Trends in Genetics

CellPress OPEN ACCESS

Review

Artificial intelligence in plant breeding

Muhammad Amjad Farooq[1,2], Shang Gao[1,2], Muhammad Adeel Hassan[3,4], Zhangping Huang[1,2], Awais Rasheed[5], Sarah Hearne[6], Boddupalli Prasanna[7], Xinhai Li[1], and Huihui Li[1,2,*]

Harnessing cutting-edge technologies to enhance crop productivity is a pivotal goal in modern plant breeding. Artificial intelligence (AI) is renowned for its prowess in big data analysis and pattern recognition, and is revolutionizing numerous scientific domains including plant breeding. We explore the wider potential of AI tools in various facets of breeding, including data collection, unlocking genetic diversity within genebanks, and bridging the genotype–phenotype gap to facilitate crop breeding. This will enable the development of crop cultivars tailored to the projected future environments. Moreover, AI tools also hold promise for refining crop traits by improving the precision of gene-editing systems and predicting the potential effects of gene variants on plant phenotypes. Leveraging AI-enabled precision breeding can augment the efficiency of breeding programs and holds promise for optimizing cropping systems at the grassroots level. This entails identifying optimal inter-cropping and crop-rotation models to enhance agricultural sustainability and productivity in the field.

Highlights

Rapid advances in genomics, phenomics, and molecular biology are accelerating crop breeding into the era of artificial intelligence (AI).

Addressing the formidable challenge of managing the 'data deluge' is crucial for transitioning from disparate datasets to an integrated data infrastructure in AI-enabled plant breeding.

Although AI has revolutionized various aspects of crop breeding, such as phenomics, variant calling models, gene discovery, genomic selection, and gene editing, there is a pressing need to synergize these components into an integrated breeding technology for future crop development.

Accelerating genetic gain in the context of emerging technologies

Enhancing crop productivity remains a formidable challenge amid surging global populations and the escalating impact of climate change-induced weather events. According to Breeder's equation, genetic gain – a measure of increased crop productivity over time – hinges on improvements in selection accuracy, selection intensity, additive genetic variance, and generation turnaround time. In our prior work we advocated the utilization of specific technologies in crop genomics, phenomics, and speed-breeding to expedite the rate of genetic gain [1]. We posit here that **artificial intelligence (AI)** (see Glossary), a ubiquitous force across scientific disciplines, holds promise for accelerating genetic gain.

The current landscape of **plant breeding** is characterized by a 'data deluge' where data generation through omic innovations far outpaces efficient management, archiving, and analysis. Compared to traditional methods, AI tools can extract more useful and less biased insights from high-throughput sequencing and imaging data [2]. For instance, large language model (LLM)-based ChatGPT serves as a potent chatbot that can generate intelligent text and images based on natural language input. This tool has been instrumental in posing thought-provoking questions relevant to plant science [3], filling gaps left by plant experts in projects such as the 'one hundred important questions facing plant science' [4].

Beyond text processing, there is growing interest in harnessing LLMs across all facets of plant breeding to expedite genetic gain. Omic data, akin to specialized language input, offer a foundation for training LLMs to comprehend biological processes at various levels. These data are poised to play a pivotal role in predicting the value of complex traits and in uncovering biological insights into genetic variants comprising different alleles and haplotypes that affect the phenotype of an individual in different ways. Recent advances in LLMs have spurred our examination of the current role of AI in plant breeding and the formulation of a roadmap for its future utilization (Figure 1).

[1]State Key Laboratory of Crop Gene Resources and Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences (CAAS), International Maize and Wheat Improvement Center (CIMMYT) China office, Beijing 100081, China
[2]Nanfan Research Institute, CAAS, Sanya, Hainan 572024, China
[3]Adaptive Cropping Systems Laboratory, Beltsville Agricultural Research Center, US Department of Agriculture, Beltsville, MD 20705, USA
[4]Oak Ridge Institute for Science and Education, Oak Ridge, TN 37830, USA
[5]Department of Plant Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan
[6]CIMMYT, KM 45 Carretera Mexico-Veracruz, El Batan, Texcoco 56237, Mexico
[7]CIMMYT, International Centre for Research in Agroforestry (ICRAF) House, Nairobi 00100, Kenya

*Correspondence:
lihuihui@caas.cn (H. Li).

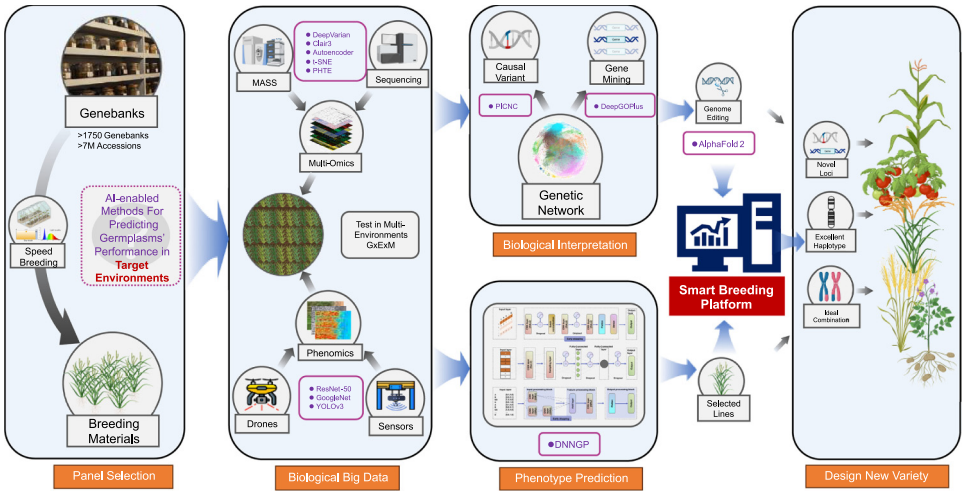

Figure 1. A roadmap of artificial intelligence (AI)-enabled plant breeding. The far left panel depicts the source of germplasm from a genebank that is either used directly or subjected to breeding to utilize omic analyses to generate big data (second panel). The big data from the germplasm could uncover genes and predict breeding values (third panel), which would characterize the gene-editing targets and optimal lines for developing next-generation cultivars. Representative AI-based methods are identified in purple boxes. The dotted box represents methods that require development.

Specifically, we delineate four key areas where AI holds the potential to expedite genetic gain: panel selection, generation of biological big data, biological interpretation, and phenotype prediction.

AI-enabled characterization of germplasm resources to generate genomic big data

High-throughput advances in remote sensing and plant 'omics' have provided scientists with large multi-dimensional datasets to work with when addressing problems. The use of **machine learning (ML)** algorithms in pre-breeding strategies, regional selection, and adaptive marker-assisted selection can boost genetic diversity and speed up the development of cultivars that are climate-resilient. These algorithms can also maintain and restore the dynamic processes that result in genetic variability, which is necessary for plant adaptation to changing climates [5]. Genetic gain stems from the vast biological diversity harbored within crops. There are >1750 genebanks worldwide containing >7 million germplasm accessions [6], including cultivars, landraces, and wild relatives, and the potential of these resources remains largely untapped. Genebank genomics, involving genome-wide genotyping of stored germplasm resources, offers a promising avenue to better understand and utilize these valuable resources. Notably, genome-wide genotyping data have been generated for substantial numbers of wheat, maize, and barley accessions, totaling >80 000 [7], 4000 [8], and 20 000 [9], respectively. These datasets can serve as a foundation for the targeted selection and optimal testing of accessions in specific environments using AI-driven predictive genomics.

A crucial initial step encompasses constructing a reference genome that serves as a foundational resource for comparing individuals within a species. This makes it easier to identify, map, and associate allelic mutations with phenotypic variety, which helps to advance crop breeding efforts. Sequencing technologies have produced substantial improvements, enabling the production of more comprehensive and precise reference genome assemblies. Initiatives such as the 10 KP plan, part of the broader Earth BioGenome Project, symbolize significant progress in this realm [10]. This ambitious project aims to sequence 10 000 plant species, spanning representatives

Glossary

Artificial intelligence (AI): a suite of techniques including ML, DL, and natural language processing to simulate human intelligence processes by machines. These processes include learning (the acquisition of information and rules for using the information), reasoning (using rules to reach approximate or definite conclusions), and self-correction.
Convolutional neural network (CNN): a type of DL algorithm designed specifically for processing and analyzing visual data such as images and videos. It mimics the structure and functioning of the human visual cortex by using layers of neurons to extract features from input images.
Data islands: this term in the AI field usually refers to isolated and disconnected sets of data that are not integrated and interconnected. A major application of AI-enabled plant breeding is to access diverse and comprehensive datasets to train robust and generalized models and predict the plant performance to select ideal plant types (Box 2).
Deep learning (DL): an AI component process to train artificial neural networks (ANNs) with multiple layers to learn a representation of the data. Each layer of the neural network processes the data in a hierarchical manner by extracting increasingly abstract and complex features from the input to automatically discover patterns and features from raw data, without the need for explicit programming.
Deep neural network (DNN): an ANN architecture characterized by multiple layers between the input and output layers. In the context of plant science, DNNs are computational models that are used to process and analyze complex data related to plants, their biology, environment, and interactions. These networks consist of interconnected nodes (neurons) organized into layers, where each layer processes and extracts features from the input data, leading to hierarchical learning of representations.
Gene editing: a technique in molecular breeding to precisely change the phenotype by addition, removal, or alteration of specific DNA sequences. The most common gene-editing technique is clustered regularly interspaced short palindromic repeats (CRISPR) with CRISPR-associated

from all known taxonomic families. For many of these species this endeavor will yield their first comprehensive genome-wide assembly that will serve as a crucial reference for subsequent resequencing efforts aimed at detecting allelic variation within species [11]. The rich diversity inherent in nature presents an opportunity to integrate beneficial traits from wild populations into cultivated elite varieties. This integration enhances the resilience of crops and forest trees to evolving climate patterns. As increasing numbers of genome sequences are compiled, greater potential to delve into genome regulation and variation arises, and consistent updates to reference genome annotations are enabled by the Ensembl browser release 93. Emphasis will be placed on disseminating genomic data to facilitate researcher-driven analyses.

