the plant journal

S E B
SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2022) **111**, 1527–1538 doi: 10.1111/tpj.15905

FOCUSED REVIEW

Unsupervised and semi-supervised learning: the next frontier in machine learning for plant systems biology

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

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

Received 20 April 2022; revised 5 July 2022; accepted 7 July 2022; published online 12 July 2022.
*For correspondence (e-mail xwang@cau.edu.cn).

SUMMARY

Advances in high-throughput omics technologies are leading plant biology research into the era of big data. Machine learning (ML) performs an important role in plant systems biology because of its excellent performance and wide application in the analysis of big data. However, to achieve ideal performance, supervised ML algorithms require large numbers of labeled samples as training data. In some cases, it is impossible or prohibitively expensive to obtain enough labeled training data; here, the paradigms of unsupervised learning (UL) and semi-supervised learning (SSL) play an indispensable role. In this review, we first introduce the basic concepts of ML techniques, as well as some representative UL and SSL algorithms, including clustering, dimensionality reduction, self-supervised learning (self-SL), positive-unlabeled (PU) learning and transfer learning. We then review recent advances and applications of UL and SSL paradigms in both plant systems biology and plant phenotyping research. Finally, we discuss the limitations and highlight the significance and challenges of UL and SSL strategies in plant systems biology.

Keywords: deep learning, machine learning, plant systems biology, semi-supervised learning, unsupervised learning.

Box 1. Summary

- Machine learning technology is a powerful tool with excellent performance and wide applicability in plant systems biology.
- Machine learning can be divided into supervised, unsupervised and semi-supervised paradigms, depending on whether the training data needs to be labeled or not.
- The unsupervised and semi-supervised learning paradigms play an indispensable role in plant systems biology when labeled data are scarce or expensive.
- Plant systems biology is benefitting from the use of unsupervised and semi-supervised learning strategies, such as clustering, dimensionality reduction, self-supervised learning and transfer learning.

Box 2. Open questions

- Are there further potential applications of unsupervised and semi-supervised learning techniques in plant systems biology?
- Can unsupervised learning algorithms be used effectively to reduce redundancy and integrate complex multi-omics and multimodal data in plants?
- Will semi-supervised learning strategies (e.g. positive-unlabeled learning) effectively help to discover causative genes associated with important plant traits?
- Can other advanced ML strategies (e.g. contrastive learning and reinforcement learning) be applied in plant systems biology?

INTRODUCTION

In the past decade, advances in high-throughput phenotyping, next-generation sequencing and mass spectrometry technologies have greatly empowered plant systems biology (Yuan et al., 2008). Rapid accumulation of plant omics data at multiple levels (i.e. multi-omics data), such as genome, metabolome, phenome, proteome and transcriptome, helps us to more comprehensively and systematically analyze complex biological changes and regulatory processes, and discover crucial genes and regulatory elements to accelerate plant breeding and improvement (Yang et al., 2021). Multi-omics data contain a wealth of information on plant physiology and developmental regulation. However, in-depth exploration and integration of this knowledge is not easy because multi-omics data are often characterized by high dimensionality, redundancy and noise, and may have different sources, follow different statistical distributions, and contain different degrees of inaccuracy and uncertainty (Li et al., 2018).

Machine learning (ML) provides a way to interpret multi-omics data in plants. As its name implies, ML describes the process in which computers have the capability to autonomously learn and master complex patterns in large-scale data sets, and then make decisions or predictions about real-world events (Greener et al., 2022). ML has achieved success in both academia and industry in recent decades (Xu & Jackson, 2019). In plants, ML has been widely used for data dimensionality reduction (DR) and visualization, feature extraction and integration, gene regulatory network construction, genotype-to-phenotype (G2P) prediction and plant phenotyping (Ma et al., 2014; Tong & Nikoloski, 2021; Yang et al., 2020). ML algorithms are commonly classified into supervised learning (SL), unsupervised learning (UL) and semi-supervised learning (SSL), with the main difference being whether the training samples are labeled or not. SL uses a labeled training data set for model training and makes predictions for an unlabeled test data set, whereas UL uses an unlabeled data set to both train a model and find relationships within the data. Classification and regression are two branches of SL, whereas clustering and DR are two categories of UL (Petegrosso et al., 2020). In SSL, labels are known for only part of the training samples, and models make predictions or decisions on the test data by learning from both labeled and unlabeled samples in the training data, as exemplified by positive-unlabeled (PU) learning (Li et al., 2022).

Deep learning (DL) is a leading class of ML methods that has emerged in recent years with the proliferation of affordable computing power. The motivation of DL is to establish a neural network simulating human brain mechanisms for learning and interpreting data (Eraslan et al., 2019). DL combines low-level features to create more abstract, high-level attributes or features for discovering intricate structure in the data; this effectively solves many complex pattern-recognition problems associated with unstructured data, such as text, images, and sound, greatly advancing artificial intelligence (AI)-related technologies (Webb, 2018). Analogous to traditional ML algorithms, DL methods can also be supervised, unsupervised or semi-supervised. In recent years, supervised DL algorithms, such as CNNS (convolutional neural networks), RNNS (recurrent neural networks), and their variants, have proliferated in various fields (Liu et al., 2020). Meanwhile, other DL approaches, such as self-SL, transfer learning and reinforcement learning, are emerging and gaining extensive attention (Esteva et al., 2019).

Unsupervised and semi-supervised approaches are indispensable under certain circumstances, however. For example, feature extraction and integration of large-scale unlabeled multi-omics data and DR and clustering of single-cell-level omics data rely on UL algorithms (Petegrosso et al., 2020). In cases where well-labeled training data are difficult or costly to obtain, SSL or self-supervised strategies can be introduced (van Dijk et al., 2021). Moreover, large-scale omics data sets have been generated and annotated for only a very limited number of model plants, such as Arabidopsis, *Oryza sativa* (rice) and *Zea mays* (maize), and it is impractical to produce well-annotated training data for all non-model plant species. One possible solution is the adoption of transfer learning to achieve cross-species prediction by considering conserved gene functions and pathways between evolutionarily related species (Cheng et al., 2021). With UL and SSL approaches playing an increasingly important role in plant research, a systematic survey of their basic concepts and application in plant research is necessary. Supervised ML and DL approaches have been widely applied in multiple fields of plant biology, and have been systematically reviewed in multiple articles (Azodi et al., 2020; Mahood et al., 2020; Silva et al., 2019; van Dijk et al., 2021). However, to our knowledge, reviews on this topic are rarely seen in plant research. Thus, our review mainly focuses on advanced algorithms for UL and SSL paradigms and their recent applications in plant systems biology, which are designed to compensate for the inability to perform SL because of the insufficient availability of labeled data.

