest, genetic algorithms operate by (1) generating a large population of random solutions, (2) selecting the subset that provides higher values of the objective function, (3) randomly recombining pair of selection solutions into a large population of solutions, (4) repeat steps 1–3 until convergence. Here, a solution ($\mathrm{b} \in \mathrm{B}$) corresponds to a vector of 0's, and 1's, indicating which individuals should be part of the calibration set.

Deterministic accuracy serves as a powerful tool for optimizing breeding programs by leveraging linear stochastic models to estimate the correlation between predicted and true genetic values, even when true values are unknown. By incorporating phenotypic data, genetic relationships, and experimental design elements, accuracy can be calculated at both population and individual levels, supported by equations that account for variance components and G×E interactions. Factors such as genetic covariance between populations and the number of experimental replicates play an important role in enhancing prediction accuracy. Optimization strategies, including clustering environments to maximize G×E correlations and using genetic algorithms to select ideal calibration sets, enable breeders to address key questions about resource allocation and envirotype classification, ultimately improving the efficiency and outcomes of genomic prediction efforts.

6 Revisiting genome editing for breeding

CRISPR-mediated genome editing has emerged as a revolutionary tool with broad implications in agriculture including molecular biology, functional analysis, and genetic modifications (Gao 2018). Moreover, CRISPR-mediated genome editing has been widely applied to develop desirable traits in various crops, including rice, maize, wheat, sugarcane, soybean, potato, sorghum, orange, cucumber, tomato, flax, and cassava, for attributes such as herbicide resistance, drought tolerance, disease resistance, and improved product quality, with some traits nearing commercial release (Gao 2018). Beyond generating novel alleles, genome editing facilitates the promotion of superior alleles and the removal of deleterious ones identified through large-scale sequencing (Gao 2018). Additionally, a reverse domestication approach has been proposed, where genes associated with domestication traits in wild species are edited to create new or improved crop varieties, enhancing stress resistance and promoting crop diversification (Gao 2018). However, the success of CRISPR applications critically relies on the accurate prediction and optimization of guide RNA (sgRNA) sequences for efficient target recognition and DNA cleavage Zong et al. (2022). Traditional approaches for designing sgRNAs often involve time-consuming empirical testing, which is limited by experimental constraints and lacks scalability (Cheng et al. 2023; Das et al. 2023). In recent years, the application of ML techniques in CRISPR-mediated genome editing has garnered significant attention as a promising avenue for enhancing the efficiency and precision of this technology (Sherkatghanad et al. 2023). ML algorithms have the potential to analyze vast amounts of genomic data, extract complex patterns, and generate predictive models that can guide the design of highly efficient sgRNAs (Hesami et al. 2021). ML-based methods can be considered a promising approach to overcome the limitations of conventional methods, accelerate the discovery of optimal sgRNA sequences, and streamline the process of CRISPR-mediated genome editing (Wang et al. 2020). In this section, common issues and possible solutions in the application of ML in CRISPR-mediated genome editing such as data labeling, data selection, readable features to ML, and selection algorithms have been reviewed and discussed.

6.1 Data labeling

In supervised ML, defining a suitable "label" is one of the initial steps (Hesami et al. 2022). In CRISPR gene-editing experiments, the label can be related to factors like fluorescence-based expression measurements, gene knockdown efficiency, or cleavage efficiency (Sherkatghanad et al. 2023). The label can be represented continuously (e.g., within a range from 0 to 1) or discretely (e.g., low or high). Different factors such as the data, the algorithm, and the desired outcome are involved in selecting the representation (Li et al. 2022). Discrete variables are typically handled using classification algorithms, while regression algorithms are used for continuous variables (Yoosefzadeh-Najafabadi et al. 2024). For instance, sgRNA cleavage efficiency falls under continuous representation, as it ranges from 0 to 100%. Using a regression algorithm, a trained model can predict the efficiencies of 4 unlabeled sgRNAs and assign each one a value within this range, such as [0%, 45%, 85%, 98%]. In this scenario, higher efficiency is desirable, making an sgRNA prediction of 98% the optimal choice (O'Brien et al. 2021).

Discrete representation can also be used for continuous values like sgRNA cleavage efficiency. For instance, in a classification model, sgRNAs with "< 50%" efficiency can be labeled as "low," while those with "≥50%" efficiency can be labeled as "high." Consequently, the developed model would classify the four sgRNA cleavage efficiency mentioned earlier as [low, low, high, high] (O'Brien et al. 2021). However, this discretization eliminates the ability to distinguish between the top two sgRNAs as they are all simply classified as "high," disregarding the specific values of 85% and 100%. (O'Brien et al. 2021) Although regression may be more informative than classification, classification offers the advantage of faster training and prediction (Sherkatghanad et al. 2023).

Existing regression algorithms for predicting the efficiency of CRISPR lack high accuracy, and there is often a weak correlation between predicted and observed efficiencies. For instance, there may not be a substantial difference in efficiency between a target predicted to be 98% and one predicted to be 85%. Predicting efficiency accurately in biological systems is challenging due to their complexity (O'Brien et al. 2021). Less information is required through categorizing sgRNAs as either "low" or "high" efficiency compared to assigning them precise efficiency values on a scale from 0 to 100% (O'Brien et al. 2021). Incomplete feature sets and limited sample sizes make the low/high classification less informative but more accurate than continuous predictions (Sherkatghanad et al. 2023). Classification-based models still hold value in distinguishing highly active sites until regression models can better predict the efficiency of sgRNA (Wang et al. 2020).

Imbalance is a common issue with genome editing dataset, where there is a disparity between positive and negative editing results (Wang et al. 2020). This can occur due to biased reporting of positive results or an overwhelming number of negative results, such as from low homology directed repair (HDR) efficiency (Sherkatghanad et al. 2023). When training a classification model, addressing data imbalance involves selecting an appropriate threshold to convert continuous efficiency into binary values (low/high) (O'Brien et al. 2021). Choosing a threshold of 50% may seem logical, but if only 2 out of 10 targets represents more than 50% efficiency, a ML algorithm could classify all targets as low efficiency and still achieve 80% accuracy. Adjusting the decision threshold, such as lowering it to 20%, can balance the number of high- and low-efficiency samples (O'Brien et al. 2021). However, this may not be ideal as targets with efficiency >20% are now classified as high efficiency (O'Brien et al. 2021). Modifying target sampling, such as using a bootstrap method to oversample the minority class, can be a potential solution to address this issue, as demonstrated by DeepCRISPR (Chuai et al. 2018) and CRISTA (Abadi et al. 2017).

