PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

Cite this article: Singhal R, Izquierdo P, Ranaweera T, Segura Abá K, Brown BN.I, Lehti-Shiu MD, Shiu S-H. 2025 Using supervised machine-learning approaches to understand abiotic stress tolerance and design resilient crops. *Phil. Trans. R. Soc. B* **380**: 20240252.
https://doi.org/10.1098/rstb.2024.0252

Received: 26 November 2024
Accepted: 15 January 2025

One contribution of 21 to a theme issue 'Crops under stress: can we mitigate the impacts of climate change on agriculture and launch the 'Resilience Revolution'?'.

Subject Areas:
plant science

Keywords:
climate change, machine learning, resilient crops, abiotic stress

Authors for correspondence:
Rajneesh Singhal
e-mail: singha24@msu.edu
Shin-Han Shiu
e-mail: shius@msu.edu

†These authors contributed equally to the study.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.7758939.

THE ROYAL SOCIETY PUBLISHING

Using supervised machine-learning approaches to understand abiotic stress tolerance and design resilient crops

Rajneesh Singhal[1,†], Paulo Izquierdo[1,2,†], Thilanka Ranaweera[1,2], Kenia Segura Abá[2,3], Brianna N.I. Brown[1], Melissa D. Lehti-Shiu[1] and Shin-Han Shiu[1,2,3,4]

[1]Department of Plant Biology, [2]DOE Great Lakes Bioenergy Research Center, [3]Genetics and Genome Sciences Program, and [4]Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI 48824, USA

RS, 0000-0002-0412-3011

Abiotic stresses such as drought, heat, cold, salinity and flooding significantly impact plant growth, development and productivity. As the planet has warmed, these abiotic stresses have increased in frequency and intensity, affecting the global food supply and making it imperative to develop stress-resilient crops. In the past 20 years, the development of omics technologies has contributed to the growth of datasets for plants grown under a wide range of abiotic environments. Integration of these rapidly growing data using machine-learning (ML) approaches can complement existing breeding efforts by providing insights into the mechanisms underlying plant responses to stressful conditions, which can be used to guide the design of resilient crops. In this review, we introduce ML approaches and provide examples of how researchers use these approaches to predict molecular activities, gene functions and genotype responses under stressful conditions. Finally, we consider the potential and challenges of using such approaches to enable the design of crops that are better suited to a changing environment.

This article is part of the theme issue 'Crops under stress: can we mitigate the impacts of climate change on agriculture and launch the 'Resilience Revolution'?'.

1. Introduction

By 2050, the human population is projected to reach 10 billion, requiring a 35–56% increase in food production compared with 2010 [1,2]. Meeting this demand through traditional means would necessitate cultivating an additional area twice the size of India [1]. This challenge is further exacerbated by the threats to food production posed by climate change: average global temperatures have risen by 1.7°C and CO_2 levels by 50% (from 280 to 420 parts per million) since the mid-1700s [3,4], with projections of a further 3–5°C increase in temperature by 2100 and an increase in CO_2 to 550 parts per million by 2050 if no action is taken [5,6]. This warming has resulted in more frequent and severe abiotic stresses, such as drought, high temperatures and flooding. While individual stresses can have varying effects on crop yields (e.g. high temperatures generally reduce yields, while increased CO_2 may benefit C3 crops), the real challenge lies in their combined impact. Studies show that co-occurring stress conditions, even at relatively low levels, can severely hinder plant growth and survival when individual conditions alone have milder effects [7–9]. The expected increase in adverse conditions associated with climate change necessitates breeding resilient crops with improved productivity that can be grown on existing arable land.

Plant breeding, which focuses on selecting plants with desirable traits, has long been the foundation of crop improvement. Because desirable traits are typically selected under optimal growing conditions, representative of those under which the crops will be cultivated, this approach may not facilitate the identification of abiotic stress-tolerant genotypes that harbour stress-tolerance genes [10–12]. Plant geneticists have conducted experiments under multiple stress conditions to identify these genes and to distinguish susceptible and tolerant genotypes and analyse DNA-level differences that may be associated with stress resilience [13]. These DNA-level differences, or genetic variants, have been applied in marker-assisted selection to identify genotypes with beneficial alleles linked to abiotic stress, thereby enhancing tolerance in crops like rice and maize [14–16]. Despite the progress in identifying stress-tolerant genotypes and the associated genetic variants, identifying the causal genes remains challenging because abiotic stress tolerance is a quantitative trait and, as such, is controlled by multiple genes and influenced by genotype-by-environment interactions.