BASIC CONCEPTS AND ALGORITHMS OF UL AND SSL

Algorithms for data clustering

The UL and SSL paradigms cover a wide range of advanced ML and DL algorithms (Figure 1; Table S1). Clustering is a major branch of the UL paradigm and refers to grouping similar objects together and separating dissimilar objects into different categories. Various distance metrics have been introduced for similarity measurement and most of these are common for different clustering

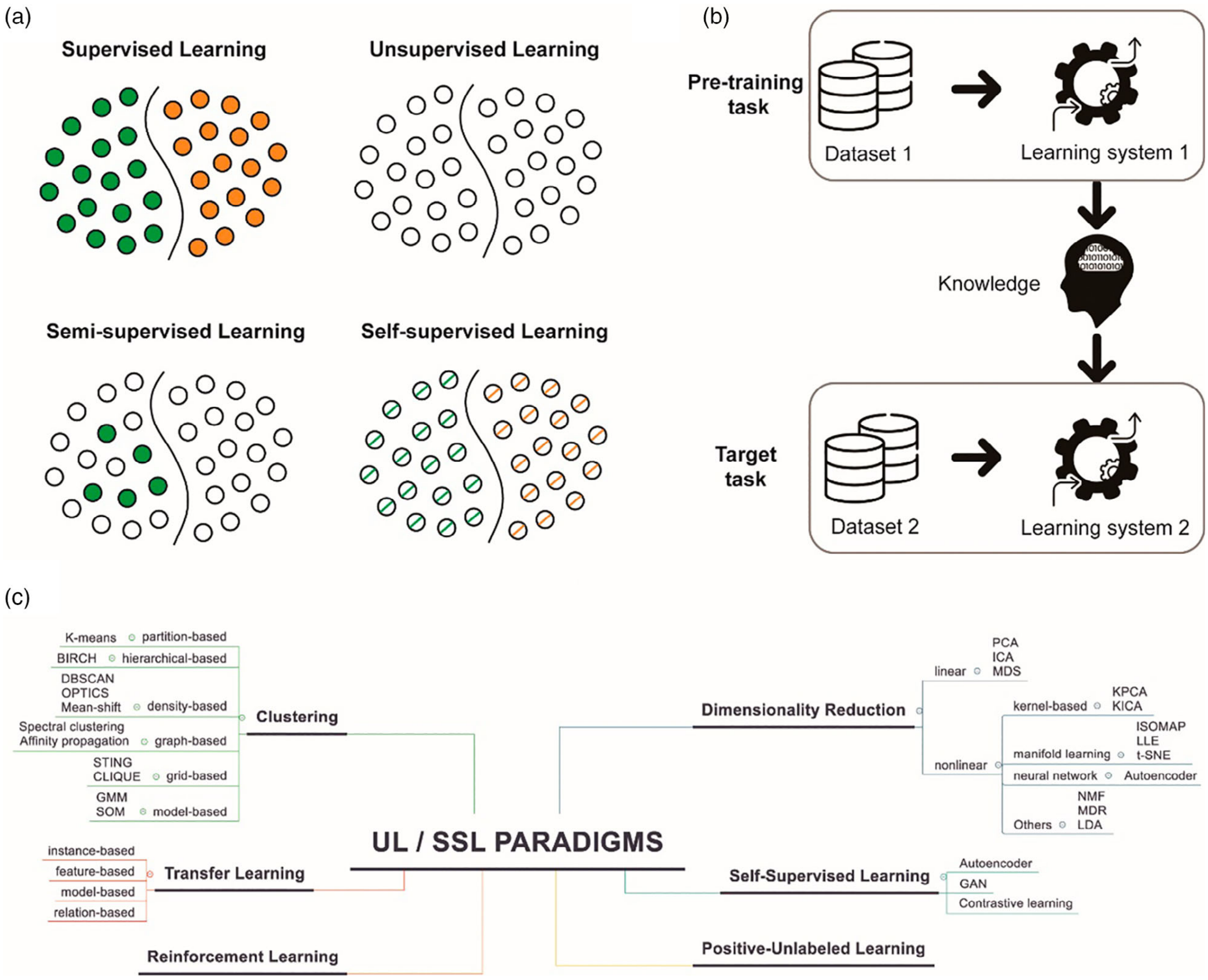

Figure 1. Classification of unsupervised learning (UL) and semi-supervised learning (SSL) paradigms. (a) Schematic of four typical machine learning strategies. White circles represent unlabeled data points, colored points represent labeled data points of different classes, striped circles represent unlabeled data points that are labeled during the model training process and black lines represent decision boundaries between classes. The SSL example diagram shows positive-unlabeled (PU) learning, with only a small number of labeled positive samples (green points) and a large number of unlabeled samples. (b) Schematic of the transfer learning strategy, which uses knowledge from a pre-training task to assist with a new task. (c) Classification of representative unsupervised and semi-supervised machine learning algorithms. Definitions for the abbreviations used in this figure can be found in Table S1.

algorithms (Berthold & Höppner, 2016; Irani et al., 2016). For example, Euclidean distance is the most common distance metric, which can be simply described as the geometric distance between objects in a multidimensional space; cosine distance is another common distance metric that calculates the cosine of the angle between two vectors in the vector space; the Pearson correlation coefficient optimizes the Euclidean distance by centralizing the value of the vector, and then calculating the cosine distance of the centralization result; the Jaccard similarity coefficient and the Jaccard distance measures the similarity and difference of two sets, respectively, and are also used as distance metrics in clustering.

Clustering algorithms can be further divided into partition-, hierarchy-, density-, graph-, grid- and model-based algorithms. One of the most popular clustering algorithms applied widely across various research areas is *K*-means, a partition-based clustering algorithm. *K*-means is easy to implement and understand, but the number of clusters must be specified and it cannot solve irregularly shaped clusters (Demidenko, 2018). Hierarchy-based clustering is a popular alternative in which the number of clusters need not be specified. This can be implemented in two ways: 'divisive hierarchical clustering', which starts by grouping all samples into one class and then gradually dividing them into smaller units; and 'agglomerative hierarchical clustering', which executes the clustering process in the opposite manner (Lorbeer et al., 2018).

Density-based clustering can identify clusters of arbitrary shape and handle noisy data better than partition- or hierarchy-based clustering. DBSCAN (density-based spatial clustering of applications with noise), OPTICS (ordering points to identify the clustering structure) and MEAN-SHIFT are representatives of this type of clustering algorithm. In DBSCAN, the maximum radius of a region and the minimum number of objects that should be accommodated in the region are first defined. Clustering then continues as long as the density (number of data points) of neighboring regions exceeds a certain threshold. Finally, objects in the same region are defined as a cluster. As DBSCAN uses fixed parameters to identify clusters, the results are highly parameter dependent (Schubert et al., 2017). OPTICS effectively reduces the parameter dependence of DBSCAN by searching for high-density areas and automatically adjusting parameter settings (Qu et al., 2018). MEAN-SHIFT is a density-based algorithm that has been applied to image segmentation (Wang & Xu, 2018). It employs the idea of moving sample points in the direction of local density increases, and points converging at the same local maximum value are assigned to the same category (Carreira-Perpinán, 2015).

Spectral clustering and affinity propagation (AP) are examples of graph-based clustering algorithms. Spectral clustering regards samples as vertices and the similarity between samples as weighted edges (Park & Zhao, 2018), whereas AP regards all objects as nodes of a network and calculates the cluster center potential of each object through message passing along each edge in the network. AP is more robust and accurate than *K*-means but has higher computational complexity and is not suitable for large sample sizes (Bodenhofer et al., 2011). Conversely, grid-based clustering algorithms (Rani, 2017) have higher efficiency and lower time complexity, making them suitable for processing a large volume of samples, but at the expense of accuracy. Model-based methods mainly refer to algorithms based on probability models and neural network models. The former group is represented by the Gaussian mixture model (GMM) (Viroli & McLachlan, 2019), whereas the latter group is represented by SOM (self-organizing map) (Quintelier et al., 2021). Although these algorithms may achieve better clustering results than other approaches, they come with the disadvantages of high computational complexity and low execution efficiency.

Algorithms for data dimensionality reduction

Dimensionality reduction (DR) is another major branch of the UL paradigm. Its essence is to use a defined method to project data points from the original high-dimensional space to another low-dimensional space. Multi-omics data are high dimensional. Taking a maize population containing 500 inbred lines as an example, the dimensionality of the gene expression data generated by transcriptome sequencing is about 20 000 expressed genes multiplied by 500 lines. The genotype data generated by whole-genome resequencing is about 5 million single-nucleotide polymorphisms (SNPs), multiplied by 500 lines. If sequencing at the single-cell scale, the dimensionality shall be further multiplied by the cell count, ranging from thousands to tens of thousands. As there is usually extensive redundancy and noise in multi-omics data, DR can extract effective information, discard useless redundancy and visualize the data (Xiang et al., 2021). DR algorithms are mainly classified into two categories: linear projection and non-linear projection. Linear methods are represented by the commonly used principal component analysis (PCA), independent component analysis (ICA) and multidimensional scaling (MDS) approaches (Anowar et al., 2021). Non-linear methods largely include kernel-based methods, manifold learning, neural network-based methods and other methods (Gisbrecht & Hammer, 2015). Here, we highlight several representative non-linear algorithms, considering their potential in dealing with complex multi-omics data.