Imbalance becomes a more challenging issue when dealing with labels that have more than 2 classes. One such example is forecasting the specific changes resulting from CRISPR outcomes, as attempted by SPROUT (Leenay et al. 2019) and FORECast (Allen et al. 2019). Although this approach enhances control over experimental results, this approach also raises the number of distinct classes, requiring a larger training dataset to effectively fit the model. For instance, in a perfectly balanced dataset of 1000 samples with binary labels (low/high), each class would have 500 samples (1000/2). However, if the same dataset is labeled based on the single-nucleotide change (G, A, C, or T) present in each sample, the sample size per class decreases to 250 (1000/4). When predicting other outcomes like indels (insertions or deletions), the sample size per class diminishes further, potentially resulting in classes with only a single sample (O'Brien et al. 2021). To address this issue and ensure an adequate number of samples for each possible scenario, FORECast (Allen et al. 2019) is trained on over 40,000 sgRNAs. However, in cases where large sample sizes are not feasible, an alternative solution is to train multiple models or limit the number of classes (O'Brien et al. 2021). For instance, instead of training an individual model with labels for every type of change (e.g., T, A, TT, AT, AA), SPROUT employs multiple models. One model may be trained on the type of single-nucleotide change, while another is trained on the length of deletions. This approach allows SPROUT to achieve successful training with 1656 sgRNAs (O'Brien et al. 2021).

6.2 Data selection

In addition to labels, each sample in ML models requires features. Data (e.g., epigenetic data, and genetic data) is presented by features that is transformed into a suitable format for training the ML model (Hesami et al. 2022). The issue lies in feeding enough data to the algorithm for generating accurate results while avoiding difficult or expensive-to-obtain data, experiment-specific information, or irrelevant data. The goal is to develop a model that not only is "generalizable" but makes also correct predictions (Sherkatghanad et al. 2023).

Genetic data is a fundamental component in all the models discussed in this section. Genetic data consists of the adjacent nucleotides, protospacer adjacent motif (PAM), and/or sgRNA sequence (Das et al. 2023). One of the reasons for incorporating genetic data is that sequence information is universal. Additionally, efficient sgRNAs have been shown to prefer certain nucleotides over other nucleotides (Li et al. 2022). The sequence of sgRNA is crucial for guiding Cas protein to its target, making it a known property in previously performed genome editing experiments (providing larger training dataset) and experiments in the planning stage (Sherkatghanad et al. 2023). Therefore, the variability in the data required by different tools primarily lies in the window size around the sgRNA target (e.g., 30 nucleotides for TUSCAN (Wilson et al. 2018), CRISPRpred (Rahman and Rahman 2017), and sgRNA design (Doench et al. 2014); 26 nucleotides for WU-CRISPR (Wong et al. 2015); 23 nucleotides for ge-CRISPR (Kaur et al. 2016). Since these tools solely rely on sequence information as input variables, they can forecast the efficiency of sgRNA irrespective of the species or cell type.

CRISPRpred (Rahman and Rahman 2017) and Azimuth (Doench et al. 2016) seek to enhance accuracy compared to baseline models by incorporating positional features such as "position of target in gene" and "exon targeted". While the inclusion of these features improves the performance of the model, it also reduces generalizability compared to sequence-only models, as genetic annotation becomes necessary for predicting sgRNA efficiency. Consequently, predictions become specific to certain species. In cases where positional information is unavailable, Azimuth resorts to using the sequence-only sgRNA design algorithm (Doench et al. 2016).

Chari et al. (2015) discovered that epigenetic status, determined through histone lysine K4 trimethylation data and DNase-seq, serves as an additional modulator of efficiency. Although incorporating epigenetic information may increase the accuracy of the model, it introduces species and cell type specificity (Chari et al. 2015). Therefore, only sequence information has been used for sgRNA scorer 2.0 (Chari et al. 2017) and sgRNA scorer (Chari et al. 2015) algorithms.

To improve information and accuracy, it is crucial to refrain from incorporating irrelevant data when selecting features (Li et al. 2022). Ideally, a set of features must only consist of properties that have a direct causal relationship with the label. The addition of irrelevant features can have negative consequences by expanding the search space and introducing noise (Sherkatghanad et al. 2023; Wang et al. 2020). This can potentially lead to a decrease in performance of the developed model (Hesami et al. 2022).

6.3 Readable features for ML

After identifying the data to be used for training, it is necessary to process it to meet specific requirements (Cheng et al. 2023; Li et al. 2022). This is particularly important for sequence data, as most ML algorithms cannot directly handle strings (Sherkatghanad et al. 2023). While an algorithm can recognize that '’CTAC'’ is different from '’CCTA'’, it cannot discern the specific nature of the difference (O'Brien et al. 2021). To capture quantitative distinctions and address this issue, sequence features need to undergo a process known as

