Trends in
Plant Science

CellPress

Review

Machine learning bridges omics sciences and plant breeding

Jun Yan [1,2] and Xiangfeng Wang [1,2,*]

Some of the biological knowledge obtained from fundamental research will be implemented in applied plant breeding. To bridge basic research and breeding practice, machine learning (ML) holds great promise to translate biological knowledge and omics data into precision-designed plant breeding. Here, we review ML for multi-omics analysis in plants, including data dimensionality reduction, inference of gene-regulation networks, and gene discovery and prioritization. These applications will facilitate understanding trait regulation mechanisms and identifying target genes potentially applicable to knowledge-driven molecular design breeding. We also highlight applications of deep learning in plant phenomics and ML in genomic selection-assisted breeding, such as various ML algorithms that model the correlations among genotypes (genes), phenotypes (traits), and environments, to ultimately achieve data-driven genomic design breeding.

Highlights

To bridge the gaps between basic research and breeding practice in plants, machine learning (ML) holds great promise to integrate biological knowledge and omics data, to ultimately achieve precision-designed plant breeding.

Recent applications of ML in plant research and breeding include data dimensionality reduction, inference of gene-regulation networks, gene discovery and prioritization, plant phenomics analysis, and genomic prediction of plant phenotypes.

High-dimensional biology denotes the integration and analysis of macroscale to microscale biological data, elevating the chance of identifying trait-causative genes utilizable for knowledge-driven molecular design breeding.

In the era of big data, ML is capable of modeling the complex relations of genotypic, phenotypic, and environmental data collected from breeding practice, to achieve data-driven genomic design breeding.

Machine learning translates knowledge and data into breeding

Over recent decades, knowledge acquired from basic research in plant biology has greatly expedited the progress of plant breeding and accelerated crop improvement, (i.e. achieving higher yields or better stress tolerance) [1]. However, existing gaps between basic research and breeding practice in plants still have to be overcome if we are to ultimately achieve the goal of precision-designed plant breeding. As a subfield of artificial intelligence technology, **ML** (see Glossary) holds great promise, because of its superior ability and flexibility for integrating various forms of biological knowledge and omics data. ML may translate biological knowledge and data into precision-designed plant breeding, mainly through two pathways (Figure 1, Key figure). One path is to facilitate omics sciences in plant biology and expedite the discovery of agronomically utilizable genes and mutations to achieve knowledge-driven molecular design breeding (Figure 1A). The other path is to directly apply ML techniques in commercial breeding programs to construct a variety of predictive models for achieving data-driven genomic design breeding (Figure 1B). These two paths have been incorporated into and are playing essential roles in modern breeding pipelines for which selection of the proper path depends on the number of genes or loci related to a trait. For example, quantitative traits are determined mostly by genetic background (i.e., yield, biomass, or environmental fitness); therefore, data-driven modeling is usually adopted to infer the correlation between phenotypes and whole-genome markers. Polygenic traits are determined by genetic foreground (specific genes with major effects, i.e., disease resistance); therefore functions of causal genes have to be explicitly characterized, so that beneficial alleles can be pyramided. Whereas for single-gene traits genome editing is the promptest way to artificially create a mutation to alter the trait. As long as sufficient knowledge and data are accumulated in plant biology and breeding, ML can facilitate precision-designed breeding.

Although most ML methods and tools can be used in both animals and plants [2,3], we focus mainly on recent developments in the application of ML to plant research and breeding in this

[1]National Maize Improvement Center, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100094, China
[2]Frontiers Science Center for Molecular Design Breeding, China Agricultural University, Beijing 100094, China

*Correspondence:
xwang@cau.edu.cn (X. Wang).

© 2022 Elsevier Ltd. All rights reserved.

review. We first introduce the family of ML models, followed by reviewing advanced ML methods utilized for data dimensionality reduction (DR), inference of gene-regulation networks (GRNs), gene discovery, and prioritization. We then review applications of **deep learning (DL)** in plant phenomics and ML methods employed in **genomic selection (GS)**-assisted breeding. We finally discuss current challenges of ML in plant research and potential future solutions.

The family of ML algorithms

Rapid advancement of high-throughput omics technologies has seen plant biology enter the era of **high-dimensional biology (HDB)** [4] (Figure 2A). However, genomic, transcriptomic, proteomic, metabolomic, and phenomic datasets are highly heterogeneous and complex, posing unprecedented challenges for data integration [5]. Multi-omics data are also extremely large, highly dimensional, and noisy, beyond the capability of conventional, model-based statistical analysis. Therefore, analytical methodologies to cope with high-dimensional biological datasets are eagerly anticipated. ML has been widely utilized in big data analytics in biology, due to its superior capability of dealing with large-scale, nonstructured, and complex datasets [6]. As a data-driven paradigm, it does not require statistical assumption and thus greatly reduces human effort in understanding the data characteristics [7].

In general, ML algorithms, whether solving a classification or regression problem, fall into three main classes: '**supervised learning**', '**unsupervised learning**', and '**semi-supervised learning**' (Figure 2B). The most frequently used supervised learning algorithms in biology are support vector machine (SVM), random forest (RF), artificial neural network (ANN), Bayesian approaches, and penalized regressions such as least absolute shrinkage and selection operator, ridge regression, and elastic net [8]. The unsupervised learning algorithms are mostly used for sample classification and DR, such as K-means and **principal component analysis (PCA)**, respectively [6]. Semi-supervised learning is a hybrid of the aforementioned two classes [9]. Notably, the recently emerged ML paradigm of DL has revolutionized the fields of computer vision, speech recognition, and natural language processing [10]. It also has become a popular ML method for solving problems in biology [11]. Among the DL family, convolutional neural network (CNN), recurrent neural network (RNN), generative adversarial network (GAN), graph convolutional network (GCN), long short-term memory (LSTM), **transfer learning**, as well as **contrastive learning**, a recent self-supervised learning method represented by SimCLR (a simple framework for contrastive learning of visual representations), MoCo (momentum contrast for unsupervised visual representation learning), and BYOL (bootstrap your own latent) [12], have been successfully implemented in many fields of life sciences and health care [13].

ML and DL will likely play critical roles in exploiting the rapidly accumulating multi-omics data in plant biology and ultimately applying the resulting knowledge to plant breeding. As illustrated in Figure 2C, ML has a wide spectrum of applications in large-scale omics research, including prediction of genetic elements such as transcription factors (TFs) and non-coding RNAs, prediction of molecular structures such as alternative splicing and protein structures, and prediction of regulatory elements such as promoters, enhancers, TF-binding sites, and epigenetically modified regions.

Data DR

The high dimensionality of multi-omics data may lead to the so-called '**curse of dimensionality**', as exemplified by the case study using multi-omics data association studies for exploiting maize (*Zea mays*) germplasm (Box 1). Therefore, application of feature selection and/or DR is an essential step prior to training an ML model, especially when the feature set is far larger than the sample set [14]. Whereas feature selection based on biological indicators requires expertise to remove

Glossary

Contrastive learning: a subset of self-supervised deep learning methods that has been successfully applied for object recognition in the field of computer vision.

Curse of dimensionality: means that feature space exponentially increases along with increased dimensionality, which may result not only in a dramatic increase of computational cost, but also in problematic overfitting, since the model may learn incorrect features from training samples.

Deep learning (DL): constructs large-scale, multilayer artificial neural networks using a framework of either supervised or unsupervised learning.