Collecting phenotypic data for such a vast array of germplasm resources across diverse environments presents a formidable challenge. However, studies such as that of Lasky *et al.* [12] have demonstrated that leveraging bioclimatic and soil gradient data of georeferenced sorghum landraces can unveil genomic signatures associated with crop adaptability. Integrating genomics with passport data, which provides genotypic identity information of accessions, offers a means to assess the breeding potential of germplasm, even in the absence of traditional phenotypic data. Enhanced AI-driven genomic prediction models, incorporating georeferenced passport data, agroclimatic variables, and soil gradients, can simulate the performance of different germplasms, thereby facilitating the selection of optimal breeding panels. Thus, this approach addresses a significant constraint in genebank germplasm utilization, namely the scarcity of phenotype information in target environments.

AI-enabled digitalization and collection of phenotyping data

Phenotypic data are essential for crop breeding, but face significant barriers that hinder full utilization. Traditional plant phenotyping methods have long been viewed as a bottleneck in crop breeding owing to their limited data acquisition capacity. However, the recent emergence of plant phenomics represents a fundamental shift in this paradigm [2,13,14]. Plant phenomics, which systematically studies phenotypes, holds promise for overcoming these limitations. Phenomic platforms equipped with advanced imaging sensors have the potential to revolutionize large-scale phenotyping of various plant traits and environmental conditions (Figure 2A). These platforms may utilize either stationary or mobile sensors. Towers and other fixed platforms are commonly used to monitor growth phases because of their simplicity and ease of maintenance. To illustrate, digital cameras mounted on a permanent phenotyping tower have been used to monitor rice growth, nitrogen content, leaf area index, and the presence of the rice bugs [15,16]. The fixed phenotyping tower is simple to set up and maintain, but it has restricted crop information within defined locations. The rail-based field scanalyzer system for field phenotyping, developed by Rothamsted Research, incorporates a sensor array comprising a visible camera, a 3D laser scanner, a thermal IR (TIR) camera, a chlorophyll fluorescence sensor, and a visible to the near-IR hyperspectral camera [17,18]. This setup enables comprehensive characterization of crop canopy development across all growth phases. In addition, a Crop3D high-throughput crop phenotyping platform has been introduced that uses multiple imaging sensors within a movable gantry system to quantify 3D plant and leaf structures, as well as leaf temperature [19]. To address field coverage limitations, sensors are installed on manually powered carts or self-propelled tractors. These platforms have successfully captured canopy traits such as plant height, normalized difference vegetation index (NDVI), temperature, reflectance spectra, and red-green-blue (RGB) imagery for soybean and wheat using a phenocart equipped with diverse sensors [20]. However, various environmental factors, such as light intensity, can impact on imaging platforms in open areas. The BreedVision technology addresses this challenge by effectively blocking ambient light and conducting imaging within a movable dark chamber [21]. This innovative system enables nondestructive measurement of

protein 9 (Cas9) which utilizes a bacterial immune system mechanism to precisely target and edit specific sequences of DNA within the genome.
Genomic selection (GS): also termed genomic prediction, GS is a modern epoch in plant and livestock breeding to predict phenotypes based on genomic information. GS outperforms traditional selection methods of in plant breeding owing to superior selection accuracy and potential for faster genetic gain.
Machine learning (ML): a subset component of AI that can develop algorithms and statistical models that enable computers to perform tasks without explicit instructions. ML uses an algorithm to learn from patterns in a large amount of labeled data and, once trained, predictions or decisions based on that learning can be made in response to new and unseen data.
Plant breeding: the science and art of manipulating plant species to generate desirable traits (Box 1)
Variant calling: the process of identifying variations in a genomic sequence by comparing it to a reference genome or another set of genomic data. The major types of the variants are single-nucleotide polymorphisms (SNPs), small insertions or deletions (indels), and larger structural variations.

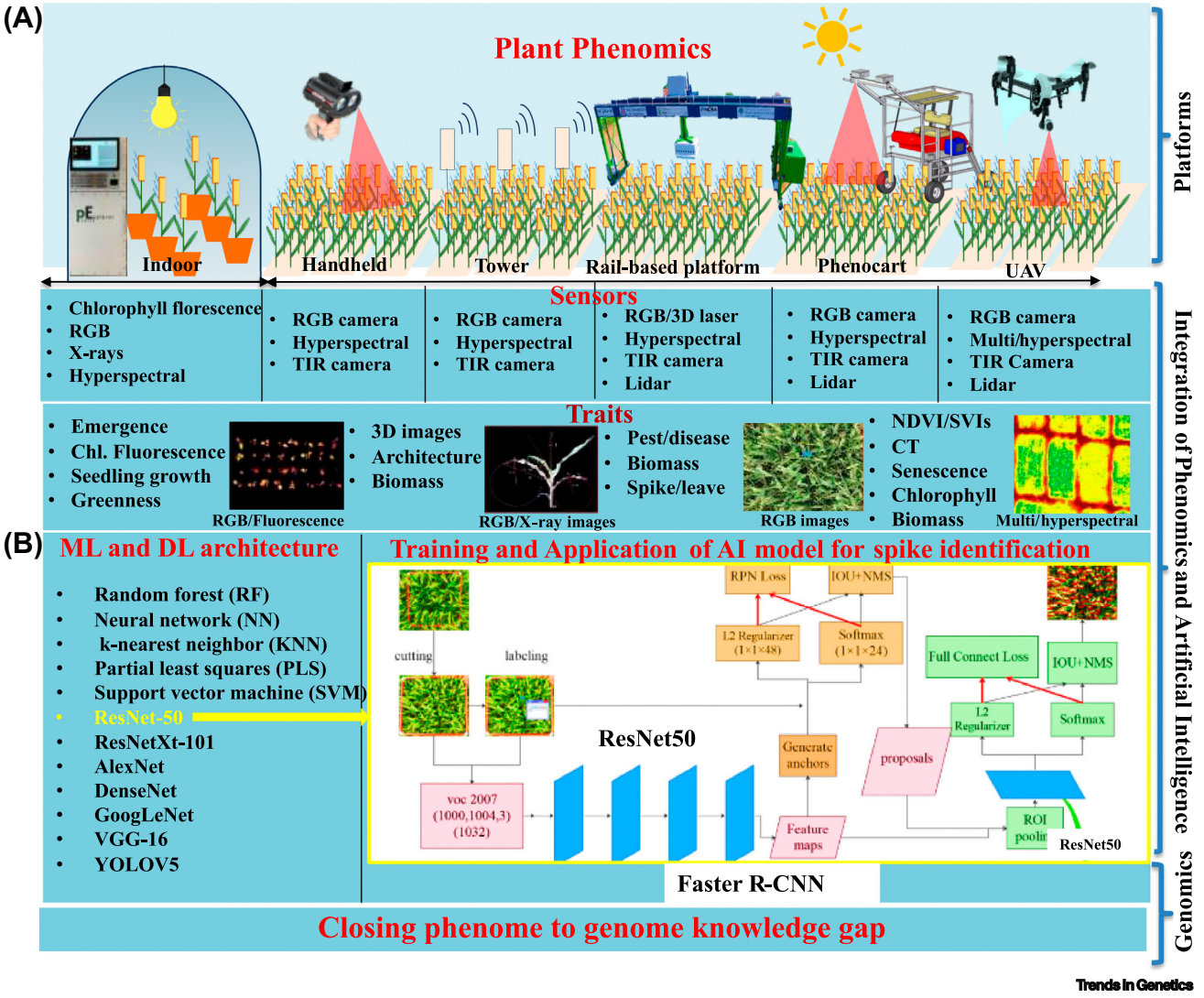

Figure 2. Techniques for plant phenotyping and data analysis. (A) Advanced plant phenomics platforms for digitalization of phenomics and the acquisition of big phenomics datasets. (B) Machine learning (ML)/deep learning (DL) architectures to predict key plant traits by an AI-based approach and a DL-based workflow to identify spike numbers to narrow the phenome-to-genome knowledge gap. Abbreviations: Chl., chlorophyll; CNN, convolutional neural network; CT, computed tomography; IOU, intersection over union; lidar, light detection and ranging; NDVI, normalized difference vegetation index; NMS, non-maximum suppression; pE, _____; RGB, red-green-blue; ROI, region of interest; RPN, region proposed network; SVI, standardized vegetation index; TIR, thermal IR; UAV, unmanned aerial vehicle.

plant traits, including plant height, tiller density, grain yield, moisture content, and leaf color. It utilizes a range of sensors such as 3D time-of-flight cameras, laser distance sensors, hyperspectral imaging, and RGB sensors. However, specific restrictions are placed on vehicle-based platforms based on field topology, weather, and soil conditions. Unmanned aerial vehicles (UAVs) offer a dynamic platform capable of rapidly gathering data over vast regions and producing high-resolution photographs with pixel densities of ~1 mm. High-resolution UAV photography has simplified phenotyping tasks by providing detailed canopy color and texture features with high spatial and temporal resolution from UAV platforms [22]. Consequently, high-resolution UAV photography has found applications in various phenotyping tasks across several crops, such as wheat ear identification and senescence quantification [23,24].