The basic principle of kernel-based non-linear DR algorithms is to project the original data into a high-dimensional space through kernel functions and then use traditional linear DR algorithms to reduce the dimensionality of the data. These algorithms are represented by Kernel PCA (KPCA) and Kernel ICA (KICA) (Pilario et al., 2019). Manifold learning is another main branch of non-linear DR algorithms, which aims to find low-dimensional manifolds in high-dimensional spaces and identify the corresponding embedding projections to achieve DR or data visualization (Moon et al., 2018). Representative manifold learning algorithms include ISOMAP (isometric feature mapping), LLE (locally linear embedding) and T-SNE (t-distributed stochastic neighbor embedding). ISOMAP is based on the MDS theoretical framework but changes the Euclidean distance in the high-dimensional space to geodesic distance, the shortest distance between two points on the manifold (Ghojogh et al., 2020b). LLE, as its name suggests, builds a local linear model between neighboring points in a high-dimensional space and then projects data to a low-dimensional space (Ghojogh et al., 2020a). T-SNE converts Euclidean distance into conditional probability to express the similarity between data points in high-dimensional space and is mainly used for data visualization (Kobak & Berens, 2019).

In a neural network model, high-dimensional input data are transformed into low-dimensional data through a smaller hidden layer. Therefore, neural network-based algorithms also provide an effective DR approach. AUTOENCODER is a type of neural network that learns the same objective as the input. It consists of two parts: an encoder and a decoder. The encoder compresses the input into a latent space representation, and the decoder then reconstructs this representation into the output (Tschannen et al., 2018). As AUTOENCODER neural networks can encode and decode in a single model, they have been used not only for DR but also for data denoising, image generation and feature extraction (Amodio et al., 2019).

Other DR algorithms, such as NMF (non-negative matrix factorization), MDR (multifactor dimensionality reduction) and LDA (latent Dirichlet allocation), are also widely used in plant research. The principle behind NMF is that for any given non-negative matrix $A_{(m \times n)}$, two lower dimensional non-negative matrices $U_{(m \times k)}$ and $V_{(k \times n)}$ can be found with $k \leq \min(m, n)$, so $A_{(m \times n)}$ can be decomposed into the product of $U_{(m \times k)}$ and $V_{(k \times n)}$ (Lin & Boutros, 2020). NMF is mainly used for clustering and feature selection of gene expression data (Ma et al., 2022). MDR uses multiple factors in combination to project high-dimensional data to low-dimensional spaces (Gola et al., 2016). MDR is mainly used for studying gene–gene or gene–environment interactions (Xu et al., 2018). LDA is a generative probabilistic model originally used for document mining, and considers a document to be a collection of unordered words. A document contains multiple topics, and each word corresponds to one of the topics, and then the dimensionality of the document can be reduced from a large number of words to a probability distribution of several topics. LDA has been extended to other data and applied in plant phenotyping for image segmentation (Wang & Xu, 2018).

The self-SL paradigm

Self-SL is a special UL paradigm that has emerged in the DL field in recent years. It does not rely on manually labeled data but learns supervised information from large-scale unsupervised data by constructing a 'pre-text task', that is, by first training an auxiliary task with unlabeled data to learn representations, and then applying the representations to the actual task (Schmarje et al., 2021). Self-SL is exemplified by AUTOENCODER, generative learning and contrastive learning. GAN (generative adversarial network) is a representative generative learning algorithm inspired mainly by zero-sum game theory. When applied to the DL category, the model is composed of two networks, namely the generator network *G* and the discriminator network *D*. The goal of *G* is to generate real data to deceive *D*, whereas the goal of *D* is to distinguish the fake data generated by *G* from the real data. Through this continuous game, *G* learns the distribution of the data. When used for data generation, after model training is completed, *G* can generate realistic data from a random number. Unlike traditional DL models, GANS comprise two different networks instead of a single network, and the gradient update information for *G* comes from the discriminator *D*, not from the data (Creswell et al., 2018). Compared with other algorithms, GAN models generate more realistic samples, prompting their wide use in the areas of data denoising, image inpainting and training data augmentation (Saxena & Cao, 2021).

Unlike generative learning, contrastive learning aims to develop an encoder that clusters similar data while making the encoding results of different kinds of data as different as possible (Chen et al., 2020). Using this approach, a DL model can be trained to distinguish between similar and different images. Contrastive learning does not need to pay attention to the tedious details of an instance, it only needs to learn to distinguish the data in the feature space at the abstract semantic level; the model and its optimization therefore become simpler, and the generalization ability is stronger. Contrastive learning has been successfully applied in computer vision for image classification, object detection and behavior recognition (Jaiswal et al., 2020).

PU learning

Traditional supervised binary classifiers rely on both positive and negative samples, but they are unable to be directly applied when the positive samples are limited and unlabeled data accounts for the majority of the data. PU learning, a type of SSL paradigm, is designed to solve such a problem. There are two commonly used strategies in PU learning. The first strategy is termed 'selecting reliable negatives', in which putative negative samples are first identified from unlabeled data followed by training the classifier using true positive samples and putative negative samples. The second strategy is termed 'adapting the base classifier', where all unlabeled samples are initially treated as negative samples to first train a base classifier, and then 'bagging' or Bayesian approaches are applied to obtain the classification probability of each unlabeled sample (Li et al., 2022). In PU learning, the base classifiers are commonly used supervised ML algorithms, such as weighted SVM (support vector machine), for the mining of functionally related genes (Shen et al., 2020), RF (random forest), for the identification of disease-associated circular RNAs (Zeng et al., 2020), ANN (artificial neural network), for the prioritization of pathogenic variants (Pejaver et al., 2020), and the ensemble method of multi-class classification, for the prioritization of genome-wide association study (GWAS) candidate genes (Kolosov et al., 2021).

Transfer learning

Transfer learning refers to the application of knowledge or patterns learned in a certain domain to other related domains, thereby accelerating or improving the learning effect of the target domains (Zhuang et al., 2020). Transfer learning can be divided into four categories: (i) instance-based transfer, where instances or samples from the source domain that are distributed close to the target domain are selected for building a new model in the target domain; (ii) feature-based transfer, in which common features between source and target domains are identified and used for model training in the target domain; (iii) model-based transfer, which utilizes pre-trained models in the source domain along with new training data to fine-tune parameters in the target domain; and (iv) relation-based transfer, in which the relationship between concepts is learned in the source domain and then analogized to the target domain to complete the transfer of knowledge (Niu et al., 2020). Transfer learning provides an effective solution to the large volume of high-quality labeled data and considerable computing resources required for training a large ML/DL model, and has thus been successfully applied to cross-species prediction and plant phenotyping (Moore et al., 2020; Nabwire et al., 2021). In theory, transfer learning can be performed between any two domains. However, insufficient similarity between the source domain and the target domain will result in less than ideal transfer results, creating a so-called negative transfer situation (Zhuang et al., 2020). Identifying source and target domains with the highest possible similarity is the most important premise of transfer learning.