Density-based spatial clustering of applications with noise (DBSCAN): a density-based nonparametric clustering algorithm widely used in ML, which separates clusters of high-density points from clusters of low-density points.

Ensemble learning: the strategy of assembling multiple weak learners and combining their outputs to enhance predictability.

Genomic selection (GS): the utilization of whole-genome genetic markers or SNPs to predict the phenotype of an individual from its genotype, in order to assist selecting individuals with high breeding values.

High-dimensional biology (HDB): the integrative analysis of multiple types of omics (multi-omics) data generated from high-throughput platforms.

Linear discriminant analysis (LDA): a supervised method; its rationale is to make related samples as compact as possible after projecting the data to a low-dimensional space.

Machine learning (ML): a data-driven paradigm in which computational algorithms can learn from data themselves, without either relying on statistical assumptions or being explicitly programmed.

Manifold learning: a class of nonlinear DR algorithms aiming to embed high-dimensional data into low-dimensional space, but maximally preserving the geometric properties of data manifolds.

Multifactor dimensionality reduction (MDR): a nonparametric approach that is used to characterize combinatory influence of multiple factors on outcome; it is commonly used to detect epistatic interactions among genes in biology.

redundant and noisy features. ML-based approaches such as wrapper feature selection methods (e.g., forward feature selection, backward feature selection, recursive feature elimination) and intrinsic feature selection methods (e.g., decision trees, regularization models) do not require domain knowledge, but may lead to the loss of important features [15]. DR relies on a spectrum of ML algorithms and provides an alternative way for the extraction of features. Two of the most widely used linear DR algorithms are PCA [16] and **linear discriminant analysis (LDA)** [17]. In biology, while PCA is widely applied for extracting and visualizing sample relatedness based on population genotype data and omics data, LDA is commonly used in feature extraction and classification tasks. For instance, LDA has been used to classify wheat varieties exhibiting different efficiency of nutrient uptake by extracting features from the images of wheat root systems [18].

Certain algorithms are designed for learning the nonlinear geometry of high-dimensional data, and **multifactor dimensionality reduction (MDR)** is among the most highly used in biology [19]. In barley, it was adopted for inferring epistatic interactions between multiple quantitative trait loci (QTLs) and computing the joint effects of multiple SNPs associated with a trait by converting multiple attributes into one [20]. Non-negative matrix factorization (NMF) is another nonlinear method that can factorize a non-negative matrix $A_{(m \times n)}$ into a feature matrix $W_{(m \times k)}$ and a coefficient matrix $H_{(n \times k)}$, in which k is the low rank approximation of A [$k \leq \min(m, n)$] [21]. The main objective of NMF is to reduce data dimensionality by reducing large numbers of features. It has been applied to classification of expression data in arabidopsis (*Arabidopsis thaliana*) and maize, with thousands of genes clustered into small sets of 'metagenes' exhibiting similar expression patterns in samples [22,23]. As previously mentioned, single-cell RNA sequencing (scRNA-seq) data exhibits extremely high dimensionality and nonlinear characteristics. Multiple **manifold learning** techniques have been introduced for applying DR to scRNA-seq data, including t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), and potential of heat-diffusion for affinity-based transition embedding (PHATE) [24]. These methods are effective for capturing nonlinear relationships of different populations of cell types embedded in tens of thousands of transcriptomes at a single-cell level by visualizing the structure in 2D or 3D space [25].

In plant germplasm research, these DR algorithms are also widely used for inferring and visualizing the genetic structure of a population based on genotypic data, an essential step prior to the application of genome-wide association studies (GWAS) and GS [26,27]. In the recently released software package MODAS (multi-omics data association analysis), Liu *et al.* developed a novel method for applying DR to genotypic data to accelerate association analysis [28]. MODAS first applies a nonlinear clustering algorithm called **DBSCAN (density-based spatial clustering of applications with noise)** to identify haplotype blocks and then employs PCA to generate a pseudo-genotype index file. The index represents a highly simplified atlas of genomic variations, which may be used for population structure analysis or association analysis with ultrahigh computing efficiency.

Inference of GRNs

The major objective of multi-omics analysis is to reconstruct GRNs, and ChIP-seq experiments are the most straightforward approach for profiling the binding sites of TFs and target genes. Compared with humans and model animals, ChIP-seq data is very scarce in plants, with only a limited number of TFs having been characterized in model plants such as arabidopsis, rice, and maize [29]. Thus, inference of GRNs in plants mostly relies on expression data, from which a potential regulatory relationship between two genes is inferred using correlation-based methods and **mutual information (MI)** algorithm [30]. Pearson's correlation coefficient and Spearman's rank correlation coefficient were among the earliest methods used in gene coexpression analysis

Mutual information (MI): a quantity developed in information theory, which measures the degree of 'mutual dependence' between two random variables.
Positive-unlabeled (PU) learning: a semi-supervised approach that uses only positive and unlabeled samples for prediction (e.g., first identifies a subset of reliable negative samples from all of the unlabeled samples and then utilizes labeled positive and predicted negative samples as two contrasting sets for classifying the remaining unlabeled samples).
Principal component analysis (PCA): maps high-dimensional data to a low-dimensional space through linear projection in an unsupervised manner, from which the maximum variance for each projected dimension is computed.
Semi-supervised learning: first uses labeled samples to predict a part of the unlabeled samples and then merges the two sets of samples to train a new model to predict the remaining unlabeled samples.
Supervised learning: an ML model is trained using labeled samples to predict unlabeled samples, in which labels are also called 'features' or 'predictors' that can be either categorical or continuous variables.
Transfer learning: a framework for transferring the network structure and/or parameters of a pretrained DL model learned from a dataset to another model applied to a new dataset.
Unsupervised learning: algorithms that do not require labels and instead learn patterns directly from the data.