To enable feature identification in some complex tasks, including stress phenotyping, ML tools use processing methods to extract patterns and features from these large amounts of data. **Deep learning (DL)** is valuable for many scientific research tasks because it is effective at identifying complex structures within high-dimensional data [25]. The wide variability of plant images obtained from diverse sensors poses a challenge to the application of particular ML techniques [26]. Hence, workflows for feature identification have become more dependent upon on DL tools. Various ML and DL architectures, including random forest (RF), neural network (NN), *k*-nearest neighbor (KNN), partial least squares (PLS), and support vector machine (SVM), alongside models such as ResNet-50, ResNetXt-101, AlexNet, DenseNet, GoogLeNet, VGG-16, and YOLOV5, have been used to predict essential crop traits (Figure 2B). For example, **convolutional neural networks (CNNs)** have been applied in RGB image analysis to segment roots of chicory, wheat, and rapeseed [27–29], leaf counting [30], and classification and quantification of biotic and abiotic stresses in leaves across various crops [31], as well as predicting spike number and yield of individual barley and wheat plants [24,32]. For spike identification, the Faster R-CCN model labeled the training set images using supervised learning to enhance model accuracy (Figure 2B). Screening plant responses to different stresses can aid in selection decisions for developing climate-resilient cultivars, a crucial task in many breeding programs. A CNN was used to provide an interpretable diagnosis of biotic stress in individual soybean plants by integrating saliency maps into the analysis of multispectral and hyperspectral data [33,34]. To characterize the response of sorghum to drought stress, researchers have used linear discriminant analysis (LDA) and PLS models [35]. A CNN was used to generate interpretable classifications of biotic and abiotic stressors in soybean leaves by isolating the top-k feature maps learned by the model [36]. The nitrogen nutrition index in rice was estimated using a combination of ML algorithms including RF, NN, KNN, PLS, and SVM to enhance nitrogen use efficiency [37,38]. In maize, above-ground biomass was estimated using PLS and RF models [34]. YOLO V5-CAcT, an innovative network architecture, has emerged as a solution for swiftly identifying diseases in field crops. Implemented on the novel CNN (NCNN) DL platform, it has become an effective tool for combating crop diseases at an industrial level. In addition, SVMs were used for 3D point cloud analysis to estimate yield and characterize the geometry of apple canopies [39], and both SVMs and Gaussian processes proved to be successful in analyzing TIR images to detect drought stress in spinach [40].

Recent studies demonstrate that combining data from different sources yields superior results than using individual sources alone. For example, integrating RGB, TIR, and multispectral data through a **deep neural network (DNN)** improved soybean yield prediction accuracy [41]. Yoosefzadeh *et al.* [42] also found significant associations between highly heritable secondary characteristics (hyperspectral vegetation indices, HVIs) and soybean seed yield and fresh biomass production. It creates ensemble-bagging (EB) and DNN methods to anticipate HVI data throughout the early phases of growth. The optimal HVI values are related to the strength Pareto evolutionary algorithm 2 (SPEA2) for maximum yield and biomass.

Similar advances are seen with the use of extreme learning machines for estimating soybean nitrogen concentration, leaf area index, aboveground biomass, and chlorophyll content [43]. To better grasp the biological mechanisms behind desirable traits and plant responses to environmental stresses, integration of phenomic and genomic datasets, initially separate, is proposed [44,45]. These integration efforts should encompass environmental data, including climate types, because the phenotype of a plant is influenced by the interaction between its genotype and its environment (G × E). Such integration is crucial for designing crop ideotypes optimized for specific environments in the face of rapid climate change [45].

AI-enabled predictions to explain genomic data

Numerous studies have demonstrated the potential applications of AI in interpreting biochemical data to advance the understanding of plant stress biology. For instance, AI has been effectively used to forecast genomic crossovers in maternal and parental maize plants, thus aiding in the identification of genomic regions with elevated mutation rates [46]. Furthermore, based on the DNA methylation patterns of maize plants growing under stress, researchers have used AI approaches to identify and characterize genomic areas, thereby differentiating between functional genes and pseudogenes [47]. Similar to this, Uygun *et al.* [48] used AI algorithms to examine the expression patterns of important genes to predict gene promoters and *cis*-regulatory elements in *Arabidopsis thaliana* and maize plants. In addition, by identifying tissue-specific variations in biosynthetic genes, such as those linked to nitrogen utilization efficiency, starch biosynthesis, and secondary metabolites in *A. thaliana* and rice, studies have demonstrated the value of AI in deciphering plant metabolic regulatory networks [49,50]. Similarly, Meena *et al.* [51] demonstrated the crucial role that AI can play in bioenergy management by utilizing AI to optimize biomass generation by using a variety of plant species and algal blooms to increase biofuel production.

Sequencing technology is advancing swiftly, and long-read sequencing has emerged as the predominant method in crop breeding. This technology offers the advantage of gaining more structural variations and haplotype data than short-read sequencing, but introduces more errors and challenges for variant identification. Marked by the ability to extract meaningful patterns from complex data, methods building on DL have been put forth to enhance the accuracy and efficiency of **variant calling** from short-read sequencing [52] and long-read sequencing [53,54]. These approaches use different strategies, including haplotype-aware modeling, image-based representation, local realignment, and full-alignment, to leverage the complex information in long reads and reduce the influence of errors, demonstrating superior performance over existing tools and enabling the discovery of novel variants in challenging-to-map regions of the genome [55]. A primary drawback is that these tools are primarily trained on human somatic cells. The variant calling identified in Table 1 requires thorough revision for use in crop plants. Clearly, the enhanced accuracy in variant calling will improve the precision of **genomic selection (GS)** and marker-assisted breeding, thus enabling breeders to make informed decisions for crop improvement.

Identifying single-nucleotide polymorphisms (SNPs) and small insertion/deletion (indel) variants remains a significant challenge using both second- and third-generation sequencing technologies. Artificial neural networks (ANNs) are emerging as a solution to this challenge. Recently, a CNN model called Clairvoyante was proposed by Luo *et al.* [56] for predicting SNP or indel variations, zygosity, and indel length from long-read alignments. They evaluated Clairvoyante using data from Illumina, PacBio, and Oxford Nanopore sequencing platforms, focusing on finding common variant sites from the 1000 Genomes Project dataset with a minor allele frequency of at least 5%.

Another significant application of ANNs in variant calling was demonstrated by Poplin *et al.* [52] and their DeepVariant package. DeepVariant uses a sophisticated approach that calculates the likelihood of three possible allele combinations at each variant site. This tool learns to distinguish between homozygous or heterozygous alleles and the reference and homozygous alleles within the variations by analyzing statistical correlations between pictures of reads surrounding putative variants and genuine genotype calls. Furthermore, ML algorithms extend beyond variant detection in long-read sequencing practices, as evidenced by their application in population genetics studies. Supervised ML approaches have successfully examined recombination rates within target genomes. For instance, Schrider and Kern [57] used a RF classifier to differentiate recombination rate levels in *Drosophila melanogaster* using DNA motifs. The

Table 1. Representative methods used in AI-enabled plant breeding

Domain	Organism	Technology	Model/method	Problem to solve	Refs
High-throughput phenotyping	Universal	Weakly supervised DL	ResNet-50 CNN Transformer	Reduce manual labeling efforts Improve the accuracy of measurement	[139] [140] [141]
Multi-omic data integration	Universal	Autoencoder neural networks	CNN GCNN	Establish a roadmap for utilizing multi-omic data in DL methods	[103]
Variant calling	Universal	DL	DeepVariant Clair3 Pepper NanoCaller	Development of fast and precise variant calling tools	[52] [54] [55] [53]
Gene discovery and functional prediction	*Spartina alterniflora*	DL	DeepGOplus	Identification of candidate genes underlying specific biological processes	[100]
Causal variant inference	Maize	Pre-trained model ML	UniRep Random forest PICNC	Prioritization of candidate causal variants	[102]
Genomic prediction	Universal	DL	CNN DNNGP	Achieve higher performance with multi-omic data	[123]
Optimizing genome-editing tools	Universal	DL	AlphaFold	Discovery or optimization of protein function	[99]
Genomic intelligent design	Model species	DL	GAN FGAN	Promoter design or optimization of protein function	[79] [80]
Transposon detection	Universal	DL	Inpactor2 TEsorter	Identification of transposable elements	[77] [78]

combination of genomics and ML to produce innovative approaches for predicting agricultural diseases and simplifying bioinformatic pipelines without coding expertise is very important. Noteworthy technology providers in this field include Trace Genomics and Sequentia Biotech. Trace Genomics is featured in soil health applications where patented ML algorithms are used to pinpoint crucial elements influencing crop performance. By contrast, Sequentia Biotech offers solutions such as AI RNA-seq (AIR) to streamline the procedures for data generation and interpretation in transcriptomic studies [58]. The future of ML may entail addressing multiple species simultaneously. DL methods may be used for objectives such as knowledge transfer from model plants to target crops or comparative genomics studies.

Integration of multi-omic big data in plant breeding

At the biochemical level, 'omics' encompasses diverse molecular data sources such as genomics, epigenomics, protein–DNA interactions, transcriptomics, proteomics, and metabolomics. In the past decade a vast amount of omic data have been generated, inundating the web with transcriptomic, genomic, proteomic, and metabolomic information [59]. Recently, biology has evolved into an information-intensive science because of the massive datasets generated by high-throughput sub-molecular biological experiments across various domains, including genomics, transcriptomics, proteomics, and metabolomics. In bioinformatics, the inventory of components at the genome, transcriptome, proteome, and metabolome levels is gradually becoming comprehensive and invaluable to researchers [60]. ML is frequently used to assess vast datasets because of their size, complexity, and the requirement for combined interpretation. Even within each omic technology, several profiling techniques have also matured, in addition to RNA-seq for genome-wide gene expression, chromatin immunoprecipitation with deep sequencing (ChIP-seq) [61],

DNA affinity purification sequencing (DAP-seq) [62], and assay for transposase-accessible chromatin sequencing (ATAC-seq) [63–65]. Access to ever more detailed information on systems biology, from single genetic elements to regulatory networks, offers us unprecedented chances to capture the true biological substance underlying phenotypic variation. However, for speeding up crop breeding, it is still challenging to properly integrate various data layers, connect them to stress reactions in the environment, and precisely model the overall system.

One area where ML has demonstrated utility is in identifying various types of genomic regions. For instance, in maize, the live genes and dead pseudogenes can be classified by an ML model trained on DNA methylation features [47]. ML-based method has also been developed to predict the crossover probability along a chromosome [66]. In addition, ML is starting to find applications in population genetics, although predominantly in humans [67]. One application of ML in plants was to predict the near-complete fixation of mutations retained by natural selection [68]. These instances highlight how ML, beyond its conventional role in the areas of gene and genome annotation [69], can be used to further explore genome function, complementing traditional comparative genomics approaches.