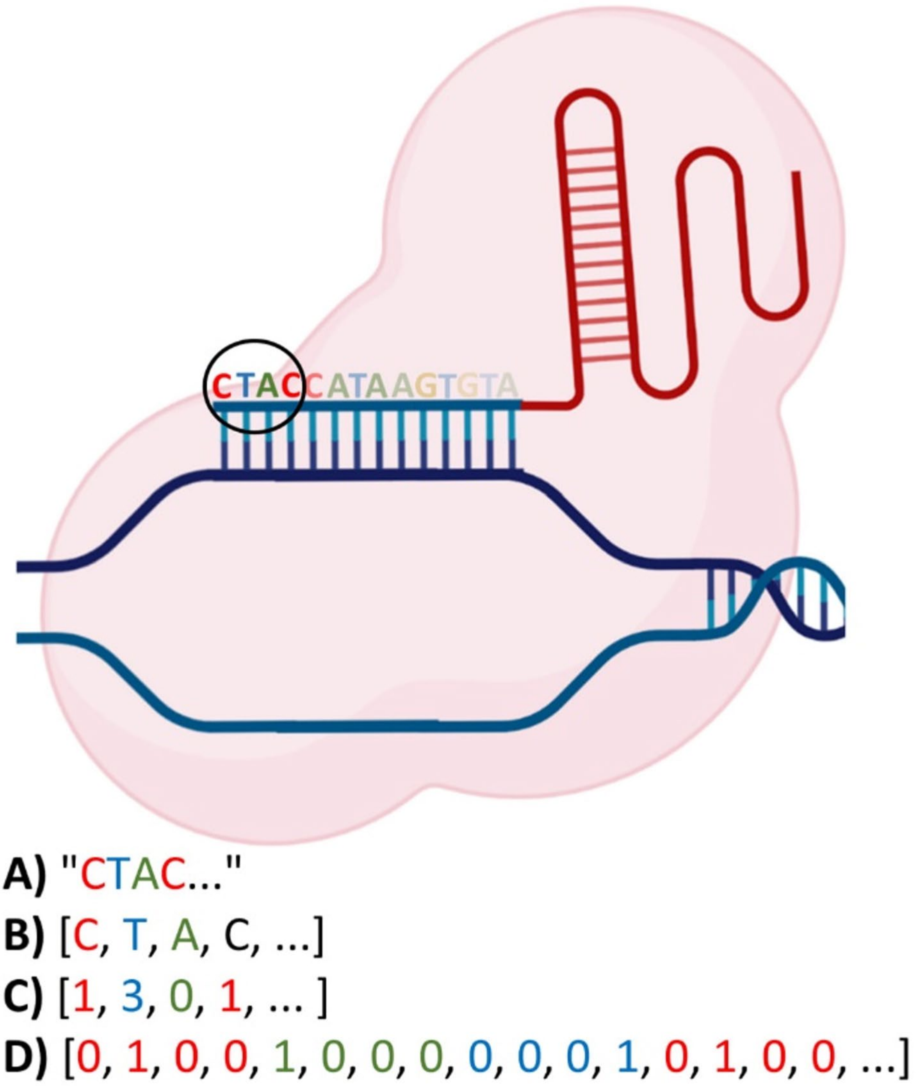

Fig. 4 Schematic representation of various methods for encoding sgRNA sequences. **A** String format, showing the sequence as a continuous text (e.g., "CTAC…"). **B** List of characters, breaking the sequence into individual nucleotides (e.g., [C, T, A, C,…]). **C** Numerical list, assigning integers to each nucleotide (e.g., [1, 3, 0, 1,…]). **D** One-hot encoding, representing each nucleotide as a binary vector (e.g., [0, 1, 0, 0, 1,…]). Created by BioRender.com

"tokenization" (O'Brien et al. 2021). Tokenization enables the representation of these features in a format suitable for analysis and modeling (Fig. 4).

Tokenization is the process of converting an item, such as a string, into a broader format, typically represented as a number array. In the context of DNA or RNA sequences, numbers can be used for the representation of nucleotides (i.e., A= 0, C= 1, G= 2, T= 3) (O'Brien et al. 2021; Xu et al. 2022a, b). For instance, 'CTAC' would become [1, 3, 0, 1], and 'CCTA' would become [1, 1, 3, 0]. This representation enables ML algorithms to identify the position where changes occur. However, it may not be suitable for algorithms that require continuous variables, as the numerical differences between nucleotides are not accurately captured (Sherkatghanad et al. 2023). To address this, we can utilize "one-hot encoding" to represent nucleotides as 0 s and 1 s (O'Brien et al. 2021). This approach uses separate columns for each possible nucleotide and each position in the sequence (Fig. 4). These processes can be extended to generate additional features, such as representing nucleotide pairs. This process entails generating additional feature columns for every possible combination of two nucleotides at each position within the sequence. Feature generation can also be guided by expertise or domain knowledge. For instance, the inclusion of a feature that represents the nucleotides surrounding the 'GG' sequence in the PAM ('NGGN') has been observed to empirically affect efficiency (O'Brien et al. 2021).

6.4 Algorithm selection

By carefully selecting appropriate labels and constructing a well-defined set of features, it becomes feasible to train a model (Yoosefzadeh Najafabadi and Torkamaneh 2025). Various ML algorithms have been employed in CRISPR prediction, each with its own strengths and limitations (Sherkatghanad et al. 2023). In this context, we discuss the commonly utilized algorithms within CRISPR prediction tools.

Logistic regression and linear regression are two frequently employed algorithms in ML. Logistic regression, as employed in sgRNA design (Doench et al. 2014), is suitable for discrete labeling, whereas linear regression, as utilized in CRISPRscan (Moreno-Mateos et al. 2015), is suitable for continuous labeling. Both models establish linear relationships among the labels and the features, but they can also be expanded to capture nonlinear relationships by applying non-linear transformations. For instance, a nonlinear association was identified between efficiency and GC content of sgRNA, where a low or high level of GC content was found to have lower activity compared to a GC content of around 50% (Doench et al. 2014). To account for this nonlinear relationship, two distinct features, one for GC content above 50% and another for GC content below, were introduced, enabling the logistic regression algorithm to model this nonlinearity (Doench et al. 2014). However, alternative algorithms capable of handling nonlinear separations can be employed to avoid the need for manual transformations.

Support-vector machines (SVMs) are an example that supports nonlinear separation. Several CRISPR prediction tools, such as CRISPRpred (Rahman and Rahman 2017), which uses SVM for sgRNA on-target activity prediction, alongside CRISPR-DT (H. Zhu and Liang 2019), TSAM (Peng et al. 2018), WU-CRISPR (Wong et al. 2015), sgRNA scorer 2.0 (Chari et al. 2017), ge-CRISPR (Kaur et al. 2016), sgRNA design (Doench et al. 2014), and sgRNA scorer (Chari et al. 2015) utilize trained SVM models for classification or regression (Table 3). Features can be transformed into a high-dimensional representation via SVM, enabling linear separation in this transformed space, thereby modeling nonlinear data (Hesami et al. 2021). While SVMs offer improved accuracy, the trade-off is the "black box" behavior (i.e., lack of transparency), as the contribution of individual features to the decision process is obscured (Greener et al. 2022; Lin and Wong 2018).