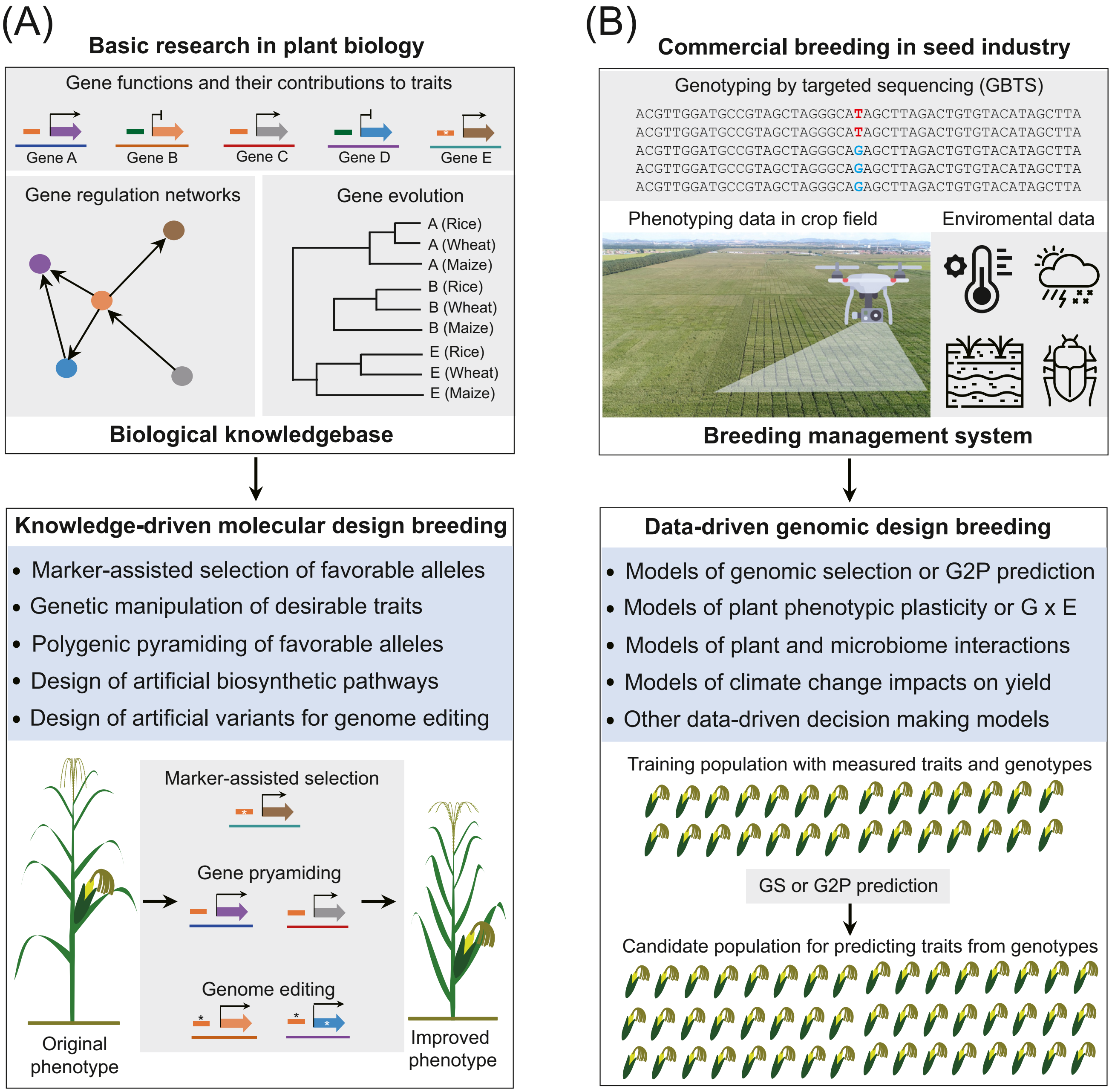
[31]. The algorithms CLR (context likelihood of relatedness), ARACNE (algorithm for the reconstruction of accurate cellular networks), and TGMI (triple-gene mutual interaction) all utilize MI as a measurement to represent the potential of a regulatory relationship between two genes based on gene coexpression patterns [32–34]. However, both correlation- and MI-based methods fail to distinguish regulatory direction and are unable to consider temporal delay between expression of genes [35]. Probabilistic graphical models (PGM) solves this problem by incorporating prior probability distribution of temporal, spatial, or environmental information, as used by the tool GENIST [30]. GENIST utilizes dynamic Bayesian networks to infer GRN by combining spatial with temporal expression data in arabidopsis root stem cells [36]. Another example is JRmGRN, based on a Gaussian graphical model, which is capable of jointly reconstructing multiple GRNs to identify common hubs or condition-specific genes using arabidopsis data collected from different tissues or under different light conditions [37]. However, one main limitation of PGM-based methods is that they require expression data at high spatiotemporal resolution to ensure the accuracy of inferred GRNs.

ML-based approaches to GRN inference have emerged recently and gained widespread attention due to their flexibility and superior performance [38]. GENIE3 is the most popular method used in plants and mainly utilizes tree-based models, such as RF and extreme randomized tree, to infer GRNs from expression data [39]. In maize, GENIE3 was used to generate 45 maize GRNs by integrating publicly available expression datasets, from which distinct association patterns of TFs and target genes in different populations were revealed [40]. In wheat, GENIE3 was used to infer a GRN-regulating senescence and identify key regulators that were functionally validated for the first time in a polyploid species [41]. BTNET is another tree-based tool, employing adaptive boosting and gradient boosting algorithms, with a focus on inferring GRNs from time-series expression data [42]. In addition to these decision-tree-based methods, Beacon utilizes an SVM algorithm for context-specific inference of GRNs at specific stages during the seed development of arabidopsis [43]. As scRNA-seq has become a popular technique in omics studies, there is a need for tools capable of inferring cell-specific GRNs. GRNBoost2 is one of these, developed on the framework of GENIE3 with stochastic gradient boosting machine (GBM) algorithms [44]. To improve performance when working with the high dimensionality of gene expression data at a single-cell scale, an analytical pipeline of SCENIC was developed by integrating multiple tools, including GRNBoost2, cisTarget, AUCell, and nonlinear projection methods, to visualize cell populations [45]. A benchmark test of SCENIC showed that the pipeline can analyze a dataset comprising 10 000 genes and 50 000 cells within 2 h.

Strictly speaking, GRNs inferred from gene expression data may not be regarded as *bona fide* regulatory relationships of genes. However, with the availability of multi-omics data, multiple layers of genetic information may be integrated to improve the power of GRN inference [46]. One example is iDREM (interactive dynamic regulatory events miner), designed to integrate static and time-series multi-omics data based on a hidden Markov model [47]. iDREM has been used in plants to reconstruct temporal GRNs and identify key regulators involved in the responses to biotic and abiotic stress using transcriptomic, proteomic, and epigenomic data [48]. Similarly, RTP-STAR (regression tree pipeline for spatial, temporal, and replicate data), developed on the framework of GENIE3, was used to integrate transcriptomic, proteomic, and phosphorproteomic data in arabidopsis to infer GRNs in response to jasmonic acid [49]. Recently, an improved version of RTP-STAR was released, named SC-ION (spatiotemporal clustering and inference of omics networks), which has been successfully applied in arabidopsis to infer TF-centered, brassinosteroid-responsive networks by separately constructing abundance and phosphosite GRNs from multi-omics data [50]. Multi-omics data at a single-cell scale has also