AI-enabled bridging of the genotype–phenotype gap

There has been growing interest in utilizing AI for precise, non-destructive estimation of crop traits and genetic studies. The development of cultivars with high yield potential that are resilient to climate change relies heavily on genetic improvements through the identification of novel alleles and genomic-assisted selections in crop breeding. Several studies have successfully applied AI-based approaches to hyperspectral and RGB datasets to predict early wheat yield and conduct quantitative genomic analysis, and identified novel alleles in wheat [70,71]. Lei *et al.* [24] implemented an automatic spike-counting system using quantitative genetic analysis with RGB images from a ground-based camera, and improved the accuracy and generalization of the Faster R-CNN model by increasing training data and annotations (Figure 2B). Accurate disease screening is crucial for developing disease-resistant cultivars and has been addressed through a genome-wide association study (GWAS) analysis using DL-based detection of sudden death syndrome (SDS) in soybean [72], and revealed significant SNPs near candidate SDS genes. In addition, studies have utilized multimodal DL for yield predictions, demonstrating that methods such as phenotype–genotype multiple instance learning (PheGeMIL), leave-one-environment-out (LOEO), support vector regression (SVR), and gradient boosting machine (GBM) leverage phenotypic observations to enhance genomic predictions for wheat yield [73–75]. In summary, integrating phenomics, genomics, and AI offers a promising approach to monitor crop productivity, assess responses to abiotic and biotic stressors, and identify novel genes and quantitative trait loci (QTLs). This integrated approach could pave the way for accelerating crop breeding efforts to develop climate-resilient crops [76].

'Missing heritability' refers to the instance where all the genetic variants cannot completely account for the heritability of complex traits, and is a major bottleneck in genetic studies. In such cases, integrating transposons and epigenetic mutations such as DNA methylation can offer more comprehensive information to explain the genetic foundation of phenotypes. There have been massive advances in the identification and classification of transposons in plants using DL, including Inpactor2 [77] and TEsorter [78]. In addition to AI-enabled gene identification to bridge the genome–phenome gap, AI has promising applications in functional genomics of the genes identified to construct a genomic intelligent design. For instance, generative adversarial networks (GANs) were used to optimize and enhance genomic functions via synthetic promoter design in *Escherichia coli* [79]. The model, guided by sequence features gleaned from natural promoters, includes interactions between nucleotides at different positions and designs novel synthetic promoters *in silico*. Moreover, the feedback GAN (FBGAN) introduces a novel

feedback-loop organization to optimize protein functions by generating synthetic DNA sequences encoding proteins of variable length and optimizing them using an external function analyzer [80]. There is still a need for further work to adapt cutting-edge synthetic methods to plant biology and crop improvement [81].

Functional genomics and gene mining using AI

Several ML methodologies have emerged to prioritize genes relevant to agronomic traits, such as leveraging gene functions [82], exploring protein interactions [83], and incorporating gene annotation and sequence variation [84]. An intriguing avenue for further exploration lies in integrating evolutionary insights into ML models. For instance, recent research has demonstrated the predictive power of leveraging knowledge from well-annotated species to infer gene functions in less-characterized species, particularly in predicting specialized metabolism genes.

What distinguishes many of the studies is their dual focus on optimizing prediction performance and unraveling biologically meaningful features underlying the data. For instance, Lin *et al.* [84] identified transcription factors as being pivotal in prioritizing genes associated with specific traits in *A. thaliana* and rice. Similarly, Demirci *et al.* [66] found that particular DNA shape features predicted crossover occurrence across various plant species such as *A. thaliana*, tomato, maize and rice, highlighting both commonalities and species-specific nuances. Some studies utilized various machine learning models to identify genes responsible for abiotic stresses [85–87]. Such insights gleaned from ML models pave the way for generating testable hypotheses, such as identifying genomic regions, candidate genes, or protein residues for further experimental validation.

An exciting frontier where ML plays a crucial role is in single-cell RNA sequencing [88,89], enabling the exploration of developmental processes and responses to environmental stimuli within complex, heterogeneous tissues. The resulting datasets are comprehensive, and encompass data from thousands of cells and tens of thousands of genes. Examining such data often encompasses the application of unsupervised ML techniques. Unlike many studies where a specific 'label' is predicted, unsupervised ML methods aim to identify patterns that can help to organize and interpret data lacking predefined labels. Examples of such approaches include clustering and manifold learning approaches that seek to uncover underlying structures within the data in a nonlinear manner similar to principal component analysis [90].

Although metabolomics encounters a challenge owing to the unknown characteristics of several components, ML offers solutions that integrate and analyze metabolomic data. ML methodologies enable the prediction of metabolic pathways, exemplified by studies such as those focusing on tomatoes [91]. ML is anticipated to make significant contributions in multi-omic analysis which integrates transcriptomic, proteomic, and metabolomic data, as illustrated by McLoughlin *et al.* [92]. They used extensive multi-omic analysis to examine maize autophagy mutants cultivated under nitrogen-replete and -starvation conditions to ascertain how cells/tissues depend on autophagy. Even in the absence of stress, plants lacking the essential autophagy component ATG12 showed broad abnormalities in the leaf metabolome, with a focus on lipid turnover products and secondary metabolites. These changes were supported by significant modifications in the transcriptome and/or proteome. By comparing the abundances of mRNA and proteins, it was possible to identify specific protein targets for autophagic clearance, as well as protein complexes and organelles, and to identify many processes regulated by this catabolism.

The mlDNA R package, short for ML-based differential network analysis, used an RF algorithm to pinpoint salt stress-related genes in *A. thaliana* [93]. Another innovative tactic involves amalgamating multiple gene regulation networks (GRNs), along with prioritization algorithms, which led to the

identification and validation of OsbHLH148, an important drought-related transcription factor in rice [94]. In contrast to gene discovery relying on GRNs, GWAS offers a comparatively direct method to detect genes associated with traits and natural variations natural variants for molecular breeding. Nonetheless, a challenge arises from the possibility that a QTL identified via GWAS may encompass numerous genes owing to linkage disequilibrium (LD). Selecting candidate genes for validation is still a major difficulty in biological research. To tackle this challenge, various ML-based methods targeting GWAS-identified QTLs have been developed for gene prioritization and causal mutation elucidation. These ML approaches encompass penalized regression, gradient boosting machines (GBMs), Bayesian approaches, and DL [95]. ML-based gene prioritization relies on compiling features from known genes and the causal variants underlying phenotypic variation [96]. Taking QTG-Finder as an example – a GWAS results explainer equipped with various ML methods [84] – this is able to utilize a feature set comprising 28 attributes derived from genomic data of *A. thaliana*, including DNA polymorphisms, functional annotations, cofunction networks, and evolutionary conservation. An enhanced version, QTG-Finder2, incorporates orthologous information from multiple model plants to enrich the feature set for comprehensive gene discovery [97]. QTG-Finder2 effectively prioritizes genes identified through GWAS in non-model plant species, thus addressing the challenge of scarce causal gene information. However, the limited prior knowledge is still a significant challenge in ML-based methods for gene discovery in plants. Such an issue may be overcome in the future by semi-supervised learning strategies such as positive-unlabeled (PU) learning [98].

Recent studies have underscored the pivotal role of ML and DL techniques in addressing specific biological challenges, particularly in the identification of salt-tolerance genes in plants. A LLM-based gene function prediction model was developed and deployed to identify differentially expressed genes under salt stress, and unraveled crucial pathways [99]. Yang *et al.* [100] also utilized a DL-based gene annotation tool, DeepGOPlus, to detect high-affinity K^+ transporters (HKTs) linked to salt tolerance, and revealed 16 *Spartina alterniflora* (Sa)HKT genes with varied expression patterns and ion transport preferences. Besides using protein sequences as the input, Gao *et al.* [85] merged ML with coevolutionary information about genes to pinpoint salt stress-related genes, and their study emphasized the enrichment of genes associated with ion transport and detoxification pathways. These findings highlight the indispensable role of computational methodologies in unraveling the molecular intricacies of salt-tolerance mechanisms in *S. alterniflora* and offer invaluable insights into plant functional genomics. These works demonstrate the flexibility of DL and ML to use various data types to study key genes underlying plant abiotic stress.

Finding excellent alleles and causal variants from omic data

Constructing genetic networks from multi-omic data is often more potent than single-omic applications in uncovering complex relationships in biological systems. Nonetheless, the ultra-high dimensionality of omic data often causes the 'curse of dimensionality' [101], posing significant challenges to single-omic data and integrated multi-omic data analyses [102]. Therefore, reducing dimensionality is always the first step in multi-omic data analysis. Kang *et al.* [103] identified a roadmap for multi-omic data integration, encompassing denoising data with autoencoders and extracting features with DL methods such as CNNs. Although this roadmap was proposed for biomedical data, it offers a valuable touchstone for harnessing large datasets in plant studies. With powerful data integration methods, multi-omic analysis can often provide more accurate predictions than single-omic analysis. This is likely because these methods include complex interactions and dependencies between various biological data types, thereby offering a more comprehensive and holistic understanding of biological systems.

Functional investigations of plant genes, particularly those underlying agronomic traits, offer essential information for breeding crops via genetic modification and genome editing. Characterizing gene functions from multi-omic data is foundational to the discovery of genes implicated in specific processes.

Leveraging the power of DL, Yang *et al.* [100] successfully used a DL-based model to examine multi-omic data, and identified a series of salt tolerance-related genes in non-model halophytic plants and validated their functions. DL has also been applied to directly characterize causal phenotypic variants, which is a challenging but valuable goal. For instance, the ML-based method PICNC (prediction of mutation impact by *calibrated nucleotide conservation*) can infer functional alterations caused by mutations in maize. It integrates mutation importance information obtained through UniRep, a long short-term memory (LSTM)-based mutation/protein structure assessment method [102]. Foundational AI tools can be used to create more advanced tools for biological interpretation.