Capturing higher-order interactions between features is another crucial aspect in the CRISPR field. Tree-based methods are a group of models that excel at capturing these inter-

Table 3 Some ML-based online tools for gRNA on- and/or off-target prediction

Tool	Model	Prediction	PAM	References
DeepCRISPR	Convolution Neural Network	Off-target and on-target	NAA, NCG, NGA, NGG, NTG, NGC, NAG, NGT	(Chuai et al. 2018)
CNN_std	Convolution Neural Network	Off-target	NAA, NCG, NGA, NGG, NTG, NGC, NAG, NGT,	(Lin and Wong 2018)
Elevation	Boost Regression Tree + Logistic Regression	Off-target	NGT, NGG, NGC, NGA, NCG, NAG	(Listgarten et al. 2018)
Predict CRISPR	ensemble SVM classifier	Off-target	NGG	(Peng et al. 2018)
CRISTA	Random Forest	Off-target	NGG	(Abadi et al. 2017)
Synergizing CRISPR	AdaBoost	Off-target	NGG	(Zhang et al. 2019)
ge-CRISPR	Support Vector Machine	On-target	NGG	(Kaur et al. 2016)
CRISPRpred	Support Vector Machine	On-target	NGG	(Rahman and Rahman 2017)
Azimuth	Gradient Boost Regression Tree	On-target	NGG	(Doench et al. 2016)
TUSCAN	Random Forest	On-target	NGG	(Wilson et al. 2018)
SgRNAScorer	Support Vector Machine	On-target	NNNNG-MTT, NNAGAAW, NAG, NGG	(Chari et al. 2017)
WU-CRISPR	Support Vector Machine	On-target	NGG	(Wong et al. 2015)
DeepCas9	Convolution Neural Network	On-target	NGG	(Xue et al. 2019)
CRISPRscan	Linear regression	On-target	NGG	(Moreno-Mateos et al. 2015)
DeepCpf1	Convolution Neural Network	On-target	TTTN	(Kim et al. 2018)

actions. Decision trees, for instance, recursively split the dataset based on features to create "pure" groups containing only low-efficiency or high-efficiency targets. If, for example, sgRNAs with a G at position 20 and a GC content of less than 20% have higher efficiencies, decision trees will create additional splits to account for these conditions (Loh 2011). Tree-based methods have the advantage of being applicable to both classification and regression tasks. Additionally, these models can be interrogated to determine the most influential features in prediction efficiency, providing interpretability to the predictions (Loh 2011).

sgRNA efficiency has been predicted by gradient boosting, as seen in Azimuth (Doench et al. 2016), which employs gradient boosting for sgRNA design, and SPROUT (Leenay et al. 2019), as well as random forests (e.g., CUNE (O'Brien et al. 2019), CRISPR-DT (H. Zhu and Liang 2019), TUSCAN (Wilson et al. 2018), CRISTA (Abadi et al. 2017), and CRISPRpred (Rahman and Rahman 2017) as the most well-known tree-based models (Table 3). These algorithms belong to the ensemble methods category. By leveraging this ensemble approach, these models can explore a larger search space and outperform individual trees by reducing errors and improving generalization.

Advancements in computational power have paved the way for deep learning, a subset of ML that involves algorithms with multiple layers of nonlinearity, such as convolutional neural networks (CNNs) (Greener et al. 2022; Sherkatghanad et al. 2023). In sgRNA prediction, deep learning has gained traction with the development of tools such as DeepCRISPR (Chuai et al. 2018), which utilizes CNNs for optimized CRISPR guide RNA design, alongside DeepCas9 (Xue et al. 2019), off_target_prediction (Lin and Wong 2018), and DeepCpf1 (Kim et al. 2018). The choice of approach depends on factors such as the task type (regression or classification), linearity or nonlinearity of the problem, the need to capture feature interactions, and the desire to identify influential features (Table 3). When multiple algorithms are applicable to a problem, it is often appropriate to conduct benchmarks and comparisons to determine the optimal solution (Sherkatghanad et al. 2023).

In conclusion, to propel progress in CRISPR ML modeling, it is crucial for data scientists and experimental researchers to collaborate closely (Greener et al. 2022). This collaboration can facilitate advancements in several ways. Firstly, by working together, they can collaboratively build large datasets specifically tailored for ML training. These datasets can be shared and made accessible through repositories, enabling broader utilization and enhancing the quality of ML models (Greener et al. 2022). Secondly, it is essential to recognize the value of both positive and negative examples in the context of CRISPR efficiency (O'Brien et al. 2021). By publishing and sharing these examples, researchers can contribute to a comprehensive understanding of target sites that exhibit varying degrees of efficiency. Lastly, the computational factors identified as influential in genome editing experiments should be translated into experimentally testable hypotheses (O'Brien et al. 2021). This approach bridges the gap between computational predictions and empirical validation, fostering a more rigorous and reliable exploration of CRISPR-Cas9 designs. By embracing this collaborative approach, data scientists and experimental researchers can collectively drive advancements in CRISPR ML modeling and pave the way for more effective CRISPR applications.

7 Current challenges and future perspectives

While ML is undeniably a powerful tool, it is still a keystone in genomics-based plant breeding. ML provides a powerful set of tools for predictive analytics, but some limitations of deploying these black-box models are not well understood. Moreover, problems may also arise at the implementation level: whereas coding ML pipelines can be an objectively simple task, other components of the whole system are not trivial, particularly regarding the validation, implementation, productionize, and curation of the models. A series of good practices and deployment checks is proposed by Sculley et al. (2015). Authors question the use of questionable data or non-reproducible features, bad coding and implementation practices that may create problems with maintaining the system, and the need for updating models and continuous monitoring the results. An additional concern of data scientists regards epistemological problems of ML (Carabantes 2020), as complex algorithms are predisposed to a problem referred to as cognitive mismatch, when the machine does not learn the intended signal. Robust applications ML has the scope limited by the quality and quantity of the data, as models project the biases from the data into the predictions. Serious problems model attributed to biased data have been previously identified on ML applied to judicial decisions, mortgage lending, and heath care (Varona et al. 2021). Biased plant breeding data leads to poor decision-making with serious economic consequences. That may include under or overrepresentation of certain genotypes or environments. Selection process per se depletes genetic variance and enlarges linkage disequilibrium blocks surrounding the major genes (Doebley et al. 2006), limiting the predictive scope of individuals for the purpose of calibrating estimation sets. For genome-wide association analysis, bias due to population structure weaken the target genetic signal (Ioannidis et al. 2009; Xu 2003). Biases in datasets may originate from systematic problems with data collection, including over-representation of certain individuals such as experimental checks; censorial bias from opportunistic notes (e.g., lodging and disease scores); and inherited bias in multi-stage analysis, where the output from certain models serve as input for downstream models (Hellström et al. 2020). Moving forward, it is crucial to focus on improving these areas to ensure ML is effectively integrated into breeding strategies. By fostering collaborations among data scientists, geneticists, and breeders, we can address these challenges and fully realize the potential of ML, paving the way for innovative solutions and substantial progress in developing resilient and productive crops.