APPLICATIONS OF UL AND SSL IN PLANT SYSTEMS BIOLOGY

In plant systems biology, UL and SSL algorithms have been applied in the fields of data clustering, DR and visualization, gene regulatory network (GRN) inference, cross-species prediction and single-cell omics data analysis (Figure 2; Table S2). Visualization of high-dimensional omics data and clustering of samples or molecular modules are the most common requirements in plant genetic and genomic research and require effective DR and clustering approaches (Rai et al., 2019). PCA is one of the most popular methods for DR and visualization of genotypic and multi-omics data (de Abreu e Lima et al., 2018, Yan et al., 2020), whereas hierarchical clustering has been widely used for clustering genes with similar expression patterns in transcriptomic and proteomic research (Klepikova et al., 2016; Xu et al., 2012). Other clustering and DR algorithms have also been employed in plant systems biology. For example, the t-SNE and OPTICS algorithms have been used for analyzing the genotypic data of a large-scale maize hybrid population to better visualize its population structure (Yan et al., 2021). DBSCAN has been combined with PCA to compress the genotypic data of a maize inbred population to assist a downstream association study (Liu

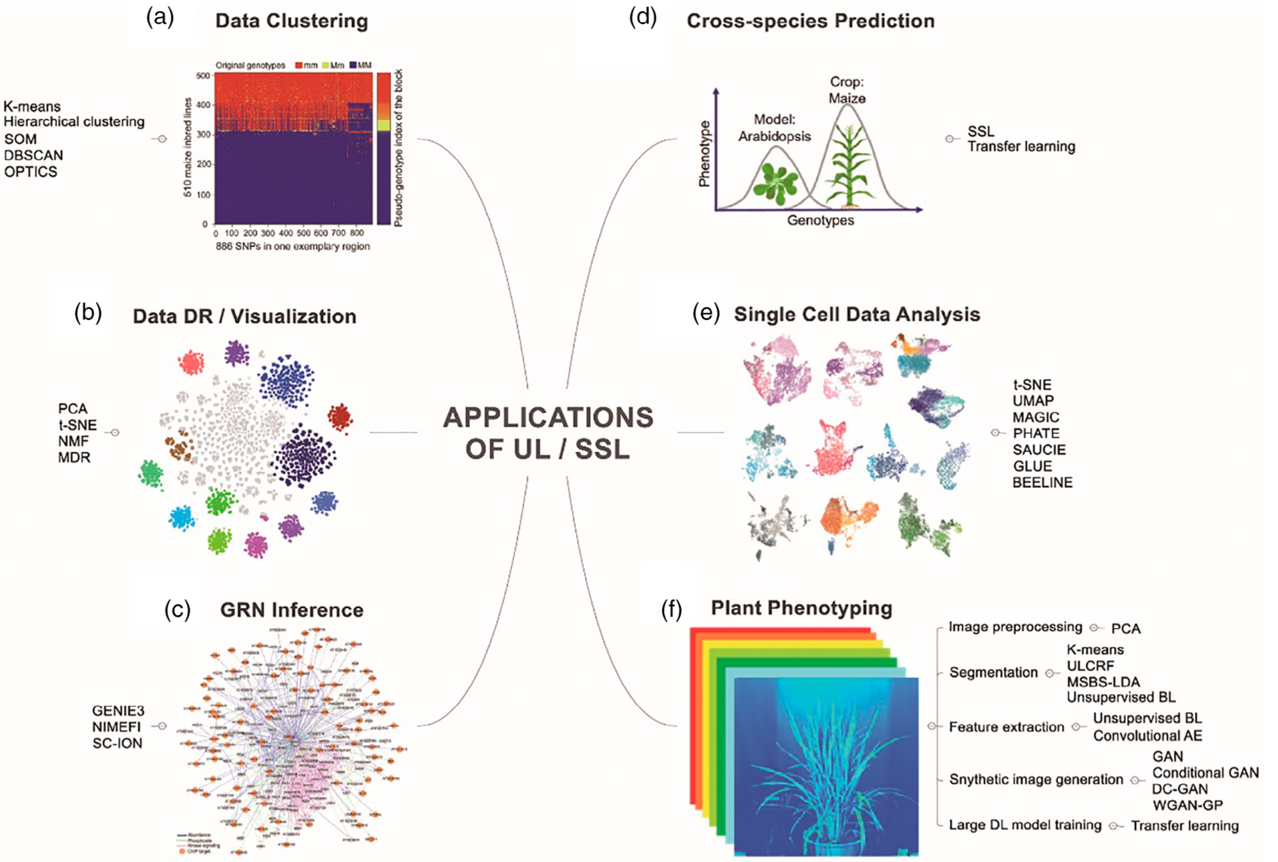

Figure 2. Application of unsupervised learning (UL) and semi-supervised learning (SSL) paradigms in plant systems biology. Representative algorithms and tools in each area are listed. Definitions for the abbreviations used in this figure can be found in Table S2. (a) A pseudo-genotype index of a maize germplasm population generated with DBSCAN (Liu et al., 2022). (b) Population structure visualization of a hybrid maize data set using t-SNE (Yan et al., 2021). (c) A merged network of transcription factor abundance, phosphosite and kinase signaling in Arabidopsis (Clark et al., 2021). (d) Using evolutionarily conserved nitrogen-responsive genes across Arabidopsis and maize to enhance the power of genotype-to-phenotype (G2P) prediction for nitrogen use efficiency traits (Cheng et al., 2021). (e) UMAP embedding of a single-nuclei chromatin accessibility data set in maize (Marand et al., 2021). (f) Hyperspectral imaging in crop phenotyping (Yang et al., 2020).

et al., 2022). The NMF algorithm has been used for decomposing expression matrices comprising thousands of genes into a small number of metagenes in Arabidopsis and maize, facilitating the exploration of downstream gene function (Ma et al., 2022; Wilson et al., 2012). Meanwhile, the MDR algorithm has been adopted for identifying multiple pairwise epistatic effects and gene–environment interactions underlying agronomic and quality traits in rice and *Hordeum vulgare* (barley) (Xu et al., 2015; Xu et al., 2018).

In addition to conventional UL approaches, tree-based ensemble learning algorithms, such as RF and gradient boosting, can also be used in an unsupervised manner for expression data-based GRN inference (Ko & Brandizzi, 2020), as they can evaluate the importance of each input feature, i.e. rank the contribution of the feature to the predicted result. The expression data of transcription factors is used to construct a feature matrix to predict the expression level of each gene. Genes for which expression levels can be predicted well are likely to be target genes of transcription factors, whereas transcription factors with high feature importance might have a regulatory relationship with target genes. Thus, transcriptome-level GRNs can be constructed through this process (Ruyssinck et al., 2014). As known gene regulatory relationships are not provided for model construction and training, this is essentially an unsupervised approach (Maetschke et al., 2014). GENIE3 (gene network inference with ensemble of trees) is an example of such a framework, based on the RF algorithm and with excellent scalability (Huynh-Thu & Geurts, 2019), and has been widely used for expression data-based GRN inference in plants such as maize and wheat (Harrington et al., 2020; Zhou et al., 2020). Recently, the application of GENIE3 has also been expanded to the construction of multi-omics-level GRNs with transcriptomic, proteomic and epigenomic data in Arabidopsis (Clark et al., 2021).

In plants, a lack of annotated genes and pathways has become a bottleneck for large-scale omics research. Semi-supervised and transfer learning strategies that consider the conservation of genes and pathways among evolutionarily related species have become an effective approach for prediction tasks. In a recent study on exploiting functionally relevant genes, a feature set was constructed to include protein–protein interactions (PPIs) of orthologous genes between species, knowledge extracted from biological literature, the conserved gene co-expression network, shared functional annotations and transcription factor binding sites of orthologous genes to synthetically train different models. The results showed that the SSL method of PU learning significantly outperformed other algorithms in the experimentally validated Arabidopsis benchmark data set (Shen et al., 2020). Another example is the use of a transfer learning strategy to predict specialized/general metabolism-related genes in *Solanum lycopersicum* (tomato) (Moore et al., 2020). In this study, multiple features, including evolutionary, structural and expression properties, were calculated for both Arabidopsis and tomato genes; the RF and SVM algorithms were then applied to different prediction tasks. Well-annotated Arabidopsis genes performed better as training data than known tomato genes. Similarly, in another so-called 'evolutionarily informed machine learning' framework, an XGBOOST model was trained using transcriptomic data and nitrogen-use efficiency (NUE) as a trait in Arabidopsis to predict NUE and related genes in maize (Cheng et al., 2021). Although this field is in a very preliminary stage, these studies have opened the door for transferring biological knowledge obtained from model plants to non-model species using ML strategies.