Practical plant breeding by predicting phenotypes using AI-enabled genomic selection

Marker-assisted selection (MAS) and GS are two primary marker-based breeding techniques used to characterize plants with desirable characteristics. MAS faces limitations in detecting genes with relatively small effects on complex traits. When markers linked to these traits capture only a fraction of genetic variation, MAS may underperform compared to phenotypic selection [104].

To address the limitation of disparity in prediction accuracy between models, the predominant approach involves conducting repeated trials using diverse statistical models to identify an optimal scenario for target phenotypic traits. GS are broadly classified into parametric and nonparametric methods based on the utilization of prior information and parameter settings [105]. Parametric methods include regularized linear regression (RLR) models, such as ridge regression (RR) and least absolute shrinkage and selection operator (LASSO) [106], which address the over-parameterization issue inherent in simple linear models. ML-based statistical models, including SVM [107], ANN [108], and RF [109] have found applications in plant breeding. There are obstacles to finding the best statistical approach because crops, cultivars, habitats, populations, and markers vary. Therefore, while using GS, breeders need to compare and choose practical statistical approaches that are tailored to each circumstance.

ML utilizes statistical approaches to allow systems to learn from data without explicit programming. Using a sample dataset, ML produces models to explore algorithms that can learn from accessible data and produce predictions for unseen data. ML-based approaches have enhanced prediction accuracy than traditional GS [110]. Unlike conventional statistical models, ML offers flexibility to allow complex relationships between input data and outcomes. As the scale of genome data expands and complexities emerge, the development of informative and predictive models becomes challenging. Therefore, the use of ML is on the rise because it offers a crucial alternative owing to its flexibility and usefulness in navigating these complexities [111] via modifying obscure patterns of unidentified structures that parametric models are unable to include [112].

Traditional statistical approaches struggle to examine the genetic foundation of plant quantitative traits, especially in complex scenarios involving pleiotropic genes, epistasis, and gene–environment (G x E) interactions. The challenge lies in identifying all marker effects, producing the 'large *P*, small *N*' issue, as well as possible over-parameterization. ML approaches offer a solution by leveraging repeated experiences to enhance prediction accuracy [113]. ML algorithms are categorized into supervised and unsupervised learning approaches: supervised learning is designed to predict target values according to input data, whereas unsupervised learning uncovers groupings and associations among input variables without output variables. ML-based GS methods mainly consist of supervised learning models [114].

SVM, a typical ML-based model, provides benefits in both classification and regression tasks. What differentiates SVM is its specialization in detecting subtle patterns within complex and

diverse datasets. SVM develops decision boundaries using various feature vectors to produce accurate predictions. This type of method enhances the non-linear formation between phenotypes and genotypes by utilizing various kernel functions [115]. In recent years, ANN methods have showed their potential in the application of GS. ANNs can identify patterns in data and generate predictions for complex functions, thus serving as universal approximators [108]. In GS, these functions accurately detect factors such as epistasis or dominance in genomic markers. Furthermore, they do not rely on assumptions about the phenotypic distribution, and using ANNs in GS enables effective estimation of the impact of complex interactions [116].

Several studies have used DL to analyze GS. Montesinos-López *et al.* [117] used a densely coupled network architecture to compare genomic best linear unbiased prediction (gBLUP) and DL models. Nine published genomic datasets (six wheat and three maize datasets) were evaluated in this study. DL demonstrated better prediction accuracy across six of nine datasets when G × E interactions were ignored. Another study uncovered that SVM and multilayer perceptron (MLP) exhibit superior computational efficiency than other approaches [118]. The current body of literature on DL-based GS methods is sparse in comparing prediction accuracy with traditional statistical approaches. Therefore, further research will be necessary to bridge this gap. DL builds upon ANNs by integrating three or more ANNs into a DNN structure [119]. Popular DL architectures in GS are MLPs, CNNs, and recurrent neural networks (RNNs) [120]. Typically supervised, MLPs integrate at least one hidden layer and are excellent for various applications owing to their simplicity and effectiveness in prediction tasks. Despite their versatility, MLPs may overfit during training, potentially reducing accuracy when applied to real-world datasets [121].

CNNs are primarily utilized in tasks associated with computer vision which take images or video data as the input. A key aspect of CNNs is their efficiency achieved through input size reduction and parameter sharing. This optimization limits the number of parameters that require estimation, thus improving computational efficiency. Typical CNN architectures are composed of three primary operations: convolution, nonlinear transformation, and pooling. These operations reduce input size without compromising pertinent information, thereby facilitating rapid training through parameter reduction [122].

RNNs do not strictly propagate in a single direction: they incorporate feedback loops that enable signals to travel both forward and backward via synaptic connections. Therefore, training RNNs requires significant computational resources [120]. DNN genomic prediction (DNNGP) is a foundational DL-based genome selection method that can integrate multi-omic data to predict plant phenotypes [123]. This method integrates a well-designed algorithm structure to limit overfitting and enhance convergence speed. It significantly outperforms conventional approaches in prediction accuracy, especially when dealing with large populations. Such AI-based tools will gradually replace traditional approaches in plant breeding, especially in the context of exponentially increasing volumes of biological data. Using prior knowledge of gene functions, expression, and interactions is helpful to guide the genomic prediction models. This will assist in reducing the dimensionality and complexity of the data, as well as elevating the biological interpretability and reliability of the predictions [124]. AI approaches can also integrate prior knowledge into genomic prediction models using diverse strategies, including defining kernel functions, partitioning genomic variance, or designing network architectures, according to the available biological information including gene ontology, transcriptomic, and GWAS data [124–126].

Although only a limited number of GS programs currently use DL, it is increasingly recognized as a promising approach for genetic prediction. First, DL models efficiently handle raw image data without preprocessing. Second, DL captures genetic diversity without additional predictor

terms, thus enabling the representation of non-additive effects and complex genetic relationships that is necessary for comprehensive genetic evaluation. Third, DL structures like CNNs acquire linkage disequilibrium throughout neighboring SNPs. Last, specific DL structures such as CNNs share parameters, thus lowering the number of parameters requiring estimation. However, using DL in GS is accompanied by caveats. DL is more susceptible to overfitting than typical statistical models, but this can be mitigated by using Bayesian methods. Moreover, implementing and optimizing DL models requires substantial experience because of the requirement for selecting various hyperparameters and the adjustment processes involved. To leverage DL effectively in GS, further iterative and collaborative examinations are required, along with the acquisition of more extensive datasets. These datasets should include phenotypic information and various omic, climate, and breeder experience data. Moreover, optimizing the DL model topology is necessary for designing an efficient framework for GS [120].

Applications of AI in gene editing

Traditional breeding methods such as mutagenesis, hybridization, and genetic engineering/transgenic breeding have made substantial contributions to improving crop yield and quality. However, they suffer from drawbacks such as extended breeding cycles, high randomness, low precision, incomplete gene function loss, and laborious screening processes [127,128]. The rise of genome sequencing technologies has opened avenues for precise and efficient molecular breeding, and has gained favor among breeders. Notably, the refinement of CRISPR/Cas9 technology has revolutionized breeding efforts and has significantly advanced research in crop quality enhancement [129].

The development of **gene editing** systems has accelerated progress in molecular biology and breeding. Site-directed nucleases (SDNs) are categorized into five classes – homing endonucleases (HEs), mega-nucleases (MNs), zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and CRISPR/Cas9 – and play a pivotal role in genome-editing technology [130,131]. HEs and MNs are rare endonucleases that recognize large DNA sequences, posing challenges in identifying their target sites [132]. ZFNs, the first generation of genome-editing nucleases, utilize small zinc-finger protein motifs, regulated by zinc ions, that bind to DNA in a sequence-specific manner. Unlike HEs and MNs, multiple ZFNs can be assembled into complexes, thereby enhancing DNA binding specificity. Similarly, TALENs were developed by fusing a TALE module to the FokI DNA cleavage domain, resulting in an effective programmable nuclease [133].

AI has been used to characterize structural information and optimize protein functionality. AlphaFold2 [134], a highly accurate protein structure prediction tool, is a remarkable application of AI in biology and now operates as a common 'infrastructure' for biological research. Another breakthrough technology, genome editing, has opened new doors for crop enhancement. AI-assisted genome editing and synthetic biology may allow the production of ideal plants through genetic modification [135]. However, producing breeding materials via genome editing may necessitate continuously optimized tools with higher efficiency. Huang *et al.* [99] utilized AlphaFold2-predicted protein structure data to unveil novel functional clusters of deaminases, and utilized this information to develop more efficient base editors. More recently, the first *de novo* generated gene editor, OpenCRISPR-1, was designed by a LLM trained on >1 million CRISPR operons [136], starting the new chapter in directly designing proteins with LLMs.

Moreover, designing genome editors with compact structures is essential for improved precision and efficiency in genetic manipulation. AI technologies can potentially revolutionize the design of compact and comprehensive genome-editing tools. Through protein structure prediction approaches, directly redesigning the proteins that underlie crucial agronomic traits is more straightforward. However, optimizing these proteins requires a large amount of specific training, and

ensuring that the redesigned protein performs adequately *in vivo* requires multidisciplinary collaboration and a comprehensive knowledge map. However, protein structure prediction tools can help to improve crop design, and it is reasonable to infer that AI-driven protein engineering will be helpful in crop enhancement in the future.