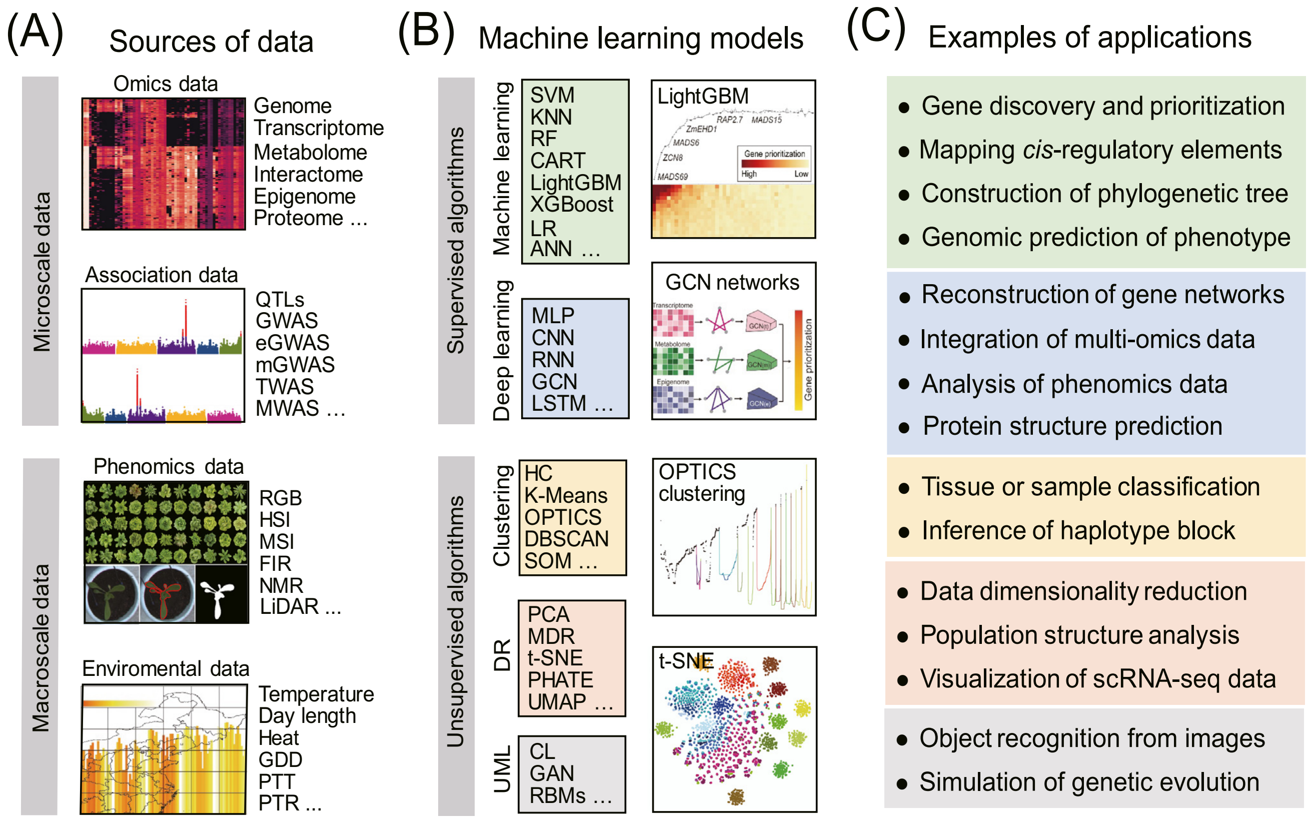
Key figure

Translation of biological data and knowledge into precision-designed breeding in plants


(A)
Basic research in plant biology
Gene functions and their contributions to traits
Gene A
Gene B
Gene C
Gene D
Gene E
Gene regulation networks
Gene evolution
A (Rice)
A (Wheat)
A (Maize)
B (Rice)
B (Wheat)
B (Maize)
E (Rice)
E (Wheat)
E (Maize)
Biological knowledgebase
Knowledge-driven molecular design breeding
• Marker-assisted selection of favorable alleles
• Genetic manipulation of desirable traits
• Polygenic pyramiding of favorable alleles
• Design of artificial biosynthetic pathways
• Design of artificial variants for genome editing
Marker-assisted selection
Gene pryamiding
Genome editing
Original phenotype
Improved phenotype
(B)
Commercial breeding in seed industry
Genotyping by targeted sequencing (GBTS)
ACGTTGGATGCCGTAGCTAGGGCATAGCTTAGACTGTGTACATAGCTTA
ACGTTGGATGCCGTAGCTAGGGCATAGCTTAGACTGTGTACATAGCTTA
ACGTTGGATGCCGTAGCTAGGGCAGAGCTTAGACTGTGTACATAGCTTA
ACGTTGGATGCCGTAGCTAGGGCAGAGCTTAGACTGTGTACATAGCTTA
ACGTTGGATGCCGTAGCTAGGGCAGAGCTTAGACTGTGTACATAGCTTA
Phenotyping data in crop field
Enviromental data
Breeding management system
Data-driven genomic design breeding
• Models of genomic selection or G2P prediction
• Models of plant phenotypic plasticity or G x E
• Models of plant and microbiome interactions
• Models of climate change impacts on yield
• Other data-driven decision making models
Training population with measured traits and genotypes
GS or G2P prediction
Candidate population for predicting traits from genotypes

Trends in Plant Science

Figure 1. (A) Understanding gene functions and regulatory mechanisms from basic research in plant biology. A biological knowledgebase will facilitate knowledge-driven molecular design breeding through multiple technologies. An example of trait improvement in a maize cultivar through marker-assisted selection, polygenic pyramiding of favorable alleles, and genome editing is shown. (B) Genotypic, phenotypic, and environmental data accumulated from commercial breeding programs will facilitate data-driven genomic design breeding by constructing a variety of decision-making models. An example of utilizing a genomic selection (GS) model to predict phenotypes from genotypes is shown. Abbreviations: G × E, genotype by environment; G2P, genotype-to-phenotype.



Trends in Plant Science

Figure 2. Applications of machine learning (ML) in plant biology. (A) Types of biological data produced from microscale to macroscale measurements. (B) Supervised and unsupervised ML methods applied in biology. The four boxes represent the applications of LightGBM to gene prioritization through feature importance analysis, the graph convolutional network (GCN) model to integration of multi-omics data, OPTICS algorithms to classification of maize lines using genotypes, and t-SNE to visualizing the structure of a maize population. (C) Examples of biological applications of ML. Abbreviations: ANN, artificial neural network; CART, classification and regression tree; CL, contrastive learning; CNN, convolutional neural network; DBSCAN, density-based spatial clustering of applications with noise; eGWAS, expression GWAS; FIR, far infrared; GAN, generative adversarial networks; GCN, graph convolutional network; GDD, growing degree days; GWAS, genome-wide association study; HC, hierarchical clustering; HIS, hyperspectral imaging; KNN, K-nearest neighbors algorithm; LiDAR, light detection and ranging light; GBM, light gradient boosting machine; LR, logistics regression; LSTM, long short-term memory; MDR, multifactor dimensionality reduction; mGWAS, metabolome GWAS; MLP, multilayer perceptron; MSI, multispectral imaging; MWAS, metabolome-wide association study; NMR, nuclear magnetic resonance; OPTICS, ordering points to identify cluster structure; PCA, principal component analysis; PHATE, potential of heat-diffusion for affinity-based trajectory embedding; PTR, photothermal ratio; PTT, photothermal time; QTLs, quantitative trait loci; RBMs, restricted Boltzmann machines; RF, random forest; RGB, red-green-blue channel camera; RNN, recurrent neural network; scRNA-seq, single-cell RNA sequencing; SOM, self-organizing map; SVM, support vector machine; t-SNE, t-distributed stochastic neighbor embedding; TWAS, transcriptome-wide association study; UMAP, uniform manifold approximation and projection; UML, unsupervised machine learning; XGBoost, extreme gradient boosting machine.

become available recently. For instance, an analytical framework for inferring cell type-specific GRNs by integrating scRNA-seq and single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) data from arabidopsis root cells has been proposed [51].

Gene discovery and prioritization

Omics analysis is often used to identify trait-related genes and causal variants, with the ultimate goal of applying marker-assisted selection and/or genome editing to improve plant traits. After reconstruction of a GRN containing a set of genes in a specific pathway of interest, the next task is to shorten the candidate list using a prioritization algorithm to assist selection of the most promising genes for functional validation. The R package mlDNA (ML-based differential

Box 1. A case study of applying DR in MODAS

Multi-omics analysis of a reference panel of germplasm can greatly enhance the mapping resolution of causative genes. However, multi-omics datasets are highly dimensional. Whole-genome resequencing of a panel containing hundreds of samples may generate tens of millions of SNPs. One single transcriptome contains tens of thousands of genes' expression per sample. If generated at single-cell scale, the sample count shall be multiplied by thousands of cell counts. If multiple conditions are included, data dimensionality will further exponentially expand. Thus, the 'curse of dimensionality' is inevitable. To solve this issue, the tool MODAS (multi-omics data association studies) utilizes multiple unsupervised learning techniques to accelerate population-scale multi-omics analysis (Figure I).