With the advancement of single-cell sequencing technology in recent years, data possessing higher dimensionality and complexity are challenging data analysis methods (Wu & Zhang, 2020). Various advanced algorithms and tools have been proposed to cope with these problems. For instance, T-SNE (Zhou & Jin, 2020), UMAP (uniform manifold approximation and projection) (Becht et al., 2019), MAGIC (Markov affinity-based graph imputation of cells) (Van Dijk et al., 2018) and PHATE (potential of heat-diffusion for affinity-based transition embedding) (Moon et al., 2019) have been implemented for understanding the structure of heterogeneous cell populations; SAUCIE (sparse autoencoder for unsupervised clustering, imputation and embedding) takes advantage of a multi-task autoencoder neural network model to simultaneously perform clustering, batch correction, visualization and denoising tasks for small conditional RNA (scRNA)-seq data (Amodio et al., 2019); and BEELINE provides a comprehensive framework integrating 12 state-of-the-art algorithms for GRN inference based on single-cell expression data (Pratapa et al., 2020). Although most of these methods were first proposed in human studies, they are also common in plant research; for example, T-SNE and UMAP have been widely employed for single-cell data visualization in plants (Marand et al., 2021).

APPLICATIONS OF UL AND SSL IN PLANT IMAGE ANALYSIS

Another important application of UL/SSL methods is in the field of plant phenotyping for image analysis. There are typically four steps in plant phenotyping: image preprocessing, segmentation, feature extraction and classification (Mochida et al., 2019). Image preprocessing enhances the target region through a series of approaches such as image cropping, contrast enhancement, denoising and DR to facilitate subsequent image analysis (Perez-Sanz et al., 2017). Segmentation is a crucial step that separates the target object or region from the rest of the image, i.e. background or noise artifacts (Singh & Misra, 2017).

Feature extraction transforms plant image data into 'feature vectors' through various algorithms to systematically describe the characteristics of the target object and enhance the predictability of the following ML-based classification. Classification refers to the establishment of models to predict the target phenotypes from image data. ML and DL techniques permeate each step of plant phenotyping, with supervised DL algorithms having gained increasing popularity because they can integrate feature-extraction and decision-making steps into one framework. For example, the CNN algorithm, and its variants such as RC-NN (REGION CNN) (Wang & Xu, 2018), FAST RCNN (Girshick, 2015), FASTER RCNN (Ren et al., 2015) and MASK RCNN (He et al., 2017), have been successfully applied to segmentation, feature extraction and classification tasks in plant phenotyping (Singh et al., 2018). However, unsupervised and semi-supervised strategies are also essential in plant phenotyping, especially when there is not enough labeled training data or a need for feature extraction without decision making.

Indeed, PCA has also been employed in plant phenotyping for image preprocessing and feature extraction (Singh et al., 2016). For example, an unsupervised Bayesian Gaussian process latent variable model (92.08% mean classification accuracy for *Daucus carota* and 94.31% for *Beta vulgaris*) and a convolutional autoencoder model (92.38% mean classification accuracy for *D. carota* and 93.28% for *B. vulgaris*) were used to extract biologically relevant features from plant leaf images followed by support vector machine (SVM) classification (Wober et al., 2021). *K*-means have been used to segment image kernels from the background based on PCA-processed image data (Hu & Zhang, 2021). For example, an unsupervised Bayesian learning approach has also been used for rice panicle segmentation, with an average recall, precision and F1 score of 96.49%, 72.31% and 82.10%, respectively (Hayat et al., 2020); the LDA-based segmentation algorithm ULCRF (unsupervised learning conditional random field) achieved an accuracy ranging from 82.41% to 99.98% for tomato fruit segmentation (Zhang & Xu, 2018), whereas MSBS-LDA (mean-shift bandwidths searching latent Dirichlet allocation) achieved a foreground–background dice (FBD) score ranging between 90.78% and 99.18% on three tomato plant image data sets (Wang & Xu, 2018).

Specific UL/SSL algorithms are also necessary in some special scenarios. As building a supervised DL model requires a large number of labeled images for training data, relying on laborious and costly manual labeling, a possible solution is to generate synthetic data through generative learning algorithms to increase the number of training samples, thereby reducing the cost and improving the accuracy of the model (van Dijk et al., 2021). GAN and its variants are the most widely used methods for synthetic image generation and have been successfully used for the segmentation of plant organs and the classification of plant diseases (Jiang & Li, 2020). ARIGAN (Arabidopsis rosette image generator through adversarial network) applies a conditional GAN for generating images of Arabidopsis plants given a condition of the number of leaves to generate (Valerio Giuffrida et al., 2017). TASSELGAN uses a deep convolutional GAN (DC-GAN) model to generate synthetic images of maize tassels and sky backgrounds separately, then merges these to produce field-based data (Shete et al., 2020). A study by Barth et al. applied a cycle GAN approach to generate more realistic synthetic plant images and improve plant part segmentation (Barth et al., 2020). However, there is a risk of overfitting when using synthetic data for model training. Bi and Hu used WGAN-GP (Wasserstein generative adversarial network with gradient penalty) with label smoothing regularization (LSR) for plant disease classification, effectively addressing the overfitting problem and improving model accuracy by 24% (Bi & Hu, 2020).

Another problem facing plant phenotyping is the enormous consumption of computational resources when training large-scale DL models. With the reuse and fine-tuning of a pre-trained model generated from a large data set associated with a similar task, transfer learning can improve the predictability of a smaller experimental data set, effectively reducing the consumption of computing resources (Nabwire et al., 2021). Transfer learning strategies have been applied in many DL architectures for plant phenotyping. Examples include ALEXNET for plant disease and pest identification (Fuentes et al., 2017), VGG-16 and VGG-19 for crop and weed segmentation (Abdalla et al., 2019), GOOGLENET for plant disease classification (Barbedo, 2018), YOLO 3 for leaf counting of Arabidopsis (Buzzy et al., 2020), and INCEPTION-RESNET and DENSENET for weed identification (Espejo-Garcia et al., 2021). Additionally, transfer learning was applied in the ARADEEPOPSIS (Arabidopsis deep learning-based optimal semantic image segmentation) pipeline by retraining a large ImageNet-based model for Arabidopsis rosette segmentation (Hüther et al., 2020). In another study, the authors adopted transfer learning and self-supervised strategies to extend a pre-trained ImageNet architecture with a triplet network for feature extraction and refinement, successfully generating feature representations from unlabeled time-series image data in Arabidopsis (Marin Zapata et al., 2021).

LIMITATIONS AND CAUTIONS WHEN APPLYING UL AND SSL

Compared with SL, UL and SSL approaches require no or only a small number of labeled samples to train a model and can be applied to a wider range of data types. However, with the absence of true labels on training samples, UL algorithms cannot predict the exact labels, and the predicted results often require additional manual curation.