Concluding remarks

AI technologies could revolutionize public sector plant breeding programs (Box 1) such as those under the One CGIAR (Consultative Group for International Agricultural Research) program. Data-driven decentralized breeding programs such as these can predict crop performance better than conventional GS [137]. AI-enabled breeding platforms are accelerating the breeding process by deploying advanced computing and analysis algorithms. The promise of AI in gene discovery and allele mining is well defined; however, the true promise of AI is in assisting the biological design of future crop cultivars that are well suited for predicted environments. Further challenges will be to (i) simulate and characterize the diversity from genebanks for predicted environments to introduce novel traits into cultivars without linkage drag, (ii) construct the capacity and trained human

Box 1. Evolution of plant breeding

Plant breeding has witnessed great developments, especially during the past two decades. Historically, breeding eras can be divided into four different groups (Figure I) [142,143]. The first breeding era started with the start of agriculture ~10 000 years ago, where a major activity was domestication and selection by the common people. Later genomic studies identified a suite of traits selected during this era, referred as 'domestication syndrome', that were controlled by few genes with major effects. The foundation of the second breeding era was built on Mendelian genetics and Fisher's quantitative genetics by understanding the genetic basis of complex traits. These principles led to the development of experimental designs, conscious hybridization, pedigree selection, and the use of more robust phenotyping methods. A major breakthrough in this era was the so-called Green Revolution in cereals by developing high-yielding wheat and rice cultivars, and by heterosis breeding in many other crops. The third era was started with the advent of molecular markers to select phenotypes based on genomic information. This era revolutionized plant breeding by marker-assisted selection and mapping genes of complex traits for use in breeding. Modern phenotyping methods were designed for automated and robust data collection. Advances in genome sequencing and genotyping technologies led to the development of genomic selection (GS) which emerged as a rewarding breeding strategy that has been particularly significant in plant and livestock breeding, especially when dealing with complex traits. The latest era of plant breeding is influenced by big data because massive phenotypic and genotypic data generation is now very cheap and convenient. This era also witnessed the precise breeding by gene-editing technologies. AI is believed to be the indispensable tool to harness the benefits of these data technologies to accelerate genetic gain.

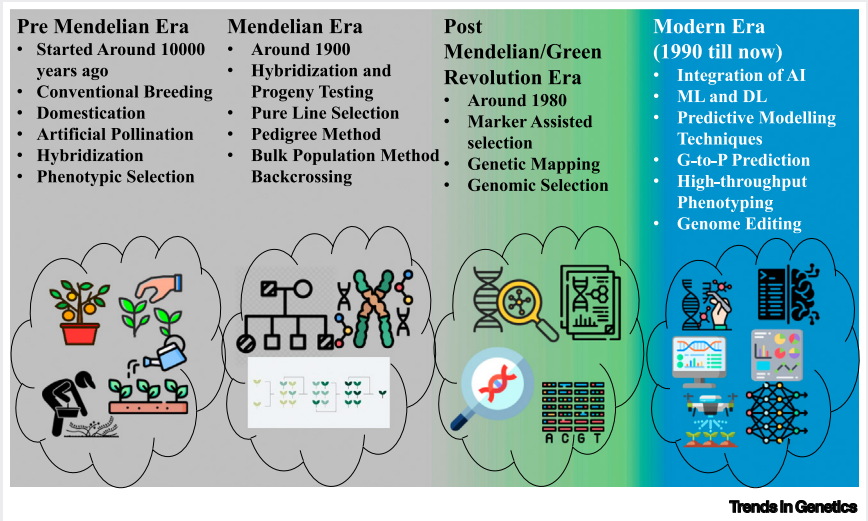

Figure I. The four plant breeding eras and the key breeding strategies for each era. Abbreviation: G-to-P, gene to phenotype.

Outstanding questions

How can AI be practically integrated into breeding programs to enable breeders and farmers to leverage advanced technologies for crop improvement?

How can AI techniques be tailored and optimized to extract meaningful insights and drive advances in plant phenomics research, especially when dealing with limited or small datasets, thus paving the way for more precise and targeted crop improvement strategies?'

How can AI-driven image analysis and computer vision techniques assist in identifying subtle phenotypic traits in plants, ultimately aiding breeders to select for complex traits such as drought tolerance or disease resistance?

What potential does AI have in utilizing ever-growing multi-omic data to facilitate the discovery of novel genetic resources and the improvement of rapid and precise GS?

How well do AI models trained on data from one environment or crop species generalize to other environments or species? What strategies can be used to improve the transferability and robustness of AI models across diverse genetic backgrounds and environmental conditions?

To what extent can plant breeding curriculums be designed to enable future plant breeders to fully harness the potential of new AI technologies?

How can we make AI technologies accessible and easy to use in public sector breeding programs by investing resources?

When will AI-enabled plant breeding products be available on the market?

Box 2. Tools for uncovering patterns in complex data

As we accumulate more and more data, the need grows for tools with which we can discover useful and possibly unexpected patterns in those data and report back to farmers in a timely and useful fashion. This can be done by using tools for data mining [144] and data visualization [145]. However, when working with big data, the Bonferroni principle [146] must be remembered: if you look hard enough for interesting patterns, you will find them. Many of these correlations may be spurious, and researchers need to ensure adequate data support. Owing to the enormity of some datasets, most available software cannot handle them entirely at once. Hence, it is crucial to devise an automated pipeline for extracting smaller subsets of data through data mining. This involves four main tasks: (i) identifying interesting relationships among variables in extensive datasets (i.e., association), (ii) segmenting datasets into discrete groups (i.e., clustering), (iii) assigning observations to these groups (i.e., classification), and (iv) predicting real-value outputs based on attributes of observational units (i.e., regression). For instance, Google developed MapReduce [147] to process vast amounts of raw web data into more manageable sets of key–value pairs. The MapReduce algorithm is actually a simple word-count algorithm that easily can be parallelized and scaled, and can be repurposed to process many different data types with very little effort. Association analysis in a data-mining context [148] is based principally on counting methods, which is not equivalent to genome-wide association analysis [149]. Clustering methods, exemplified by the work of [150], aim to categorize items into cohesive groups where items within a group share similarity, whereas those in separate groups exhibit dissimilarities. In partitional clustering, items are allocated to individual groups, as in the case of calling SNP genotypes. On the other hand, hierarchical clustering arranges items in a hierarchical or tree-like structure, allowing them to belong to multiple groups, a technique that is commonly used to illustrate species relationships. Classification models [142] use rules to assign individuals into classes based on their attributes, and typically involve training and validation steps. Numerous classification techniques are available, encompassing Bayesian belief networks, decision trees, nearest-neighbor classification, neural networks, rule-based classification, and support vector machines. All these methods are adaptable to MapReduce frameworks. The outcomes generated could then be utilized in regression models [151] that are commonly used in animal breeding for predicting real-value outputs such as breeding values and feed intake.

resources to efficiently utilize computational power for AI-enabled predictive breeding, (iii) offer multidisciplinary AI training to breeding teams engaged in designing future crop cultivars (Box 2), and (iv) produce a unified plant breeding cyber-infrastructure instead of **data islands** (see Outstanding questions). The utilization of AI has also demonstrated great promise in predicting the best cropping patterns and systems by integrating big data obtained from physical sensors, UAV platforms, and Internet of Things (IoT) devices under diverse genotype × environment × management practices [138]. This would complement the AI-based genetic innovations to obtain the required rate of genetic gain to meet the food and nutrition challenges of the next decade.

Acknowledgments

This work was supported by grants from the Sustainable Development International Cooperation Program from the Bill and Melinda Gates Foundation (2022YFAG1002), the National Natural Science Foundation of China (32261143757), Nanfan special project, CAAS (YBXM2407), the Key R&D Programs of Hainan Province (ZDYF2024XDNY210), the Bill and Melinda Gates Foundation (INV-030574) on mining useful alleles for climate change adaptation from CGIAR gene banks, and the Innovation Program of Chinese Academy of Agricultural Sciences (CAAS-CSIAF-202303).

Declaration of interests

The authors declare no competing interests.

References

1. Li, H. *et al.* (2018) Fast-forwarding genetic gain. *Trends Plant Sci.* 23, 184–186
2. Yang, W. *et al.* (2020) Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol. Plant* 13, 187–214
3. Agathokleous, E. *et al.* (2023) One hundred important questions facing plant science derived using a large language model. *Trends Plant Sci.* 29, 210–219
4. Armstrong, E.M. *et al.* (2023) One hundred important questions facing plant science: an international perspective. *New Phytol.* 238, 470–481
5. Yoosefzadeh-Najafabadi, M. *et al.* (2024) Machine learning-enhanced utilization of plant genetic resources. In *Sustainable Utilization and Conservation of Plant Genetic Diversity* (Al-Khayi, J.M. *et al.*, eds), pp. 619–639, Springer
6. McCouch, S. *et al.* (2020) Mobilizing crop biodiversity. *Mol. Plant* 13, 1341–1344
7. Juliana, P. *et al.* (2019) Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. *Nat. Genet.* 51, 1530–1539
8. Romero Navarro, J.A. *et al.* (2017) Application of genome wide association and genomic prediction for improvement of cacao