8 Conclusion

Machine learning brings transformative potential to genomics-based plant breeding by enhancing the integration and analysis of complex multi-omics data. Through sophisticated algorithms, ML can significantly improve prediction accuracy and accelerate the development of crops with desirable traits. This advancement allows breeders to tackle pressing global challenges such as food insecurity by expediting crop improvement processes. ML's ability to uncover intricate genetic patterns and interactions offers a dynamic approach to selecting and developing high-performing genotypes. The integration of ML into plant breeding pipelines enables breeders to optimize resources, reduce costs, and increase breeding efficiency. Looking ahead, the focus should be on leveraging ML to harness diverse data types, including genomics, phenomics, and environmental inputs, to drive robust breeding decisions. By fostering collaborations among data scientists, geneticists, and breeders, the potential of ML can be fully realized, paving the way for innovative solutions and substantial progress in developing resilient and productive crops. This review paper contributes by evaluating ML's role across the five G's (germplasm characterization, genome assembly, gene-assisted breeding, genomic prediction, and gene editing), assessing its strengths and limitations, proposing strategies to address data biases and implementation challenges, and identifying future research pathways to enhance ML-driven breeding for improved crop resilience and productivity.

Author contributions MYN conceptualized, proposed, prepared, and organized the contents, wrote the abstract, introduction, revisiting genetic variability (germplasm characterization and genome assembly), revisiting gene assisted breeding, revisiting genomics - based breeding approaches, conclusion and future perspective sections, review, and edited the entire paper. AC organized the contents and wrote the optimizing genomic-based plant breeding section, contributed revisiting genetic variability (germplasm characterization and genome assembly) section, illustrations, and edited the entire paper, ME organized the contents and wrote the revisiting genomics - based breeding approaches section and edited the entire paper, MH organized the contents and wrote the revesting genome editing- based breeding section, illustrations and edited the entire paper.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Abadi S, Yan WX, Amar D, Mayrose I (2017) A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. PLoS Comput Biol 13(10):e1005807. https://doi.org/10.1371/journal.pcbi.1005807

Akdemir D, Sanchez JI, Jannink J-L (2015) Optimization of genomic selection training populations with a genetic algorithm. Genet Selection Evol 47(1):1–10

Allen F, Crepaldi L, Alsinet C, Strong AJ, Kleshchevnikov V, De Angeli P, Parts L (2019) Predicting the mutations generated by repair of Cas9-induced double-strand breaks. Nat Biotechnol 37(1):64–72. https://doi.org/10.1038/nbt.4317

Azodi CB, Bolger E, McCarren A, Roantree M, de los Campos G, Shiu S-H (2019) Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3 Genes|Genomes|Genetics 9(11):3691–3702. https://doi.org/10.1534/g3.119.400498

Belgiu M, Drăguţ L (2016) Random forest in remote sensing: A review of applications and future directions. ISPRS J Photogrammetry Remote Sens 114:24–31

Bernardo R (2016) Bandwagons I, too, have known. Theor Appl Genet 129(12):2323–2332. https://doi.org/10.1007/s00122-016-2772-5

Bogard M, Hourcade D, Piquemal B, Gouache D, Deswartes J-C, Throude M, Cohan J-P (2021) Marker-based crop model-assisted ideotype design to improve avoidance of abiotic stress in bread wheat. J Exp Bot 72(4):1085–1103

Borg I, Groenen PJ, Mair P (2018) Applied multidimensional scaling and unfolding

Brown AV, Grant D, Nelson RT (2021) Using crop databases to explore phenotypes: from QTL to candidate genes. Plants 10(11):2494

Carabantes M (2020) Black-box artificial intelligence: an epistemological and critical analysis. AI Soc 35(2):309–317

Chapman SC (2008) Use of crop models to understand genotype by environment interactions for drought in real-world and simulated plant breeding trials. Euphytica 161(1):195–208

Chari R, Mali P, Moosburner M, Church GM (2015) Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nat Methods 12(9):823–826. https://doi.org/10.1038/nmeth.3473

Chari R, Yeo NC, Chavez A, Church GM (2017) SgRNA scorer 2.0: A species-independent model to predict CRISPR/Cas9 activity. ACS Synth Biol 6(5):902–904. https://doi.org/10.1021/acssynbio.6b00343

Cheng X, Li Z, Shan R, Li Z, Wang S, Zhao W, Li W (2023) Modeling CRISPR-Cas13d on-target and off-target effects using machine learning approaches. Nat Commun 14(1):752. https://doi.org/10.1038/s41467-023-36316-3

Chuai G, Ma H, Yan J, Chen M, Hong N, Xue D, Liu Q (2018) DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biol 19(1):80. https://doi.org/10.1186/s13059-018-1459-4

Costa-Neto G, Crespo-Herrera L, Fradgley N, Gardner K, Bentley AR, Dreisigacker S, Crossa J (2022) Envirome-wide associations enhance multi-year genome-based prediction of historical wheat breeding data. G3 Genes|Genomes|Genetics. https://doi.org/10.1093/g3journal/jkac313

Das J, Kumar S, Mishra DC, Chaturvedi KK, Paul RK, Kairi A (2023) Machine learning in the Estimation of CRISPR-Cas9 cleavage sites for plant system. Front Genet 13:1085332. https://doi.org/10.3389/fgene.2022.1085332

Delicado P (2011) Dimensionality reduction when data are density functions. Comput Stat Data Anal 55(1):401–420

Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. Cell 127(7):1309–1321

Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Root DE (2014) Rational design of highly active SgRNAs for CRISPR-Cas9–mediated gene inactivation. Nat Biotechnol 32(12):1262–1267. https://doi.org/10.1038/nbt.3026

Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Root DE (2016) Optimized SgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol 34(2):184–191. https://doi.org/10.1038/nbt.3437

Emanuelli F, Lorenzi S, Grzeskowiak L, Catalano V, Stefanini M, Troggio M, Moreira FM (2013) Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. BMC Plant Biol 13(1):1–17

Falck F, Zhang H, Willetts M, Nicholson G, Yau C, Holmes CC (2021) Multi-facet clustering variational autoencoders. Adv Neural Inf Process Syst 34:8676–8690

Fitzpatrick MC, Keller SR (2015) Ecological genomics Meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. Ecol Lett 18(1):1–16

François O, Durand E (2010) Spatially explicit bayesian clustering models in population genetics. Mol Ecol Resour 10(5):773–784

Gao C (2018) The future of CRISPR technologies in agriculture. Nat Rev Mol Cell Biol 19(5):275–276. https://doi.org/10.1038/nrm.2018.2

Geleta M, Montserrat DM, Giro-i-Nieto X, Ioannidis AG (2023) Deep variational autoencoders for population genetics. Biorxiv 20232009:2027–558320

Gianola D (2013) Priors in whole-genome regression: the bayesian alphabet returns. Genetics 194(3):573–596

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173(3):1761–1776

Giglio C, Brown SD (2018) Using elastic net regression to perform spectrally relevant variable selection. J Chemom 32(8):e3034

Goddard M (1998) Consensus and debate in the definition of breeding objectives. J Dairy Sci 81:6–18

Greener JG, Kandathil SM, Moffat L, Jones DT (2022) A guide to machine learning for biologists. Nat Rev Mol Cell Biol 23(1):40–55. https://doi.org/10.1038/s41580-021-00407-0

Griffel L, Delparte D, Edwards J (2018) Using support vector machines classification to differentiate spectral signatures of potato plants infected with potato virus Y. Comput Electron Agric 153:318–324

Grünwald N, Kamvar Z, Everhart S (2010) Discriminant analysis of principal components (DAPC)

Habier D, Fernando RL, Dekkers J (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389–2397

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12(1):1–12

Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194(3):597–607

Harfouche AL, Jacobson DA, Kainer D, Romero JC, Harfouche AH, Mugnozza GS, Altman A (2019) Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. Trends Biotechnol 37(11):1217–1235

Heffner EL, Lorenz AJ, Jannink JL, Sorrells ME (2010) Plant breeding with genomic selection: gain per unit time and cost. Crop Sci 50(5):1681–1690

Hellström T, Dignum V, Bensch S (2020) Bias in machine learning–what is it good for? ArXiv Preprint. arXiv:2004.00686

Henryon M, Berg P, Sørensen A (2014) Animal-breeding schemes using genomic information need breeding plans designed to maximise long-term genetic gains. Livest Sci 166:38–47

Hesami M, Yoosefzadeh Najafabadi M, Adamek K, Torkamaneh D, Jones AM (2021) Synergizing off-target predictions for in silico insights of CENH3 knockout in cannabis through CRISPR/Cas. Molecules 26(7):2053. https://doi.org/10.3390/molecules26072053

Hesami M, Alizadeh M, Jones AMP, Torkamaneh D (2022) Machine learning: its challenges and opportunities in plant system biology. Appl Microbiol Biotechnol 106(9):3507–3530. https://doi.org/10.1007/s00253-022-11963-6

Hickey LT, Hafeez N, Robinson A, Jackson H, Leal-Bertioli SA, Tester SC, Wulff M, B. B (2019) Breeding crops to feed 10 billion. Nat Biotechnol 37(7):744–754

Holland JB (2004) Implementation of molecular markers for quantitative traits in breeding programs—challenges and opportunities. Paper presented at the Proceedings of the 4th international crop science congress

Hong H, Najafabadi MY, Torkamaneh D, Rajcan I (2022) Identification of quantitative trait loci associated with seed quality traits between Canadian and Ukrainian mega-environments using genome-wide association study. Theor Appl Genet 135(7):2515–2530. https://doi.org/10.1007/s00122-022-04134-8

Hout MC, Papesh MH, Goldinger SD (2013) Multidimensional scaling. Wiley Interdisciplinary Reviews: Cogn Sci 4(1):93–103

Ioannidis JP, Thomas G, Daly MJ (2009) Validating, augmenting and refining genome-wide association signals. Nat Rev Genet 10(5):318–329

Jahufer M, Arojju SK, Faville MJ, Ghamkhar K, Luo D, Arief V, Griffiths AG (2021) Deterministic and stochastic modelling of impacts from genomic selection and phenomics on genetic gain for perennial ryegrass dry matter yield. Sci Rep 11(1):1–18

Jiao W-B, Schneeberger K (2017) The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol 36:64–70

Jombart T, Collins C (2015) A tutorial for discriminant analysis of principal components (DAPC) using adegenet 2.0. 0. London: Imperial College London, MRC Centre for Outbreak Analysis and Modelling

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11(1):94. https://doi.org/10.1186/1471-2156-11-94

Kaur K, Gupta AK, Rajput A, Kumar M (2016) ge-CRISPR - An integrated pipeline for the prediction and analysis of SgRNAs genome editing efficiency for CRISPR/Cas system. Sci Rep 6(1):30870. https://doi.org/10.1038/srep30870

Kim HK, Min S, Song M, Jung S, Choi JW, Kim Y, Kim H (2018) Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. Nat Biotechnol 36(3):239–241. https://doi.org/10.1038/nbt.4061

Kuhn M, Silge J (2022) Tidy modeling with R. O'Reilly Media, Inc.

Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Team RC (2020) Package 'caret'. R J 223(7):48

Leenay RT, Aghazadeh A, Hiatt J, Tse D, Roth TL, Apathy R, Zou J (2019) Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells. Nat Biotechnol 37(9):1034–1037. https://doi.org/10.1038/s41587-019-0203-2

Lenaerts B, Collard BC, Demont M (2019) Improving global food security through accelerated plant breeding. Plant Sci 287:110207

Li X, Zhu C, Wang J, Yu J (2012) Computer simulation in plant breeding. Adv Agron 116:219–264

Li R, Li L, Xu Y, Yang J (2022) Machine learning Meets omics: applications and perspectives. Brief Bioinform 23(1):bbab460. https://doi.org/10.1093/bib/bbab460

Liang M, Cao S, Deng T, Du L, Li K, An B, Gao X (2023) MAK: a machine learning framework improved genomic prediction via multi-target ensemble regressor chains and automatic selection of assistant traits. Brief Bioinform 24(2):bbad043

Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. Nat Rev Genet 16(6):321–332

Lin J, Wong K-C (2018) Off-target predictions in CRISPR-Cas9 gene editing using deep learning. Bioinformatics 34(17):i656–i663

Listgarten J, Weinstein M, Kleinstiver BP, Sousa AA, Joung JK, Crawford J, Doench JG (2018) Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. Nat Biomedical Eng 2(1):38–47

Liu S, Xu F, Xu Y, Wang Q, Yan J, Wang J, Wang X (2022) MODAS: exploring maize germplasm with multi-omics data association studies. Sci Bull 67(9):903–906

Loh W-Y (2011) Classification and regression trees. WIREs Data Min Knowl Discov 1(1):14–23. https://doi.org/10.1002/widm.8

López-Cortés XA, Matamala F, Maldonado C, Mora-Poblete F, Scapim CA (2020) A deep learning approach to population structure inference in inbred lines of maize. Front Genet 11:543459

Lorenz A, Nice L (2017) Training population design and resource allocation for genomic selection in plant breeding. Genomic selection for crop improvement. Springer, pp 7–22

McClung AM, Edwards JD, Jia MH, Huggins TD, Bockelman HE, Ali ML, Eizenga GC (2020) Enhancing the searchability, breeding utility, and efficient management of germplasm accessions in the USDA–ARS rice collection. Crop Sci 60(6):3191–3211

Meuwissen TH, Hayes BJ, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829

Miao J, Wu Y, Sun Z, Miao X, Lu T, Zhao J, Lu Q (2024) Valid inference for machine learning-assisted genome-wide association studies. Nat Genet 56:1–9

Moreno-Mateos MA, Vejnar CE, Beaudoin J-D, Fernandez JP, Mis EK, Khokha MK, Giraldez AJ (2015) CRISPRscan: designing highly efficient SgRNAs for CRISPR-Cas9 targeting in vivo. Nat Methods 12(10):982–988. https://doi.org/10.1038/nmeth.3543

Morota G (2017) ShinyGPAS: interactive genomic prediction accuracy simulator based on deterministic formulas. Genet Selection Evol 49(1):1–5

Musa AB (2013) Comparative study on classification performance between support vector machine and logistic regression. Int J Mach Learn Cybernet 4:13–24

Najafabadi MY, Heidari A, Rajcan I (2023) AllInOne Pre-processing: A comprehensive preprocessing framework in plant field phenotyping. SoftwareX 23:101464

O'Brien AR, Wilson LOW, Burgio G, Bauer DC (2019) Unlocking HDR-mediated nucleotide editing by identifying high-efficiency target sites using machine learning. Sci Rep 9(1):2788. https://doi.org/10.1038/s41598-019-39142-0

O'Brien AR, Burgio G, Bauer DC (2021) Domain-specific introduction to machine learning terminology, pitfalls and opportunities in CRISPR-based gene editing. Brief Bioinform 22(1):308–314. https://doi.org/10.1093/bib/bbz145

Onogi A, Nurimoto M, Morita M (2011) Characterization of a bayesian genetic clustering algorithm based on a dirichlet process prior and comparison among bayesian clustering methods. BMC Bioinformatics 12(1):1–16

Pang B, Nijkamp E, Wu YN (2020) Deep learning with tensorflow: A review. J Educational Behav Stat 45(2):227–248

Parmley K, Nagasubramanian K, Sarkar S, Ganapathysubramanian B, Singh AK (2019) Development of optimized phenomic predictors for efficient plant breeding decisions using phenomic-assisted selection in soybean. Plant Phenomics. https://doi.org/10.34133/2019/5809404

Peng H, Zheng Y, Blumenstein M, Tao D, Li J (2018a) CRISPR/Cas9 cleavage efficiency regression through boosting algorithms and Markov sequence profiling. Bioinformatics 34(18):3069–3077. https://doi.org/10.1093/bioinformatics/bty298

Peng H, Zheng Y, Zhao Z, Liu T, Li J (2018b) Recognition of CRISPR/Cas9 off-target sites through ensemble learning of uneven mismatch distributions. Bioinformatics 34(17):i757–i765

Pong-Wong R, Woolliams J (2014) Bayes U: a genomic prediction method based on the horseshoe prior. Paper presented at the World Congress of Genetics Applied to Livestock Production

Qi Y (2012) Random forest for bioinformatics. Ensemble mach learning: methods Appl, 307–323

Quintana FA, Iglesias PL (2003) Bayesian clustering and product partition models. J Royal Stat Society: Ser B (Statistical Methodology) 65(2):557–574

Rahman MK, Rahman MS (2017) CRISPRpred: A flexible and efficient tool for SgRNAs on-target activity prediction in CRISPR/Cas9 systems. PLoS ONE 12(8):e0181943. https://doi.org/10.1371/journal.pone.0181943

Ribaut J, De Vicente M, Delannay X (2010) Molecular breeding in developing countries: challenges and perspectives. Curr Opin Plant Biol 13(2):213–218

Rincent R, Charcosset A, Moreau L (2017) Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. Theor Appl Genet 130(11):2231–2247

Roelofs R, Shankar V, Recht B, Fridovich-Keil S, Hardt M, Miller J, Schmidt L (2019) A meta-analysis of overfitting in machine learning. Adv Neural Inf Process Syst 32:1

Schopp P, Müller D, Wientjes YC, Melchinger AE (2017) Genomic prediction within and across biparental families: means and variances of prediction accuracy and usefulness of deterministic equations. G3: Genes Genomes Genet 7(11):3571–3586

Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, Dennison D (2015) Hidden technical debt in machine learning systems. Adv Neural Inf Process Syst 28:1

Sherkatghanad Z, Abdar M, Charlier J, Makarenkov V (2023) Using traditional machine learning and deep learning methods for on- and off-target prediction in CRISPR/Cas9: a review. Brief Bioinform 24(3):bbad131. https://doi.org/10.1093/bib/bbad131