Step 1. DR on genotypic data: using a maize population as an example, MODAS first uses the Jaccard index to compute genotypic similarity of any pair of SNPs, followed by DBSCAN to cluster SNPs with high similarity on genotypes as a genomic block. PCA is then applied on each block of clustered SNPs and a pseudo-genotype index file is generated. The file contains ~60 000 genomic blocks as a highly simplified variation atlas to represent the genotypes of the original 2 million SNPs in maize.

Step 2. DR on omics data: taking metabolomic data as an example, about 30 000 compounds are profiled in one set of metabolome. However, a substantial proportion of the data is redundancy and noise, which must be removed before conducting association analysis. MODAS first uses mutual information to cluster redundant compounds exhibiting similar pattern across the samples and then performs DR within each cluster using the NMF algorithm. NMF maps the matrix of compounds × samples into one dimension of 'meta-compound' and one dimension of 'meta-sample'. The weights of meta-compound for the samples can classify the 300 maize samples into two groups in accordance with the two major haplotypes (H1 and H2) of the genomic block 30519.

Step 3. Association analysis: the weights of meta-compounds can be then regarded as phenotypic traits to perform GWAS either using the genotypes of 2 million SNPs or the pseudo-genotypes of 60 000 blocks. Both ways can map the same QTL and identify the peak SNP Chr3_182858806. Compared with GWAS between the 194 compounds and 2 million SNPs, association analysis between meta-compounds and pseudo-genotypes can reduce 45.6 h to only 0.63 min and the same result is produced. This strategy is applicable to any type of omics data for gene mining at the population-scale.

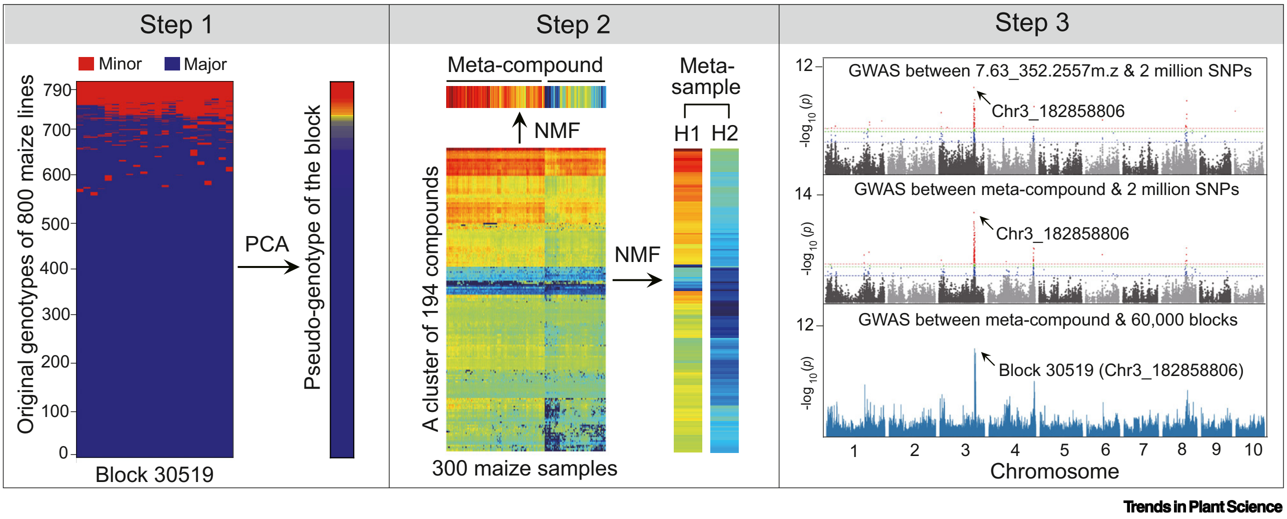


Figure I. Multi-omics data association studies. Abbreviations: GWAS, genome-wide association study; NMF, non-negative matrix factorization; PCA, principal component analysis.

network analysis) is a tool utilizing the RF algorithm to prioritize stress-responsive genes in arabidopsis and two candidate genes were successfully validated to function in the salt stress response [52]. Another novel strategy is to combine multiple GRN algorithms to prioritize genes, which was used to identify and validate the key drought-related TF *OsbHLH148* in rice [53].

Compared with GRN-based gene discovery, identifying trait-related genes and natural variants with potential for use in molecular breeding is more straightforward using GWAS. However, a QTL identified by GWAS may include dozens to hundreds of genes as a result of linkage disequilibrium, and selection of candidate genes for further validation still requires human judgement based on prior knowledge. To cope with this issue, multiple ML methods have been implemented to prioritize candidate genes and infer causal mutations within a QTL summarized from GWAS results, including penalized regression, Bayesian approaches, GBMs, and DL [54]. The key rationale of ML-based gene prioritization is compilation of a set of features obtained from genes with known function and their causal variants affecting phenotypes [55]. QTG-Finder is a tool specifically designed for plants, performing gene prioritization in post-GWAS analysis using an array of ML models [56]. The feature set of QTG-finder includes 28 features summarized from published genomic data in the model plant arabidopsis, such as DNA polymorphisms, functional annotations, cofunction networks, and evolutionary conservation. QTG-Finder2 is an upgraded version that integrates orthologous information among arabidopsis, rice, *Sorghum*, and *Setaria* into the feature set [57]. By this means, QTG-Finder2 is able to prioritize genes discovered by GWAS analysis in a non-model species where few causal genes are known. However, the major obstacle for performing ML-based gene discovery and prioritization in plants is a lack of prior knowledge, with a very limited number of genes having been functionally characterized in plants besides arabidopsis. Here, semi-supervised learning strategies relying on only a small number of labeled samples offer a possible solution. **Positive-unlabeled (PU) learning** is suitable for situations where unlabeled samples represent the majority, and PU learning has been successfully applied to causative gene prioritization [58].

Deep learning in plant phenomics

Employment of multiple types of computer vision equipment for high-throughput phenotyping (HTP) over recent decades has fostered rapid advances in plant phenomics [59]. DL is particularly suitable for large volumes of complex, unstructured imaging data because of its superior ability to perform feature extraction and model training simultaneously [60]. Applications of supervised DL algorithms, including CNN, RNN, and LSTM, to plant phenomics and precision agriculture has been well reviewed [59,60]. Here, we focus on the greatest obstacle to the development of plant phenomics. To ensure the precision and robustness of DL prediction, a sufficient number of accurately labeled samples for model training is essential. However, sample labeling is a laborious and tedious process, especially when conducted in the open field, and it is impractical in most cases for a single team to collect a sufficient number of labeled samples [8]. Of the multiple strategies proposed for solving this issue, the most promising is application of self-supervised DL algorithms, a specific class of unsupervised learning, to HTP data without absolute reliance on labeled samples. GAN utilizes either self-supervised or semi-supervised strategies to generate synthetic training data programmatically through DL models [61]. For example, DCGAN (deep convolutional GAN) was used to produce field level maize tassel synthetic images [62], StyleGAN was applied to create images for plant disease detection [63], and semi-supervised GAN was adopted to generate training samples of plant seedlings [64]. Application of GAN to segmentation of different portions of plant organs and recognition of plant diseases indicates that this strategy can significantly enhance model predictability and reduce the workload of sample labeling [65]. Because the labels of synthetic training samples are predicted by the algorithm, there is inevitably some risk of overfitting arising from incorrect prediction. To solve this problem, the label smoothing regularization (LSR) algorithm has been adopted to generalize a predictive model by replacing one-hot encoded labels with smoothed labels [66].