productivity and resistance to black and frosty pod diseases. *Front. Plant Sci.* 8, 1905
9. Milner, S.G. *et al.* (2019) Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* 51, 319–326
10. Lewin, H.A. *et al.* (2018) Earth BioGenome Project: sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* 115, 4325–4333
11. Varshney, R.K. *et al.* (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* 27, 522–530
12. Lasky, J.R. *et al.* (2015) Genome-environment associations in sorghum landraces predict adaptive traits. *Sci. Adv.* 1, e1400218
13. Dhondt, S. *et al.* (2013) Cell to whole-plant phenotyping: the best is yet to come. *Trends Plant Sci.* 18, 428–439
14. Tardieu, F. *et al.* (2017) Plant phenomics, from sensors to knowledge. *Curr. Biol.* 27, R770–R783
15. Fukatsu, T. *et al.* (2012) Field monitoring support system for the occurrence of *Leptocorisa chinensis* Dallas (Hemiptera: Alydidae) using synthetic attractants, field servers, and image analysis. *Comput. Electron. Agric.* 80, 8–16
16. Shibayama, M. *et al.* (2011) Estimating paddy rice leaf area index with fixed point continuous observation of near infrared reflectance using a calibrated digital camera. *Plant Prod. Sci.* 14, 30–46
17. Sadeghi-Tehran, P. *et al.* (2017) Automated method to determine two critical growth stages of wheat: heading and flowering. *Front. Plant Sci.* 8, 233406
18. Virlet, N. *et al.* (2016) Field Scanalyzer: an automated robotic field phenotyping platform for detailed crop monitoring. *Funct. Plant Biol.* 44, 143–153
19. Guo, Q. *et al.* (2018) Crop 3D – a LiDAR based platform for 3D high-throughput crop phenotyping. *Sci. China Life Sci.* 61, 328–339
20. Bai, G. *et al.* (2016) A multi-sensor system for high throughput field phenotyping in soybean and wheat breeding. *Comput. Electron. Agric.* 128, 181–192
21. Busemeyer, L. *et al.* (2013) BreedVision – a multi-sensor platform for non-destructive field-based phenotyping in plant breeding. *Sensors* 13, 2830–2847
22. Yue, J. *et al.* (2019) Estimate of winter-wheat above-ground biomass based on UAV ultrahigh-ground-resolution image textures and vegetation indices. *ISPRS J. Photogramm. Remote Sens.* 150, 226–244
23. Hassan, M.A. *et al.* (2021) Quantifying senescence in bread wheat using multispectral imaging from an unmanned aerial vehicle and QTL mapping. *Plant Physiol.* 187, 2623–2636
24. Li, L. *et al.* (2022) Development of image-based wheat spike counter through a Faster R-CNN algorithm and application for genetic studies. *Crop J.* 10, 1303–1311
25. Bodner, G. *et al.* (2018) Hyperspectral imaging: a novel approach for plant root phenotyping. *Plant Methods* 14, 84
26. Kaissis, G. *et al.* (2021) End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intel.* 3, 473–484
27. Gaggion, N. *et al.* (2021) ChronoRoot: high-throughput phenotyping by deep segmentation networks reveals novel temporal parameters of plant root system architecture. *GigaScience* 10, giab052
28. Smith, A.G. *et al.* (2020) Segmentation of roots in soil with U-Net. *Plant Methods* 16, 13
29. Yasrab, R. *et al.* (2019) RootNav 2.0: deep learning for automatic navigation of complex plant root architectures. *GigaScience* 8, giz123
30. Ubbens, J.R. and Stavness, I. (2017) Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks. *Front. Plant Sci.* 8, 276225
31. Geetharamani, G. and Pandian, A. (2019) Identification of plant leaf diseases using a nine-layer deep convolutional neural network. *Comput. Electr. Eng.* 76, 323–338
32. Nevavuori, P. *et al.* (2019) Crop yield prediction with deep convolutional neural networks. *Comput. Electron. Agric.* 163, 104859
33. Nagasubramanian, K. *et al.* (2019) Plant disease identification using explainable 3D deep learning on hyperspectral images. *Plant Methods* 15, 98
34. Shu, M. *et al.* (2021) The application of UAV-based hyperspectral imaging to estimate crop traits in maize inbred lines. *Plant Phenom.* 2021, 9890745
35. Tauro, F. *et al.* (2022) Latent heat flux variability and response to drought stress of black poplar: A multi-platform multi-sensor remote and proximal sensing approach to relieve the data scarcity bottleneck. *Remote Sens. Environ.* 268, 112771
36. Ghosal, S. *et al.* (2018) An explainable deep machine vision framework for plant stress phenotyping. *Proc. Natl. Acad. Sci.* 115, 4613–4618
37. Qiu, Z. *et al.* (2021) Estimation of nitrogen nutrition index in rice from UAV RGB images coupled with machine learning algorithms. *Comput. Electron. Agric.* 189, 106421
38. Shi, P. *et al.* (2021) Rice nitrogen nutrition estimation with RGB images and machine learning methods. *Comput. Electron. Agric.* 180, 105860
39. Gené-Mola, J. *et al.* (2020) Fruit detection, yield prediction and canopy geometric characterization using LiDAR with forced air flow. *Comput. Electron. Agric.* 168, 105121
40. Raza, S.-e.-A. *et al.* (2014) Automatic detection of regions in spinach canopies responding to soil moisture deficit using combined visible and thermal imagery. *PLoS One* 9, e97612
41. Maimaitijiang, M. *et al.* (2020) Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sens. Environ.* 237, 111599
42. Yoosefzadeh-Najafabadi, M. *et al.* (2021) Using hybrid artificial intelligence and evolutionary optimization algorithms for estimating soybean yield and fresh biomass using hyperspectral vegetation indices. *Remote Sens.* 13, 2555
43. Maimaitijiang, M. *et al.* (2017) Unmanned aerial system (UAS)-based phenotyping of soybean using multi-sensor data fusion and extreme learning machine. *ISPRS J. Photogramm. Remote Sens.* 134, 43–58
44. Harfouche, A.L. *et al.* (2019) Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. *Trends Biotechnol.* 37, 1217–1235
45. Streich, J. *et al.* (2020) Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the United Nations sustainable development goals? *Curr. Opin. Biotechnol.* 61, 217–225
46. Demirci, M. *et al.*, Comparative dissolved gas analysis with machine learning and traditional methods, *In: 2021 3rd International Congress on Human–Computer Interaction, Optimization and Robotic Applications (HORA)*, 2021, IEEE, 1–6
47. Sartor, R.C. *et al.* (2019) Identification of the expressome by machine learning on omics data. *Proc. Natl. Acad. Sci.* 116, 18119–18125
48. Uygun, S. *et al.* (2019) Cis-regulatory code for predicting plant cell-type transcriptional response to high salinity. *Plant Physiol.* 181, 1739–1751
49. Li, Z. *et al.* (2020) Identification and biotechnical potential of a Gcn5-related N-acetyltransferase gene in enhancing microalgal biomass and starch production. *Front. Plant Sci.* 11, 544827
50. Varala, K. *et al.* (2018) Temporal transcriptional logic of dynamic regulatory networks underlying nitrogen signaling and use in plants. *Proc. Natl. Acad. Sci.* 115, 6494–6499
51. Meena, M. *et al.* (2021) Production of biofuels from biomass: predicting the energy employing artificial intelligence modelling. *Bioresour. Technol.* 340, 125642
52. Poplin, R. *et al.* (2018) A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* 36, 983–987
53. Ahsan, M.U. *et al.* (2021) NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks. *Genome Biol.* 22, 261
54. Zheng, Z. *et al.* (2022) Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat. Comput. Sci.* 2, 797–803
55. Shafin, K. *et al.* (2021) Haplotype-aware variant calling enables high accuracy in nanopore long-reads using deep neural networks. *BioRxiv*, Published online March 5, 2021. https://doi.org/10.1101/2021.03.04.433952
56. Luo, R. *et al.* (2019) A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nat. Commun.* 10, 998
57. Schrider, D.R. and Kern, A.D. (2018) Supervised machine learning for population genetics: a new paradigm. *Trends Genet.* 34, 301–312