Sinha P, Singh VK, Bohra A, Kumar A, Reif JC, Varshney RK (2021) Genomics and breeding innovations for enhancing genetic gain for climate resilience and nutrition traits. Theor Appl Genet 134(6):1829–1843

Stuber CW, Lincoln SE, Wolff D, Helentjaris T, Lander E (1992) Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics 132(3):823–839

Tao H, Xu S, Tian Y, Li Z, Ge Y, Zhang J, Zhang Z (2022) Proximal and remote sensing in plant phenomics: twenty years of progress, challenges and perspectives. Plant Commun 3:100344

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414–4423

Varona D, Lizama-Mue Y, Suárez JL (2021) Machine learning's limitations in avoiding automation of bias. AI Soc 36(1):197–203

Varshney RK, Sinha P, Singh VK, Kumar A, Zhang Q, Bennetzen JL (2020) 5Gs for crop genetic improvement. Curr Opin Plant Biol 56:190–196

Voss-Fels KP, Stahl A, Hickey LT (2019) Q&A: modern crop breeding for future food security. BMC Biol 17(1):1–7

Wang D, Gu J (2018) VASC: dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. Genomics Proteom Bioinf 16(5):320–331

Wang J, Zhang X, Cheng L, Luo Y (2020) An overview and metanalysis of machine and deep learning-based CRISPR gRNA design tools. RNA Biol 17(1):13–22. https://doi.org/10.1080/15476286.2019.1669406

Wang K, Abid MA, Rasheed A, Crossa J, Hearne S, Li H (2023) DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. Mol Plant 16(1):279–293

Wientjes YC, Bijma P, Calus MP (2020) Optimizing genomic reference populations to improve crossbred performance. Genet Selection Evol 52(1):1–18

Wilson LOW, Reti D, O'Brien AR, Dunne RA, Bauer DC (2018) High activity target-site identification using phenotypic independent CRISPR-Cas9 core functionality. CRISPR J 1(2):182–190. https://doi.org/10.1089/crispr.2017.0021

Wong N, Liu W, Wang X (2015) WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. Genome Biol 16(1):218. https://doi.org/10.1186/s13059-015-0784-0

Xu S (2003) Theoretical basis of the Beavis effect. Genetics 165(4):2259–2268

Xu F, Yang X, Zhao N, Hu Z, Mackenzie SA, Zhang M, Yang J (2022a) Exploiting sterility and fertility variation in cytoplasmic male sterile vegetable crops. Hortic Res 9:uhab039

Xu Y, Zhang X, Li H, Zheng H, Zhang J, Olsen MS, Qian Q (2022b) Smart breeding driven by big data, artificial intelligence and integrated genomic-enviromic prediction. Mol Plant 15:1664

Xue L, Tang B, Chen W, Luo J (2019) Prediction of CRISPR SgRNA activity using a deep convolutional neural network. J Chem Inf Model 59(1):615–624. https://doi.org/10.1021/acs.jcim.8b00368

Yan J, Wang X (2022) Machine learning bridges omics sciences and plant breeding

Yang H-W, Hsu H-C, Yang C-K, Tsai M-J, Kuo Y-F (2019) Differentiating between morphologically similar species in genus Cinnamomum (Lauraceae) using deep convolutional neural networks. Comput Electron Agric 162:739–748

Yang F, Liu N, Crossley MS, Wang P, Ma Z, Guo J, Zhang R (2021) Cropland connectivity affects genetic divergence of Colorado potato beetle along an invasion front. Evol Appl 14(2):553–565

Yang Z, Wang Z, Wang W, Xie X, Chai L, Wang X, Su Z (2022) GgComp enables dissection of germplasm resources and construction of a multiscale germplasm network in wheat. Plant Physiol 188(4):1950–1965

Yoosefzadeh Najafabadi M (2021) Using advanced proximal sensing and genotyping tools combined with bigdata analysis methods to improve soybean yield. University of Guelph

Yoosefzadeh Najafabadi M, Torkamaneh D (2025) Machine learning-enhanced multi-trait genomic prediction for optimizing cannabinoid profiles in cannabis. Plant J 121(1):e17164. https://doi.org/10.1111/tpj.17164

Yoosefzadeh Najafabadi M, Hesami M, Eskandari M (2023) Machine learning-assisted approaches in modernized plant breeding programs. Genes 14(4):777

Yoosefzadeh-Najafabadi M, Earl HJ, Tulpan D, Sulik J, Eskandari M (2021) Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. Front Plant Sci 11:624273

Yoosefzadeh-Najafabadi M, Eskandari M, Torabi S, Torkamaneh D, Tulpan D, Rajcan I (2022) Machine-learning-based genome-wide association studies for Uncovering QTL underlying soybean yield and its components. Int J Mol Sci 23(10):5538

Yoosefzadeh-Najafabadi M, Hesami M, Eskandari M (2024) Machine learning-enhanced utilization of plant genetic resources. Sustainable utilization and conservation of plant genetic diversity. Springer, pp 619–639

Zhang S, Li X, Lin Q, Wong K-C (2019) Synergizing CRISPR/Cas9 off-target predictions for ensemble insights and practical applications. Bioinformatics 35(7):1108–1115

Zhao W, Lai X, Liu D, Zhang Z, Ma P, Wang Q, Pan Y (2020) Applications of support vector machine in genomic prediction in pig and maize populations. Front Genet 11:598318

Zhu H, Liang C (2019) CRISPR-DT: designing gRNAs for the CRISPR-Cpf1 system with improved target efficiency and specificity. Bioinformatics 35(16):2783–2789. https://doi.org/10.1093/bioinformatics/bty1061

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. plant genome 1(1):1

Zingaretti LM, Gezan SA, Ferrão LFV, Osorio LF, Monfort A, Muñoz PR, Pérez-Enciso M (2020) Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. Front Plant Sci. https://doi.org/10.3389/fpls.2020.00025

Zong Y, Liu Y, Xue C, Li B, Li X, Wang Y, Gao C (2022) An engineered prime editor with enhanced editing efficiency in plants. Nat Biotechnol 40(9):1394–1402. https://doi.org/10.1038/s41587-022-01254-w

Zou H, Hastie T (2003) Regression shrinkage and selection via the elastic net, with applications to microarrays. JR Stat Soc Ser B 67:301–320

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Royal Stat Soc Ser B: Stat Methodol 67(2):301–320