However, not all tasks in plant phenomics can be solved by unsupervised or semi-supervised learning and the complicated nature of plant phenomics requires these strategies to be validated using actual HTP data. An alternative strategy is to establish a joint-research community for plant phenomics where members share their labeled HTP data for public use in model training [67]. A third strategy is to adopt transfer learning [60], which has been successfully applied in plant organ segmentation and disease identification [68,69]. Multiple pretrained DL model architectures have been applied to plant phenomics using transfer learning, such as VGG-16, ResNet-50, DenseNet, GoogLeNet, and YOLOv3 [60]. Transfer learning can accelerate the training of a new model with the assistance of a pretrained model and may also benefit from large out-of-domain information [70]. However, it also must be noted that the two datasets should share high commonality to avoid overfitting.

Genomic prediction of phenotypes

The most straightforward way of achieving genomics-assisted plant breeding is to predict phenotypes through genotypes and/or omics features, known as GS or genotype-to-phenotype (G2P) prediction [71]. This is especially suitable for polygenic traits for which it is difficult to design molecular markers showing major effects or for stress-related traits associated with high phenotyping costs to achieve stress-induction conditions [72]. GS-assisted breeding has been widely employed in major crops, such as for prediction of yield heterosis in maize and rice, nutritional quality of soybean, and drought tolerance in wheat [73–75]. Most GS methods are based on best linear unbiased prediction (BLUP) or Bayesian models. However, these statistical methods are not capable of modeling the nonlinear effects and epistatic interactions of polygenic traits, especially for crops such as hybrid maize that exhibit strong effects of heterosis in the F_1 generation [76]. Multiple nonlinear ML approaches, such as RF, SVM, and ANN, for improving prediction accuracy and computing efficiency have been tested [77]. However, these classic ML methods do not produce significant improvement compared with BLUP or Bayesian models, especially when the number of training samples is far fewer than the number of samples to be predicted [71]. As a result, DL paradigms were introduced to the field of GS. The software DeepGS uses CNN to predict eight agronomic traits in wheat with a predictive accuracy surpassing those of ANN and ridge-regression BLUP (rrBLUP) [78]. CNN and Bayesian models were compared for five agronomic traits in strawberry and blueberry, which are hybrid plant species [79]. The CNN model outperformed the Bayesian model, presumably because the former is more sensitive in detecting epistatic effects in a hybrid genome. Although the CNN model exhibits outstanding performance in the research field, it is seldom used in practical breeding for three main reasons: first, the procedure for constructing and tuning a CNN model is very complicated and time consuming; second, a CNN model requires a large number of training samples to ensure model accuracy and robustness; third, actual scenarios in the seed industry are far more complicated than those in the research setting due to the genetic complexity of breeding populations. A recent study systematically evaluated 12 statistics- and ML-based algorithms for predicting 18 traits in six plant species [80]. No single method performed best across all species and traits and model tuning for automated optimization of hyper-parameters using grid search was the key to achieving the best performance for all ML methods. However, it significantly increased consumption of computing resources, with limited improvement of prediction accuracy.

Precision prediction is not the only goal of GS-assisted plant breeding pursued in industrial practice owing to the complicated nature of breeding programs and the complex composition of genetic materials. Instead, robustness, extendibility, and efficiency of a GS model, and the ability to predict a large set of unlabeled samples using a much smaller set of training samples, are the most important factors to consider [81]. To meet these demands, a one-stop toolbox called CropGBM (genomic breeding machine for crops) based on the LightGBM (light gradient boosting machine) algorithm was recently developed [82]. LightGBM is a member of the tree-based **ensemble learning** paradigm [83]. In addition to ultrafast efficiency in coping with large sample sets, LightGBM outperforms other ML methods and rrBLUP in terms of model precision and robustness, especially in situations with a small training population versus a large candidate population [82].

G2P prediction in plants also has to consider environmental influence, which may be modeled by genotype by environment (G × E) interaction. Using statistical methods like genomic BLUP or rrBLUP, G × E effect can be directly modeled by adding environmental factors as covariates in addition to genotypes [84]. By contrast, ML is more flexible in terms of integrating various types of features. Westhues *et al.* utilized gradient boosting (GB) frameworks, namely XGBoost and LightGBM, to integrate genotypes and environmental factors, including weather, location (longitude and latitude), and year, for predicting grain yield and plant height of maize. Compared with results from a linear random effect model considering the same environmental factors, the GB methods produced equivalent precision with significantly enhanced computing efficiency [85]. Another strategy to further improve predictability is to predict phenotypes by integration of genotypes and omics features from transcriptomic, metabolomic, and proteomic data [86]. The feasibility of this approach was validated by improved predictability of yield-related traits in hybrid rice [87]. However, caution is required in multiple aspects to prevent overfitting. First, DR must be applied to omics data prior to model training. Second, spatial-temporal features of the omics data must be consistent with the target traits, as tissue-specificity of gene expression will seriously affect predictability. Third, feature selection is required to select genes showing tissue-specific patterns, as expression of housekeeping genes may also introduce bias. It is worth noting that DL is perhaps a suitable solution for integrating genotypes, environmental factors, and omics data for genomic prediction of phenotypes, as these different types of features may be designed as different layers of a neural network to achieve better predictability [88].

Concluding remarks and future perspectives

Application of biotechnology and information technology will accelerate plant breeding. With the rapid advancement of various high-throughput omics technologies, plant research has entered the era of HDB. As a multidisciplinary field, HDB integrates and analyzes macroscale to microscale biological data to identify genes and regulatory networks associated with phenotypic traits. ML is the best current solution to interpret HDB data and obtain biological knowledge, given its superior capability for big data analytics. However, challenges in the field of ML analytics in both basic research and plant breeding still exist, with some of them common to both animals and plants and some specific to plants (see Outstanding questions). At last, it is foreseeable that plant breeding may finally enter the generation of the '5Gs (genome, germplasm, gene, genomic breeding, and gene editing)' [89]. To bridge basic sciences and applied breeding in plants, ML holds great promise to translate biological knowledge and omics data into precision-designed plant breeding, through knowledge-driven molecular design breeding and data-driven genomic design breeding, respectively.

A shift of breeding paradigm toward a data-driven perspective has occurred in the seed industry [90], attributed to integrated applications of doubled haploid (DH) technology, big data analytics, high-throughput genotyping, and phenotyping platforms. Taking maize as an example, a breeding enterprise commonly produces tens to hundreds of thousands of DH lines per year. Screening of this amount of DH lines cannot be done by phenotypes but instead has to rely on genotypes. Genotyping by targeted sequencing is a cost-effective technique that drastically reduces the cost by multiplexing thousands of DNA samples in one sequencing library [91]. Nevertheless, the new bottleneck has become the high cost and low efficiency for tissue collection, sample preparation, and DNA extraction, if all of these steps are done manually. In the near future, industrial-scale workflow utilizing automated liquid-handling robots are urgently required to be integrated into a modern breeding pipeline.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (31871706), Hainan Yazhou Bay Seed Laboratory (B21HJ0505), the Chinese Universities Scientific Fund (2022TC139), and the 2115 Talent Development Program of China Agricultural University.