58. Vara, C. *et al.* (2019) Three-dimensional genomic structure and cohesin occupancy correlate with transcriptional activity during spermatogenesis. *Cell Rep.* 28, 352–367
59. Bansal, A. and Srivastava, P.A. (2018) Transcriptomics to metabolomics: a network perspective for big data. In *Applying Big Data Analytics in Bioinformatics and Medicine* (Lytras, M.D. and Papadopoulou, P., eds), pp. 188–206, IGI Global
60. Kanaya, S. *et al.* (2015) Big data and network biology 2015. *Biomed. Res. Int.* 2015, 604523
61. Johnson, D.S. *et al.* (2007) Genome-wide mapping of in vivo protein–DNA interactions. *Science* 316, 1497–1502
62. Boyle, A.P. *et al.* (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132, 311–322
63. Buenrostro, J.D. *et al.* (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–1218
64. Raj, A. *et al.* (2015) msCentipede: modeling heterogeneity across genomic sites and replicates improves accuracy in the inference of transcription factor binding. *PLoS One* 10, e0138030
65. Sherwood, R.I. *et al.* (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* 32, 171–178
66. Demirci, S. *et al.* (2018) DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. *Plant J.* 95, 686–699
67. Schrag, T.A. *et al.* (2018) Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. *Genetics* 208, 1373–1385
68. Bourgeois, Y. *et al.* (2018) Genome-wide scans of selection highlight the impact of biotic and abiotic constraints in natural populations of the model grass *Brachypodium distachyon*. *Plant J.* 96, 438–451
69. Yip, K.Y. *et al.* (2013) Machine learning and genome annotation: a match meant to be? *Genome Biol.* 14, 205
70. Fei, S. *et al.* (2022) Application of multi-layer neural network and hyperspectral reflectance in genome-wide association study for grain yield in bread wheat. *Field Crop Res.* 289, 108730
71. Li, L. *et al.* (2023) UAV-based RGB imagery and ground measurements for high-throughput phenotyping of senescence and QTL mapping in bread wheat. *Crop Sci.* 63, 3292–3309
72. Rairdin, A. *et al.* (2022) Deep learning-based phenotyping for genome wide association studies of sudden death syndrome in soybean. *Front. Plant Sci.* 13, 966244
73. Montesinos-López, A. *et al.* (2024) Deep learning methods improve genomic prediction of wheat breeding. *Front. Plant Sci.* 15, 1324090
74. Montesinos-López, A. *et al.* (2023) Multimodal deep learning methods enhance genomic prediction of wheat breeding. *G3: Genes Genomes Genet.* 13, jkad045
75. Togninalli, M. *et al.* (2023) Multi-modal deep learning improves grain yield prediction in wheat breeding by fusing genomics and phenomics. *Bioinformatics* 39, btad336
76. Maes, W.H. and Steppe, K. (2019) Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture. *Trends Plant Sci.* 24, 152–164
77. Orozco-Arias, S. *et al.* (2023) Inpactor2: a software based on deep learning to identify and classify LTR-retrotransposons in plant genomes. *Brief. Bioinform.* 24, bbac511
78. Zhang, R.-G. *et al.* (2022) TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic. Res.* 9, uhac017
79. Wang, Y. *et al.* (2020) Synthetic promoter design in *Escherichia coli* based on a deep generative network. *Nucleic Acids Res.* 48, 6403–6412
80. Gupta, A. and Zou, J. (2019) Feedback GAN for DNA optimizes protein functions. *Nat. Mach. Intel.* 1, 105–111
81. Yasmeen, E. *et al.* (2023) Designing artificial synthetic promoters for accurate, smart, and versatile gene expression in plants. *Plant Commun.* 4, 100558
82. Bargsten, J.W. *et al.* (2014) Prioritization of candidate genes in QTL regions based on associations between traits and biological processes. *BMC Plant Biol.* 14, 330
83. Liu, S. *et al.* (2017) A computational interactome for prioritizing genes associated with complex agronomic traits in rice (*Oryza sativa*). *Plant J.* 90, 177–188
84. Lin, F. *et al.* (2019) QTG-Finder: a machine-learning based algorithm to prioritize causal genes of quantitative trait loci in *Arabidopsis* and rice. *G3: Genes Genomes Genet.* 9, 3129–3138
85. Gao, S. *et al.* (2023) Mining salt stress-related genes in *Spartina alterniflora* via analyzing co-evolution signal across 365 plant species using phylogenetic profiling. *Abiotech* 4, 291–302
86. Yang, M. *et al.* (2023) Deep learning-enabled discovery and characterization of HKT genes in Spartina alterniflora. *Plant J.* 116, 690–705
87. Huang, Z. *et al.* (2024) Exploring salt tolerance mechanisms using machine learning for transcriptomic insights: case study in *Spartina alterniflora*. *Hortic. Res.* 11, uhae082
88. Denyer, T. *et al.* (2019) Spatiotemporal developmental trajectories in the *Arabidopsis* root revealed using high-throughput single-cell RNA sequencing. *Dev. Cell* 48, 840–852
89. Jean-Baptiste, K. *et al.* (2019) Dynamics of gene expression in single root cells of *Arabidopsis thaliana*. *Plant Cell* 31, 993–1011
90. Luecken, M.D. and Theis, F.J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* 15, e8746
91. Toubiana, D. *et al.* (2019) Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. *Commun. Biol.* 2, 214
92. McLoughlin, F. *et al.* (2018) Maize multi-omics reveal roles for autophagic recycling in proteome remodelling and lipid turnover. *Nat. Plants* 4, 1056–1070
93. Ma, C. *et al.* (2014) Machine learning-based differential network analysis: a study of stress-responsive transcriptomes in *Arabidopsis*. *Plant Cell* 26, 520–537
94. Gupta, C. *et al.* (2021) Using network-based machine learning to predict transcription factors involved in drought resistance. *Front. Genet.* 12, 652189
95. Sun, S. *et al.* (2021) Revisiting genome-wide association studies from statistical modelling to machine learning. *Brief. Bioinform.* 22, bbaa263
96. Broekema, R. *et al.* (2020) A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.* 10, 190221
97. Lin, F. *et al.* (2020) QTG-Finder2: a generalized machine-learning algorithm for prioritizing QTL causal genes in plants. *G3: Genes Genomes Genet.* 10, 2411–2421
98. Kolosov, N. *et al.* (2021) Prioritization of disease genes from GWAS using ensemble-based positive-unlabeled learning. *Eur. J. Hum. Genet.* 29, 1527–1535
99. Huang, J. *et al.* (2023) Discovery of deaminase functions by structure-based protein clustering. *Cell* 186, 3182–3195
100. Yang, M. *et al.* (2023) Deep learning-enabled discovery and characterization of HKT genes in *Spartina alterniflora*. *Plant J.* 116, 690–705
101. Ramstein, G.P. *et al.* (2019) Breaking the curse of dimensionality to identify causal variants in Breeding 4. *Theor. Appl. Genet.* 132, 559–567
102. Ramstein, G.P. and Buckler, E.S. (2022) Prediction of evolutionary constraint by genomic annotations improves functional prioritization of genomic variants in maize. *Genome Biol.* 23, 183
103. Kang, M. *et al.* (2022) A roadmap for multi-omics data integration using deep learning. *Brief. Bioinform.* 23, bbab454
104. Xu, Y. and Crouch, J.H. (2008) Marker-assisted selection in plant breeding: from publications to practice. *Crop Sci.* 48, 391–407
105. Budhlakoti, N. *et al.* (2022) Genomic selection: a tool for accelerating the efficiency of molecular breeding for development of climate-resilient crops. *Front. Genet.* 13, 832153
106. Meuwissen, T.H. *et al.* (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829
107. Long, N. *et al.* (2011) Application of support vector regression to genome-assisted prediction of quantitative traits. *Theor. Appl. Genet.* 123, 1065–1074
108. Gianola, D. *et al.* (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet.* 12, 87
109. Holliday, J.A. *et al.* (2012) Predicting adaptive phenotypes from multilocus genotypes in Sitka spruce (*Picea sitchensis*) using random forest. *G3: Genes Genomes Genet.* 2, 1085–1093

110. Yoosefzadeh-Najafabadi, M. *et al.* (2022) Optimizing genomic selection in soybean: an important improvement in agricultural genomics. *Heliyon* 8, e11873
111. Greener, J.G. *et al.* (2022) A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* 23, 40–55
112. Gianola, D. (2013) Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 194, 573–596
113. Gianola, D. *et al.* (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173, 1761–1776
114. González-Camacho, J.M. *et al.* (2018) Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. *Plant Genome* 11, 170104
115. Noble, W.S. (2006) What is a support vector machine? *Nat. Biotechnol.* 24, 1565–1567
116. Rosado, R.D.S. *et al.* (2020) Artificial neural networks in the prediction of genetic merit to flowering traits in bean cultivars. *Agriculture* 10, 638
117. Montesinos-López, O.A. *et al.* (2018) Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. *G3: Genes Genomes Genet.* 8, 3829–3840
118. Montesinos-López, O.A. *et al.* (2019) A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding. *G3: Genes Genomes Genet.* 9, 601–618
119. LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
120. Montesinos-López, O.A. *et al.* (2021) A review of deep learning applications for genomic selection. *BMC Genomics* 22, 19
121. Abdollahi-Arpanahi, R. *et al.* (2020) Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genet. Sel. Evol.* 52, 12
122. Pook, T. *et al.* (2020) Using local convolutional neural networks for genomic prediction. *Front. Genet.* 11, 561497
123. Wang, K. *et al.* (2023) DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Mol. Plant* 16, 279–293
124. Farooq, M. *et al.* (2021) Prior biological knowledge improves genomic prediction of growth-related traits in *Arabidopsis thaliana*. *Front. Genet.* 11, 609117
125. Wu, C. *et al.* (2024) A transformer-based genomic prediction method fused with knowledge-guided module. *Brief. Bioinform.* 25, bbad438
126. Li, H. *et al.* (2024) Smart Breeding Platform: a web-based tool for high-throughput population genetics, phenomics, and genomic selection. *Mol. Plant* 17, 677–681
127. Labroo, M.R. *et al.* (2021) Heterosis and hybrid crop breeding: a multidisciplinary review. *Front. Genet.* 12, 643761
128. Yali, W. and Mitiku, T. (2022) Mutation breeding and its importance in modern plant breeding. *J. Plant Sci.* 10, 64–70
129. Wang, T. *et al.* (2021) CRISPR/Cas9-mediated gene editing revolutionizes the improvement of horticulture food crops. *J. Agric. Food Chem.* 69, 13260–13269
130. Adli, M. (2018) The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* 9, 1911
131. Gaj, T. *et al.* (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* 31, 397–405
132. Rasheed, A. *et al.* (2021) A critical review: recent advancements in the use of CRISPR/Cas9 technology to enhance crops and alleviate global food crises. *Curr. Issues Mol. Biol.* 43, 1950–1976
133. Christian, M. *et al.* (2010) Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* 186, 757–761
134. Jumper, J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589
135. Huang, X. *et al.* (2022) The integrated genomics of crop domestication and breeding. *Cell* 185, 2828–2839
136. Ruffolo, J.A. *et al.* (2024) Design of highly functional genome editors by modeling the universe of CRISPR-Cas sequences. *BioRxiv*, Published online April 22, 2024. https://doi.org/10.1101/2024.04.22.590591
137. de Sousa, K. *et al.* (2021) Data-driven decentralized breeding increases prediction accuracy in a challenging crop production environment. *Commun. Biol.* 4, 944
138. Jung, J. *et al.* (2021) The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems. *Curr. Opin. Biotechnol.* 70, 15–22
139. Ghosal, S. *et al.* (2019) A weakly supervised deep learning framework for sorghum head detection and counting. *Plant Phenom.* 2019, 1525874
140. Vaswani, A. *et al.* (2017) Attention is all you need. *Adv. Neural Inf. Proces. Syst.* 30,
141. He, K. *et al.* (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, IEEE
142. Hand, D.J. and Henley, W.E. (1997) Statistical classification methods in consumer credit scoring: a review. *J R Stat. Soc. Ser. A Stat. Soc.* 160, 523–541
143. Wallace, J. *et al.* (2018) On the road to Breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Annu. Rev. Genet.* 52, 421–444
144. Tan, K.C. *et al.* (2005) A distributed evolutionary classifier for knowledge discovery in data mining. In *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* (35), pp. 131–142, IEEE
145. Cole, J.B. and VanRaden, P.M. (2010) Visualization of results from genomic evaluations. *J. Dairy Sci.* 93, 2727–2740
146. Shaffer, J.P. (1995) Multiple hypothesis testing. *Annu. Rev. Psychol.* 46, 561–584
147. Lin, J. and Dyer, C. (2022) *Data-Intensive Text Processing with MapReduce*, Springer
148. Aggarwal, C.C. and Yu, P.S. (1998) Online generation of association rules. In *Proceedings 14th International Conference on Data Engineering*, pp. 402–411, IEEE
149. Maltecca, C. *et al.* (2011) A genome-wide association study of direct gestation length in US Holstein and Italian Brown populations. *Anim. Genet.* 42, 585–591
150. Everitt, B.S. *et al.* (2001) *Cluster Analysis*, Arnold
151. Gorjanc, G. *et al.* (2015) Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genet. Sel. Evol.* 47, 12