Moreover, in the case of integrating data from different sources and batches, bias arising from batch effects may also influence the prediction (Kiselev et al., 2019). Similarly, as SSL uses a limited number of labeled data for model training, prediction accuracy is not expected to be as high as that achieved using SL with sufficient labeled samples (Zhou, 2017). It also must be noted that both SSL and UL may suffer a high risk of overfitting if the data features are not uniformly distributed. The same rule also applies to transfer learning, which largely depends on the similarity of source and target domains and is not applicable for tasks with low similarity or for species with high evolutionary distance (Aromolaran et al., 2021).

Additionally, different algorithms may have their own innate problems. For example, *K*-means can only cluster spherical clusters, and the number of clusters completely depends on the initial setting of *K* (Demidenko, 2018). DBSCAN requires arbitrarily set thresholds of distance and neighborhood number, and different parameter combinations have a great impact on the final clustering result (Schubert et al., 2017). SOM is even more complicated to apply because the weights of the model are difficult to determine (Quintelier et al., 2021). PCA may lead to the loss of valuable information, and the biological significance of each principal component is difficult to explain (Anowar et al., 2021). T-SNE is computationally expensive, and with poor stability and consistency (Kobak & Berens, 2019). Therefore, no algorithm is perfect for all the problems, and we must select the most suitable one according to the characteristics of the data itself.

CONCLUDING REMARKS AND FUTURE PERSPECTIVES

Advanced UL and SSL strategies, such as novel clustering and DR algorithms, transfer learning and GAN models, together with other supervised ML and DL approaches, have permeated multiple areas of plant systems biology. In this review, we introduced representative unsupervised and semi-supervised ML algorithms and highlighted their recent advances in plant systems biology research. The successful application of UL and SSL has effectively alleviated the problems of high data dimensionality, insufficient training data and scarcity of labeled data in both plant genomics and phenomics. However, UL or SSL algorithms are not panaceas. If accurate sample classification or prediction is required, the SL approaches are still essential. In addition, the traditional UL and SSL algorithms also face new challenges when analyzing complex plant multi-omics data. The joint utilization of multiple algorithms and the development of new algorithms may be useful. For example, a multi-omics data association study is computationally intensive, but by clustering the genotypic data with DBSCAN and then performing dimensionality reduction through PCA, the genotypic data including millions of SNPs can be compressed into tens of thousands of pseudo-genotypes, so that the genomic regions associated with target traits can be rapidly located (Liu et al., 2022). In the integration analysis of single-cell multi-omics data, data heterogeneity and batch effects pose great challenges to traditional algorithms. The use of graph-linked unified embedding (GLUE) introduced the generative learning strategy of the variational graph autoencoder (VGAE) algorithm to realize the unsupervised integration and regulatory inference of millions of single-cell multi-omics data, with high efficiency and accuracy (Cao & Gao, 2022). Still, a number of promising ML approaches, such as PU-learning (Li et al., 2022), contrastive learning (Chen et al., 2020) and reinforcement learning (Eckardt et al., 2021), which have succeeded in the areas of human genomics, computer vison and games, respectively, need further exploration in plant research. There is no doubt that the increasing application of advanced algorithms will further promote plant systems biology research.

ACKNOWLEDGEMENTS

This work was supported by the Key Research and Development Program of China (2018YFA0901000 and 2018YFA0901003), Hainan Yazhou Bay Seed Laboratory (B21HJ0505), Fundamental Research Funds for the Central Universities and the 2115 Talent Development Program of China Agricultural University (2662019QD050).

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest associated with this work.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Table S1. List of definitions for the abbreviations used in Figure 1.

Table S2. List of definitions for the abbreviations used in Figure 2.

REFERENCES

Abdalla, A., Cen, H., Wan, L., Rashid, R., Weng, H., Zhou, W. *et al.* (2019) Fine-tuning convolutional neural network with transfer learning for semantic segmentation of ground-level oilseed rape images in a field with high weed pressure. *Computers and Electronics in Agriculture*, **167**, 105091.

Amodio, M., van Dijk, D., Srinivasan, K., Chen, W.S., Mohsen, H., Moon, K.R. *et al.* (2019) Exploring single-cell data with deep multitasking neural networks. *Nature Methods*, **16**, 1139–1145.

Anowar, F., Sadaoui, S. & Selim, B. (2021) Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, **40**, 100378.

Aromolaran, O., Aromolaran, D., Isewon, I. & Oyelade, J. (2021) Machine learning approach to gene essentiality prediction: a review. *Briefings in Bioinformatics*, **22**. https://doi.org/10.1093/bib/bbab128

Azodi, C.B., Tang, J. & Shiu, S.H. (2020) Opening the black box: interpretable machine learning for geneticists. *Trends in Genetics*, **36**, 442–455.

Barbedo, J.G.A. (2018) Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Computers and Electronics in Agriculture*, **153**, 46–53.

Barth, R., Hemming, J. & Van Henten, E.J. (2020) Optimising realism of synthetic images using cycle generative adversarial networks for improved part segmentation. *Computers and Electronics in Agriculture*, **173**, 105378.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W., Ng, L.G. *et al.* (2019) Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, **37**, 38–44.

Berthold, M.R. & Höppner, F. (2016) On clustering time series using Euclidean distance and Pearson correlation. https://doi.org/10.48550/arXiv.1601.02213

Bi, L. & Hu, G. (2020) Improving image-based plant disease classification with generative adversarial network under limited training set. *Frontiers in Plant Science*, **11**, 583438.

Bodenhofer, U., Kothmeier, A. & Hochreiter, S. (2011) APcluster: an R package for affinity propagation clustering. *Bioinformatics*, **27**, 2463–2464.

Buzzy, M., Thesma, V., Davoodi, M. & Mohammadpour Velni, J. (2020) Real-time plant leaf counting using deep object detection networks. *Sensors*, **20**, 6896.

Cao, Z.J. & Gao, G. (2022) Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*. https://doi.org/10.1038/s41587-022-01284-4

Carreira-Perpinán, M.A. (2015) A review of mean-shift algorithms for clustering. https://doi.org/10.48550/arXiv.1503.00687

Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. (2020) A simple framework for contrastive learning of visual representations. https://doi.org/10.48550/arXiv.2002.05709

Cheng, C.Y., Li, Y., Varala, K., Bubert, J., Huang, J., Kim, G.J. *et al.* (2021) Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships. *Nature Communications*, **12**, 5627.

Clark, N.M., Nolan, T.M., Wang, P., Song, G., Montes, C., Valentine, C.T. *et al.* (2021) Integrated omics networks reveal the temporal signaling events of brassinosteroid response in Arabidopsis. *Nature Communications*, **12**, 5858.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B. & Bharath, A.A. (2018) Generative adversarial networks: an overview. *IEEE Signal Processing Magazine*, **35**, 53–65.

de Abreu e Lima, F., Li, K., Wen, W., Yan, J., Nikoloski, Z., Willmitzer, L. *et al.* (2018) Unraveling lipid metabolism in maize with time-resolved multi-omics data. *The Plant Journal*, **93**, 1102–1115.

Demidenko, E. (2018) The next-generation k-means algorithm. *Statistical Analysis and Data Mining*, **11**, 153–166.

Eckardt, J.N., Wendt, K., Bornhauser, M. & Middeke, J.M. (2021) Reinforcement learning for precision oncology. *Cancers*, **13**(18), 4624.

Eraslan, G., Avsec, Z., Gagneur, J. & Theis, F.J. (2019) Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, **20**, 389–403.

Espejo-Garcia, B., Mylonas, N., Athanasakos, L., Vali, E. & Fountas, S. (2021) Combining generative adversarial networks and agricultural transfer learning for weeds identification. *Biosystems Engineering*, **204**, 79–89.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K. *et al.* (2019) A guide to deep learning in healthcare. *Nature Medicine*, **25**, 24–29.

Fuentes, A., Yoon, S., Kim, S.C. & Park, D.S. (2017) A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors*, **17**, 2022.