Developments in omics technology have enabled researchers to generate multiple types of data (e.g. genomic, transcriptomic, proteomic, metabolomic and phenomic), revealing the molecular processes underlying changes in plant growth and development under diverse conditions. Using these multiomics data, researchers have gained a multidimensional view of how genetic information (i.e. a DNA sequence) is processed to generate functional traits [15–19]. These omics datasets, although containing a wealth of information, are highly heterogeneous. A major challenge lies in integrating data from different sources and formats to generate additional insights. Machine learning (ML) addresses this challenge, as its algorithms can analyse, interpret and create hypotheses from large-scale, heterogeneous data [20].

2. Machine learning: a brief overview

ML involves the development of algorithms that learn from data (*training*) to identify patterns in the form of a *model* [21]. In an ML workflow, the first step is to frame the research question as an ML problem. For example, to identify genes important for drought tolerance, the ML problem is to build a model trained on input data (such as known drought tolerance genes and their expression patterns) to predict whether other genes are important for drought tolerance or not. This is an example of *supervised learning*, where a model is trained to learn patterns from input data (referred to as *features*) linked to a specific outcome of interest (important for drought tolerance or not, referred to as the *label*; figure 1). Other types of ML include unsupervised, semi-supervised and reinforcement learning. Unsupervised learning (e.g. hierarchical clustering, principal component analysis, rule-based data analysis) uses unlabelled data to train the models with the goal of identifying patterns within the datasets [22]. Semi-supervised learning is used when the labelled data are limited and the model is trained using a combination of labelled and unlabelled data [22]. Reinforcement learning mimics human trial-and-error learning; algorithms learn from feedback, interacting with the environment and making the best decision to achieve a task [22]. This review is focused on using supervised learning approaches, and the reader is directed to other in-depth reviews of the use of other ML approaches in plant biology [23–28].

In supervised learning, the first step is to collect features (i.e. predictors), such as k-mers derived from gene sequences, and labels (i.e. responses), such as important for drought tolerance or not, for each known instance (e.g. *Arabidopsis* genes). The dataset is then split into subsets for training and testing. The training data are used to train the model, while the testing data are reserved to evaluate how well the model generalizes to unseen data. Frequently, the training set is further divided into training and validation subsets to fine-tune the model and reduce overfitting (i.e. the model works well for training data but not new cases). In the drought tolerance example, a label is categorical (e.g. important for drought tolerance or not), and this type of ML task is referred to as *classification*. To evaluate a classification model, we assess the consistency between the predicted and true labels using metrics such as the area under the receiver operating characteristic curve (AUC–ROC), the area under the precision-recall curve (AUC–PR) and the F1 score to determine the model performance (see electronic supplementary material, table S1 for descriptions and the pros and cons of various metrics used for evaluating the models). In a *regression* task, the labels are continuous values (e.g. degree of drought tolerance) [22]. Performance of regression models is measured by how well the predicted values correlate with the true values using measures such as the Pearson correlation coefficient and the coefficient of determination.

Once a model reaches acceptable performance, it can be interpreted to obtain insights into what features are important for making the predictions and why some instances are poorly predicted [29,30]. There are two major ways to interpret the model: *global interpretation* and *local interpretation*. Global interpretation strategies, using measures such as permutation importance [31], identify features that contribute to model predictions of most instances, capturing patterns reflective of the input data that influence the model's decision-making. For example, in a model predicting drought tolerance genes, the important features may be relevant to a high degree of differential expression under drought conditions. In contrast, *local interpretation* strategies reveal the contributions of features and/or feature interactions to the predictions of labels for a specific instance or a small set of instances [29]. Tree-based models such as random forest (RF) can provide local explanations by identifying feature combinations that are important for specific predictions (e.g. how k-mers contribute to the prediction of drought tolerance importance for an individual gene) [32]. Shapley additive explanations (SHAP) is a local interpretation approach that provides a detailed breakdown of feature contributions for individual predictions [33]. Understanding which features or feature interactions are most influential in maximizing model performance is essential for identifying features (e.g. genes) that can be targeted in breeding decisions.