Outstanding questions

The 'big-*p*, small-*n* (*p* >> *n*)' issue, in which *p* stands for predictors or features and *n* stands for samples, is a common challenge for all the ML methods when coping with high-dimensional omics data. What kind of automated feature extraction methods can solve this problem?

Most ML studies in plants possess 'self-validation' status, meaning that model predictability is tested only using a cross-validation scheme rather than by applying the trained model to an external dataset for validation. How can the effectiveness of ML prediction be objectively validated?

Scarcity of labeled training samples is a major obstacle limiting the application of ML in plants. Will unsupervised or self-supervised ML models like contrastive learning, transfer learning, and/or generative adversarial networks offer a solution to this issue?

Current statistical or ML methods only infer 'association' between gene and trait, rather than 'causation'. Will ML be able to infer the 'causal relationships' between mutations, genes, biomolecules, and traits, increasing the efficiency and accuracy of assisting biologists in formulating testable hypotheses to validate?

Large-scale omics datasets have been generated only for a very limited number of model plants for ML analysis. Is it possible to transfer biological knowledge obtained from model plants to non-model species by considering evolutionary conservation of gene sequences, functions, and pathways? This idea has recently been implemented by a so-called 'evolutionarily informed machine learning' framework to predict traits in maize using the data and knowledge from arabidopsis to train a XGBoost model.

Declaration of interests

The authors declare no conflict of interest.

References

1. Wallace, J.G. *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
2. Liu, J. *et al.* (2020) Application of deep learning in genomics. *Sci. China Life Sci.* 63, 1860–1878
3. Xu, C. and Jackson, S.A. (2019) Machine learning and complex biological data. *Genome Biol.* 20, 76
4. Mehta, T.S. *et al.* (2006) Epistemological issues in omics and high-dimensional biology: give the people what they want. *Physiol. Genomics* 28, 24–32
5. Yang, Y. *et al.* (2021) Applications of multi-omics technologies for crop improvement. *Front. Plant Sci.* 12, 563953
6. Ma, C. *et al.* (2014) Machine learning for big data analytics in plants. *Trends Plant Sci.* 19, 798–808
7. Bzdok, D. *et al.* (2018) Statistics versus machine learning. *Nat. Methods* 15, 233–234
8. van Dijk, A.D.J. *et al.* (2021) Machine learning in plant science and plant breeding. *iScience* 24, 101890
9. Greener, J.G. *et al.* (2022) A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* 23, 40–55
10. Esteva, A. *et al.* (2019) A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29
11. Eraslan, G. *et al.* (2019) Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* 20, 389–403
12. Khan, A. *et al.* (2022) Contrastive self-supervised learning: a survey on different architectures. In *2nd International Conference on Artificial Intelligence (ICAI)*, pp. 1–6, IEEE
13. Webb, S. (2018) Deep learning for biology. *Nature* 554, 555–557
14. Li, Y. *et al.* (2018) A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.* 19, 325–340
15. Mariammal, G. *et al.* (2022) Predicting the suitable fertilizer for crop based on soil and environmental factors using various feature selection techniques with classifiers. *Expert Syst.* Published online July 13, 2022. https://doi.org/10.1111/exsy.13024
16. Jolliffe, I.T. and Cadima, J. (2016) Principal component analysis: a review and recent developments. *Philos. Trans. A Math. Phys. Eng. Sci.* 374, 20150202
17. Meng, C. *et al.* (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* 17, 628–641
18. Kenobi, K. *et al.* (2017) Linear discriminant analysis reveals differences in root architecture in wheat seedlings related to nitrogen uptake efficiency. *J. Exp. Bot.* 68, 4969–4981
19. Gola, D. *et al.* (2016) A roadmap to multifactor dimensionality reduction methods. *Brief. Bioinform.* 17, 293–308
20. Xu, Y. *et al.* (2018) Capturing pair-wise epistatic effects associated with three agronomic traits in barley. *Genetica* 146, 161–170
21. Lee, D.D. and Seung, H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791
22. Wilson, T.J. *et al.* (2012) Identification of metagenes and their interactions through large-scale analysis of *Arabidopsis* gene expression data. *BMC Genomics* 13, 237
23. Ma, W. *et al.* (2022) easyMF: a web platform for matrix factorization-based biological discovery from large-scale transcriptome data. *Interdiscip. Sci.* 14, 746–758
24. Moon, K.R. *et al.* (2019) Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* 37, 1482–1492
25. Xiang, R. *et al.* (2021) A comparison for dimensionality reduction methods of single-cell RNA-seq data. *Front. Genet.* 12, 646936
26. Yan, J. *et al.* (2020) SR4R: an integrative SNP resource for genomic breeding and population research in rice. *Genomics Proteomics Bioinforma.* 18, 173–185
27. Yang, C. *et al.* (2021) Resequencing 250 soybean accessions: new insights into genes associated with agronomic traits and genetic networks. *Genomics Proteomics Bioinforma.* Published online July 24, 2021. https://doi.org/10.1016/j.gpb.2021.02.009
28. Liu, S. *et al.* (2022) MODAS: exploring maize germplasm with multi-omics data association studies. *Sci. Bull.* 67, 903–906
29. Bubb, K.L. and Deal, R.B. (2020) Considerations in the analysis of plant chromatin accessibility data. *Curr. Opin. Plant Biol.* 54, 69–78
30. Haque, S. *et al.* (2019) Computational prediction of gene regulatory networks in plant growth and development. *Curr. Opin. Plant Biol.* 47, 96–105
31. Redekar, N. *et al.* (2017) Inference of transcription regulatory network in low phytic acid soybean seeds. *Front. Plant Sci.* 8, 2029
32. Gunasekara, C. *et al.* (2018) TGMI: an efficient algorithm for identifying pathway regulators through evaluation of triple-gene mutual interaction. *Nucleic Acids Res.* 46, e67
33. Lachmann, A. *et al.* (2016) ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* 32, 2233–2235
34. Faith, J.J. *et al.* (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 5, e8
35. Banf, M. and Rhee, S.Y. (2017) Computational inference of gene regulatory networks: approaches, limitations and opportunities. *Biochim. Biophys. Acta Gene Regul. Mech.* 1860, 41–52
36. de Luis Balaguer, M.A. *et al.* (2017) Predicting gene regulatory networks by combining spatial and temporal gene expression data in *Arabidopsis* root stem cells. *Proc. Natl. Acad. Sci. U. S. A.* 114, E7632–E7640
37. Deng, W. *et al.* (2018) JRmGRN: joint reconstruction of multiple gene regulatory networks with common hub genes using data from multiple tissues or conditions. *Bioinformatics* 34, 3470–3478
38. Ko, D.K. and Brandizzi, F. (2020) Network-based approaches for understanding gene regulation and function in plants. *Plant J.* 104, 302–317
39. Huynh-Thu, V.A. and Sanguinetti, G. (2015) Combining tree-based and dynamical systems for the inference of gene regulatory networks. *Bioinformatics* 31, 1614–1622
40. Zhou, P. *et al.* (2020) Meta gene regulatory networks in maize highlight functionally relevant regulatory interactions. *Plant Cell* 32, 1377–1396
41. Harrington, S.A. *et al.* (2020) The wheat GENIE3 network provides biologically-relevant information in polyploid wheat. *G3 (Bethesda)* 10, 3675–3686
42. Park, S. *et al.* (2018) BTNET: boosted tree based gene regulatory network inference algorithm using time-course measurement data. *BMC Syst. Biol.* 12, 20
43. Ni, Y. *et al.* (2016) A machine learning approach to predict gene regulatory networks in seed development in *Arabidopsis*. *Front. Plant Sci.* 7, 1936
44. Moerman, T. *et al.* (2019) GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* 35, 2159–2161
45. Van de Sande, B. *et al.* (2020) A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* 15, 2247–2276
46. Walley, J.W. *et al.* (2016) Integration of omic networks in a developmental atlas of maize. *Science* 353, 814–818
47. Ding, J. *et al.* (2018) iDREM: interactive visualization of dynamic regulatory networks. *PLoS Comput. Biol.* 14, e1006019
48. Mishra, B. *et al.* (2021) Dynamic regulatory event mining by iDREM in large-scale multi-omics datasets during biotic and abiotic stress in plants. *Methods Mol. Biol.* 2328, 191–202
49. Zander, M. *et al.* (2020) Integrated multi-omics framework of the plant response to jasmonic acid. *Nat. Plants* 6, 290–302
50. Clark, N.M. *et al.* (2021) Integrated omics networks reveal the temporal signaling events of brassinosteroid response in *Arabidopsis*. *Nat. Commun.* 12, 5858
51. Dorrity, M.W. *et al.* (2021) The regulatory landscape of *Arabidopsis thaliana* roots at single-cell resolution. *Nat. Commun.* 12, 3334