Ghojogh, B., Ghodsi, A., Karray, F. & Crowley, M. (2020a) Locally linear embedding and its variants: tutorial and survey. https://doi.org/10.48550/arXiv.2011.10925

Ghojogh, B., Ghodsi, A., Karray, F. & Crowley, M. (2020b) Multidimensional scaling, sammon mapping, and isomap: tutorial and survey. https://doi.org/10.48550/arXiv.2009.08136

Girshick, R. (2015) Fast R-CNN. https://doi.org/10.48550/arXiv.1504.08083

Gisbrecht, A. & Hammer, B. (2015) Data visualization by nonlinear dimensionality reduction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **5**, 51–73.

Gola, D., Mahachie John, J.M., van Steen, K. & Konig, I.R. (2016) A roadmap to multifactor dimensionality reduction methods. *Briefings in Bioinformatics*, **17**, 293–308.

Greener, J.G., Kandathil, S.M., Moffat, L. & Jones, D.T. (2022) A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, **23**, 40–55.

Harrington, S.A., Backhaus, A.E., Singh, A., Hassani-Pak, K. & Uauy, C. (2020) The wheat genie3 network provides biologically-relevant information in polyploid wheat. *G3: Genes, Genomes, Genetics*, **10**, 3675–3686.

Hayat, M.A., Wu, J. & Cao, Y. (2020) Unsupervised Bayesian learning for rice panicle segmentation with UAV images. *Plant Methods*, **16**, 18.

He, K., Gkioxari, G., Dollár, P. & Girshick, R. (2017) Mask R-CNN. https://doi.org/10.48550/arXiv.1703.06870

Hu, Y. & Zhang, Z. (2021) GridFree: a python package of imageanalysis for interactive grain counting and measuring. *Plant Physiology*, **186**, 2239–2252.

Hüther, P., Schandry, N., Jandrasits, K., Bezrukov, I. & Becker, C. (2020) ARADEEPOPSIS, an automated workflow for top-view plant phenomics using semantic segmentation of leaf states. *Plant Cell*, **32**, 3674–3688.

Huynh-Thu, V.A. & Geurts, P. (2019) Unsupervised gene network inference with decision trees and random forests. In: Sanguinetti, G. & Huynh-Thu, V. (Eds.) *Gene regulatory networks*. Methods in Molecular Biology, Vol. **1883**. New York, NY: Humana Press. https://doi.org/10.1007/978-1-4939-8882-2_8

Irani, J., Pise, N. & Phatak, M. (2016) Clustering techniques and the similarity measures used in clustering: a survey. *International Journal of Computer Applications*, **134**, 9–14.

Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D. & Makedon, F. (2020) A survey on contrastive self-supervised learning. *Technologies*, **9**, 2.

Jiang, Y. & Li, C. (2020) Convolutional neural networks for image-based high-throughput plant phenotyping: a review. *Plant Phenomics*, **2020**, 1–22.

Kiselev, V.Y., Andrews, T.S. & Hemberg, M. (2019) Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, **20**, 273–282.

Klepikova, A.V., Kasianov, A.S., Gerasimov, E.S., Logacheva, M.D. & Penin, A.A. (2016) A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. *The Plant Journal*, **88**, 1058–1070.

Ko, D.K. & Brandizzi, F. (2020) Network-based approaches for understanding gene regulation and function in plants. *The Plant Journal*, **104**, 302–317.

Kobak, D. & Berens, P. (2019) The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, **10**, 1–14.

Kolosov, N., Daly, M.J. & Artomov, M. (2021) Prioritization of disease genes from gwas using ensemble-based positive-unlabeled learning. *European Journal of Human Genetics*, **29**, 1527–1535.

Li, F., Dong, S., Leier, A., Han, M., Guo, X., Xu, J. *et al.* (2022) Positive-unlabeled learning in bioinformatics and computational biology: a brief review. *Briefings in Bioinformatics*, **23**. https://doi.org/10.1093/bib/bbab461

Li, Y., Wu, F.X. & Ngom, A. (2018) A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, **19**, 325–340.

Lin, X. & Boutros, P.C. (2020) Optimization and expansion of non-negative matrix factorization. *BMC Bioinformatics*, **21**, 1–10.

Liu, J., Li, J., Wang, H. & Yan, J. (2020) Application of deep learning in genomics. *Science China Life Sciences*, **63**, 1860–1878.

Liu, S., Xu, F., Xu, Y., Wang, Q., Yan, J., Wang, J. *et al.* (2022) MODAS: Exploring maize germplasm with multi-omics data association studies. *Science Bulletin*, **67**(9), 903–906.

Lorbeer, B., Kosareva, A., Deva, B., Softic, D., Ruppel, P. & Kupper, A. (2018) Variations on the clustering algorithm birch. *Big Data Research*, **11**, 44–53.

Ma, C., Zhang, H.H. & Wang, X. (2014) Machine learning for big data analytics in plants. *Trends in Plant Science*, **19**, 798–808.

Ma, W., Chen, S., Zhai, J., Qi, Y., Xie, S., Song, M. & Ma, C. (2022) EasyMF: a web platform for matrix factorization-based biological discovery from large-scale transcriptome data. *Interdisciplinary Sciences: Computational Life Sciences*, **2022**, 1–13.

Maetschke, S.R., Madhamshettiwar, P.B., Davis, M.J. & Ragan, M.A. (2014) Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in Bioinformatics*, **15**, 195–211.

Mahood, E.H., Kruse, L.H. & Moghe, G.D. (2020) Machine learning: a powerful tool for gene function prediction in plants. *Applications in Plant Sciences*, **8**, e11376.

Marand, A.P., Chen, Z., Gallavotti, A. & Schmitz, R.J. (2021) A cis-regulatory atlas in maize at single-cell resolution. *Cell*, **184**, 3041–3055 e3021.

Marin Zapata, P.A., Roth, S., Schmutzler, D., Wolf, T., Manesso, E. & Clevert, D.-A. (2021) Self-supervised feature extraction from image time series in plant phenotyping using triplet networks. *Bioinformatics*, **37**, 861–867.

Mochida, K., Koda, S., Inoue, K., Hirayama, T., Tanaka, S., Nishii, R. *et al.* (2019) Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective. *GigaScience*, **8**, giy153.

Moon, K.R., Stanley, J.S., III, Burkhardt, D., van Dijk, D., Wolf, G. & Krishnaswamy, S. (2018) Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, **7**, 36–46.

Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S. *et al.* (2019) Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, **37**, 1482–1492.

Moore, B.M., Wang, P., Fan, P., Lee, A., Leong, B., Lou, Y.-R. *et al.* (2020) Within-and cross-species predictions of plant specialized metabolism genes using transfer learning. *silico Plants*, **2**, diaa005.

Nabwire, S., Suh, H.K., Kim, M.S., Baek, I. & Cho, B.K. (2021) Application of artificial intelligence in phenomics. *Sensors*, **21**(13), 4363.

Niu, S., Liu, Y., Wang, J. & Song, H. (2020) A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence*, **1**, 151–166.

Park, S. & Zhao, H. (2018) Spectral clustering based on learning similarity matrix. *Bioinformatics*, **34**, 2069–2076.

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K.A., Lin, G.N., Nam, H.J. *et al.* (2020) Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature Communications*, **11**, 5918.

Perez-Sanz, F., Navarro, P.J. & Egea-Cortines, M. (2017) Plant phenomics: an overview of image acquisition technologies and image data analysis algorithms. *Gigascience*, **6**, 1–18.

Petegrosso, R., Li, Z. & Kuang, R. (2020) Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Briefings in Bioinformatics*, **21**, 1209–1223.

Pilario, K.E., Shafiee, M., Cao, Y., Lao, L. & Yang, S.-H. (2019) A review of kernel methods for feature extraction in nonlinear process monitoring. *Processes*, **8**, 24.

Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A. & Murali, T. (2020) Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, **17**, 147–154.

Qu, Y., Deng, W. & Hu, F. (2018) Algorithm for ordering points to identify clustering structure based on spark. *Computer Science*, **45**, 97–102.

Quintelier, K., Couckuyt, A., Emmaneel, A., Aerts, J., Saeys, Y. & Van Gassen, S. (2021) Analyzing high-dimensional cytometry data using FlowSOM. *Nature Protocols*, **16**, 3775–3801.

Rai, A., Yamazaki, M. & Saito, K. (2019) A new era in plant functional genomics. *Current Opinion in Systems Biology*, **15**, 58–67.

Rani, P. (2017) A survey on STING and CLIQUE grid based clustering methods. *International Journal of Advanced Research in Computer Science*, **8**, 1510–1512.

Ren, S., He, K., Girshick, R. & Sun, J. (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, **28**, 1–9.

Ruyssinck, J., Huynh-Thu, V.A., Geurts, P., Dhaene, T., Demeester, P. & Saeys, Y. (2014) NIMEFI: gene regulatory network inference using multiple ensemble feature importance algorithms. *PLoS One*, **9**, e92709.

Saxena, D. & Cao, J. (2021) Generative adversarial networks (GANs) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*, **54**, 1–42.

Schmarje, L., Santarossa, M., Schröder, S.-M. & Koch, R. (2021) A survey on semi-, self-and unsupervised learning for image classification. *IEEE Access*, **9**, 82146–82168.

Schubert, E., Sander, J., Ester, M., Kriegel, H.P. & Xu, X. (2017) DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, **42**(3), 1–21.

Shen, K., Bunescu, R. & Wyatt, S.E. (2020) Mining functionally related genes with semi-supervised learning. https://doi.org/10.48550/arXiv.2011.03089

Shete, S., Srinivasan, S. & Gonsalves, T.A. (2020) TASSELGAN: an application of the generative adversarial model for creating field-based maize tassel data. *Plant Phenomics*, **2020**, 1–15.

Silva, J.C.F., Teixeira, R.M., Silva, F.F., Brommonschenkel, S.H. & Fontes, E.P.B. (2019) Machine learning approaches and their current application in plant molecular biology: a systematic review. *Plant Science*, **284**, 37–47.

Singh, A., Ganapathysubramanian, B., Singh, A.K. & Sarkar, S. (2016) Machine learning for high-throughput stress phenotyping in plants. *Trends in Plant Science*, **21**, 110–124.

Singh, A.K., Ganapathysubramanian, B., Sarkar, S. & Singh, A. (2018) Deep learning for plant stress phenotyping: trends and future perspectives. *Trends in Plant Science*, **23**, 883–898.

Singh, V. & Misra, A.K. (2017) Detection of plant leaf diseases using image segmentation and soft computing techniques. *Information processing in Agriculture*, **4**, 41–49.

Tong, H. & Nikoloski, Z. (2021) Machine learning approaches for crop improvement: leveraging phenotypic and genotypic big data. *Journal of Plant Physiology*, **257**, 153354.

Tschannen, M., Bachem, O. & Lucic, M. (2018) Recent advances in autoencoder-based representation learning. https://doi.org/10.48550/arXiv.1812.05069

Valerio Giuffrida, M., Scharr, H. & Tsaftaris, S.A. (2017) ARIGAN: synthetic Arabidopsis plants using generative adversarial network. https://doi.org/10.48550/arXiv.1709.00938

van Dijk, A.D.J., Kootstra, G., Kruijer, W. & de Ridder, D. (2021) Machine learning in plant science and plant breeding. *iScience*, **24**, 101890.

Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J. *et al.* (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell*, **174**(716-729), e727.

Viroli, C. & McLachlan, G.J. (2019) Deep Gaussian mixture models. *Statistics and Computing*, **29**, 43–51.

Wang, Y. & Xu, L. (2018) Unsupervised segmentation of greenhouse plant images based on modified latent dirichlet allocation. *PeerJ*, **6**, e5036.

Webb, S. (2018) Deep learning for biology. *Nature*, **554**, 555–557.

Wilson, T.J., Lai, L., Ban, Y. & Ge, S.X. (2012) Identification of metagenes and their interactions through large-scale analysis of Arabidopsis gene expression data. *BMC Genomics*, **13**, 237.

Wober, W., Mehnen, L., Sykacek, P. & Meimberg, H. (2021) Investigating explanatory factors of machine learning models for plant classification. *Plants*, **10**(12), 2674.

Wu, Y. & Zhang, K. (2020) Tools for the analysis of high-dimensional single-cell RNA sequencing data. *Nature Reviews Nephrology*, **16**, 408–421.

Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C. & Chen, X. (2021) A comparison for dimensionality reduction methods of single-cell RNA-seq data. *Frontiers in Genetics*, **12**, 646936.

Xu, C. & Jackson, S.A. (2019) Machine learning and complex biological data. *Genome Biology*, **20**, 76.

Xu, H., Jiang, B., Cao, Y., Zhang, Y., Zhan, X., Shen, X. *et al.* (2015) Detection of epistatic and gene-environment interactions underlying three quality traits in rice using high-throughput genome-wide data. *BioMed Research International*, **2015**, 1–7.

Xu, H., Zhang, W., Gao, Y., Zhao, Y., Guo, L. & Wang, J. (2012) Proteomic analysis of embryo development in rice (*Oryza sativa*). *Planta*, **235**, 687–701.

Xu, Y., Wu, Y. & Wu, J. (2018) Capturing pair-wise epistatic effects associated with three agronomic traits in barley. *Genetica*, **146**, 161–170.

Yan, J., Xu, Y., Cheng, Q., Jiang, S., Wang, Q., Xiao, Y. *et al.* (2021) LightGBM: accelerated genomically designed crop breeding through ensemble learning. *Genome Biology*, **22**, 271.

Yan, J., Zou, D., Li, C., Zhang, Z., Song, S. & Wang, X. (2020) SR4R: an integrative snp resource for genomic breeding and population research in rice. *Genomics, Proteomics & Bioinformatics*, **18**, 173–185.

Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J.H., Batchelor, W.D. *et al.* (2020) Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Molecular Plant*, **13**, 187–214.

Yang, Y., Saand, M.A., Huang, L., Abdelaal, W.B., Zhang, J., Wu, Y. *et al.* (2021) Applications of multi-omics technologies for crop improvement. *Frontiers in Plant Science*, **12**, 563953.

Yuan, J.S., Galbraith, D.W., Dai, S.Y., Griffin, P. & Stewart, C.N., Jr. (2008) Plant systems biology comes of age. *Trends in Plant Science*, **13**, 165–171.

Zeng, X., Zhong, Y., Lin, W. & Zou, Q. (2020) Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Briefings in Bioinformatics*, **21**, 1425–1436.

Zhang, P. & Xu, L. (2018) Unsupervised segmentation of greenhouse plant images based on statistical method. *Scientific Reports*, **8**, 4465.

Zhou, B. & Jin, W. (2020) Visualization of single cell RNA-seq data using t-SNE in R. In: Kidder, B. (Eds.) *Stem cell transcriptional networks*. Methods in Molecular Biology, Vol. **2117**. New York, NY: Humana. https://doi.org/10.1007/978-1-0716-0301-7_8

Zhou, P., Li, Z., Magnusson, E., Gomez Cano, F., Crisp, P.A., Noshay, J.M. *et al.* (2020) Meta gene regulatory networks in maize highlight functionally relevant regulatory interactions. *Plant Cell*, **32**, 1377–1396.

Zhou, Z.-H. (2017) A brief introduction to weakly supervised learning. *National Science Review*, **5**, 44–53.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H. *et al.* (2020) A comprehensive survey on transfer learning. *Proceedings of the IEEE*, **109**, 43–76.