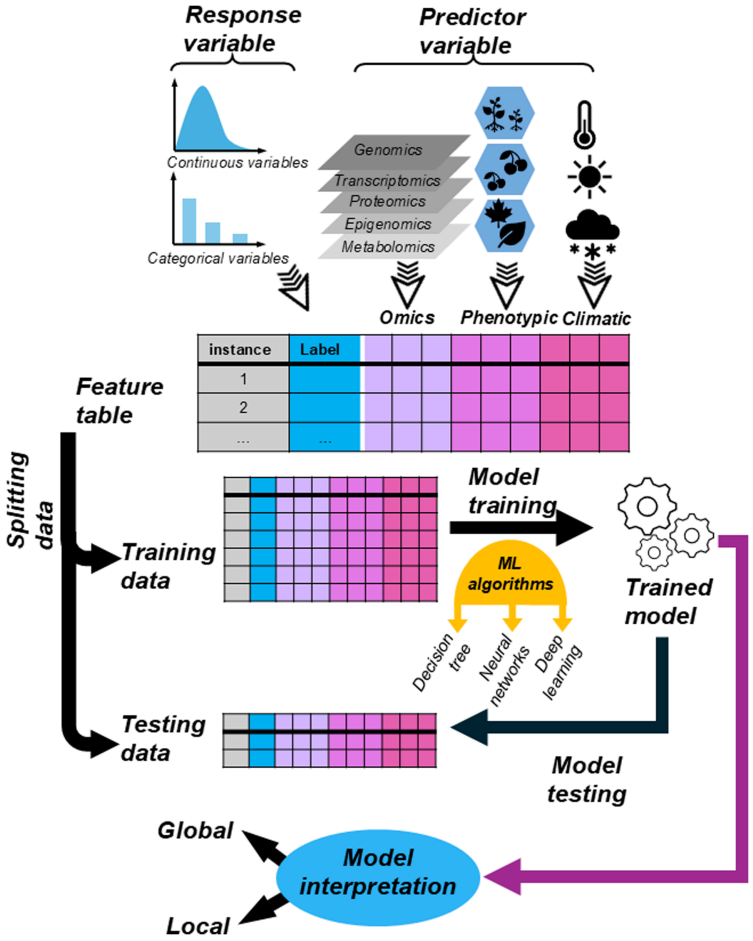

Figure 1. General ML workflow used in supervised approaches. The first step is constructing a feature table, which includes the response variable (label) to be predicted and the predictor variables (features) for each data point (instance). Both the label and feature data can be categorical or continuous values. Multiple feature types (e.g. omics, phenotypic, climatic, imaging, sound) can be used in plant biology-related predictions. The next step is splitting the full dataset into training and testing sets. The training data are used to train models, which are then tested on the testing data to evaluate their ability to predict new, unseen data. If a model performs well on the testing data, it is interpreted using different metrics (such as permutation importances in the case of global interpretation and Shapley additive explanations (SHAP) values in the case of local interpretation) to identify the key features influencing predictions. These features can provide insights into the mechanisms underlying stress responses.

3. Identifying genes associated with abiotic stress tolerance

Breeding for abiotic stress tolerance can be facilitated by introducing specific genes from resistant germplasm into elite susceptible cultivars or by modifying endogenous genes through genome editing. However, identifying these genes remains a challenge. Hundreds of genomic regions have been associated with abiotic stress resistance in plants through quantitative trait locus (QTL) mapping and genome-wide association studies (GWAS) [13], underscoring the need to prioritize candidate genes for experimental validation. ML approaches offer advantages in the identification of candidate genes because, by integrating multi-omics data, they enable the detection of complex interactions among genes, pathways and environmental factors [34]. These interactions are crucial because gene functions often rely on them [35], yet traditional methods might overlook this complexity.