52. Ma, C. *et al.* (2014) Machine learning-based differential network analysis: a study of stress-responsive transcriptomes in *Arabidopsis*. *Plant Cell* 26, 520–537
53. Gupta, C. *et al.* (2021) Using network-based machine learning to predict transcription factors involved in drought resistance. *Front. Genet.* 12, 652189
54. Sun, S. *et al.* (2021) Revisiting genome-wide association studies from statistical modelling to machine learning. *Brief. Bioinform.* 22, bbaa263
55. Broekema, R.V. *et al.* (2020) A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.* 10, 190221
56. 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 (Bethesda)* 9, 3129–3138
57. Lin, F. *et al.* (2020) QTG-Finder2: a generalized machine-learning algorithm for prioritizing QTL causal genes in plants. *G3 (Bethesda)* 10, 2411–2421
58. Kolosov, N. *et al.* (2021) Prioritization of disease genes from GWAS using ensemble-based positive-unlabeled learning. *Eur. J. Hum. Genet.* 29, 1527–1535
59. Yang, W. *et al.* (2020) Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol. Plant* 13, 187–214
60. Nabwire, S. *et al.* (2021) Review: application of artificial intelligence in phenomics. *Sensors (Basel)* 21, 4363
61. de Melo, C.M. *et al.* (2022) Next-generation deep learning based on simulators and synthetic data. *Trends Cogn. Sci.* 26, 174–187
62. Shete, S. *et al.* (2020) TasselGAN: an application of the generative adversarial model for creating field-based maize tassel data. *Plant Phenomics* 2020, 8309605
63. Arsenovic, M. *et al.* (2019) Solving current limitations of deep learning based approaches for plant disease detection. *Symmetry* 11, 939
64. Madsen, S.L. *et al.* (2019) Disentangling information in artificial images of plant seedlings using semi-supervised GAN. *Remote Sens.* 11, 2671
65. Wen, J. *et al.* (2020) Crop disease classification on inadequate low-resolution target images. *Sensors (Basel)* 20, 4601
66. Bi, L. and Hu, G. (2020) Improving image-based plant disease classification with generative adversarial network under limited training set. *Front. Plant Sci.* 11, 583438
67. Zheng, Y.Y. *et al.* (2019) CropDeep: the crop vision dataset for deep-learning-based classification and detection in precision agriculture. *Sensors (Basel)* 19, 1058
68. Yang, S. *et al.* (2021) High-throughput soybean seeds phenotyping with convolutional neural networks and transfer learning. *Plant Methods* 17, 1–17
69. Abbas, A. *et al.* (2021) Tomato plant disease detection using transfer learning with C-GAN synthetic images. *Comput. Electron. Agric.* 187, 106279
70. Kotar, K. *et al.* (2021) Contrasting contrastive self-supervised representation learning pipelines. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9949–9959, IEEE, Montreal, QC, Canada
71. Crossa, J. *et al.* (2017) Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975
72. Varshney, R.K. *et al.* (2005) Genomics-assisted breeding for crop improvement. *Trends Plant Sci.* 10, 621–630
73. Cui, Y. *et al.* (2020) Hybrid breeding of rice via genomic selection. *Plant Biotechnol. J.* 18, 57–67
74. Qin, J. *et al.* (2019) Genome wide association study and genomic selection of amino acid concentrations in soybean seeds. *Front. Plant Sci.* 10, 1445
75. Beukert, U. *et al.* (2020) Comparing the potential of marker-assisted selection and genomic prediction for improving rust resistance in hybrid wheat. *Front. Plant Sci.* 11, 594113
76. Xiao, Y. *et al.* (2021) The genetic mechanism of heterosis utilization in maize improvement. *Genome Biol.* 22, 148
77. Tong, H. and Nikoloski, Z. (2021) Machine learning approaches for crop improvement: leveraging phenotypic and genotypic big data. *J. Plant Physiol.* 257, 153354
78. Ma, W. *et al.* (2018) A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* 248, 1307–1318
79. Zingaretti, L.M. *et al.* (2020) Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Front. Plant Sci.* 11, 25
80. Azodi, C.B. *et al.* (2019) Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3 (Bethesda)* 9, 3691–3702
81. Jiang, S. *et al.* (2020) Genome optimization for improvement of maize breeding. *Theor. Appl. Genet.* 133, 1491–1502
82. Yan, J. *et al.* (2021) LightGBM: accelerated genomically designed crop breeding through ensemble learning. *Genome Biol.* 22, 271
83. Ke, G. *et al.* (2017) LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3149–3157, ACM, Long Beach, CA, USA
84. Jarquin, D. *et al.* (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127, 595–607
85. Westhues, C.C. *et al.* (2021) Prediction of maize phenotypic traits with genomic and environmental predictors using gradient boosting frameworks. *Front. Plant Sci.* 12, 699589
86. Arouisse, B. *et al.* (2021) Improving genomic prediction using high-dimensional secondary phenotypes. *Front. Genet.* 12, 667358
87. Wang, S. *et al.* (2019) Identification of optimal prediction models using multi-omic data for selecting hybrid rice. *Heredity (Edinb)* 123, 395–406
88. Shook, J. *et al.* (2021) Crop yield prediction integrating genotype and weather variables using deep learning. *PLoS One* 16, e0252402
89. Varshney, R.K. *et al.* (2020) 5Gs for crop genetic improvement. *Curr. Opin. Plant Biol.* 56, 190–196
90. Crossa, J. *et al.* (2021) The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. *Front. Plant Sci.* 12, 651480
91. Guo, Z. *et al.* (2021) Development of high-resolution multiple-SNP arrays for genetic analyses and molecular breeding through genotyping by target sequencing and liquid chip. *Plant Commun.* 2, 100230