One way to assess the utility of various types of existing data in narrowing down the number of candidate genes for further validation is to use causal genes (i.e. those experimentally validated to influence phenotypes) as labels. For example, one study showed that features such as functional categories, polymorphism types and paralogue number variations could correctly predict 80% of the causal genes related to abiotic stresses in *Arabidopsis* and rice [36] (for details on algorithms and parameters used for modelling see electronic supplementary material, table S2), indicating that integrating these data types using ML models can facilitate prioritizing QTL candidate genes for further studies. In another application, RF models for predicting cold-responsive genes in rice, *Arabidopsis* and cotton were established by integrating functional annotations, gene sequences and evolutionary features, achieving AUC–ROC values of 0.67, 0.70 and 0.81, respectively [37]. For context, an AUC–ROC of 0.5 indicates random performance, while 1.0 represents a perfect model. AUC–ROC values between 0.7 and 0.8 are considered acceptable and values above 0.8 are considered excellent [38]. This study also demonstrated the transferability of a cold-responsive gene prediction model trained on data from one cotton species to two other cotton species (AUC–ROC > 0.79). Although model performance can be further optimized, these examples demonstrate that ML can facilitate the discovery of genes involved in plant responses to abiotic stress conditions.

Plants adopt both specific and general response mechanisms to adapt to abiotic stresses [39]. Using ML to identify common genes responding to multiple abiotic stressors and those responding to specific ones can provide insights into the general and

specific mechanisms underlying abiotic stress responses in plants. For example, an RF model using gene expression data from 10 *Arabidopsis* accessions exposed to salt, heat, cold and high light stress predicted the stress conditions experienced by the accessions with an accuracy of 0.99 [40], and through model interpretation, three genes that may be important for general abiotic stress responses in *Arabidopsis* and rice were identified [41–43]. In another study, an RF model classified rice plants as experiencing one of 13 abiotic/biotic stress conditions with a 0.99 accuracy using gene expression data [44], demonstrating the ability of RF models to identify gene expression patterns associated with different stresses. Together, these studies illustrate the potential of ML to identify genes involved in both general and specific stress responses, which will facilitate the discovery of stress-tolerance genes for crop improvement.

In addition to identifying stress-tolerance genes, ML has also been applied to uncover genes associated with other physiological processes. For instance, XGBoost regression models based on gene expression profiles predicted nitrogen use efficiency under varying nitrogen conditions with a correlation coefficient (r) of 0.79 in maize and 0.65 in *Arabidopsis* [45]. Eight genes important for nitrogen use efficiency prediction were validated through loss-of-function mutants in both species, demonstrating ML's capacity to uncover novel genetic mechanisms. In another study aimed at identifying genes related to flowering time, an RF model was used on a global collection of 383 *Arabidopsis* accessions, combining genetic variants, gene expression and methylation data as input features [46]. This model yielded a Pearson's correlation greater than 0.6 between observed and predicted flowering times. Interestingly, the model identified both known and novel genes influencing flowering time, with 9 out of 21 novel genes confirmed to affect flowering in the Col-0 accession through loss-of-function analysis. Local interpretation revealed that the predictive importance of features (e.g. genes) varied across genetic backgrounds, highlighting the genotype-dependent contributions of genes to flowering time. The genotype dependence of gene effects may also explain why most novel genes tested did not affect flowering time: these genes were identified as important for predicting flowering time in a global collection of *Arabidopsis* accessions, but their effects were only tested in a single accession (Col-0). Local interpretation also identified 7186 interactions between features. Among these interactions was an interaction between a transcription factor (TF) and genes that the TF was reported to regulate, suggesting that these interactions may represent potential biological relationships between genes.

In summary, these studies demonstrate that ML models can identify important features for predictions that may be associated with phenotypic variability. The ML models were able to identify known causal genes and gene interactions, as well as novel candidate genes and interactions controlling phenotypic responses. This suggests that although ML models do not perfectly predict causal genes or traits, they can be further interpreted to prioritize genes that are likely associated with stress tolerance, enabling more efficient experimental validation and acceleration of resilient crop development. Future studies should consider integrating multiple datasets to increase the sample size, thereby enhancing the generalizability of the models [44]. Additionally, ML models can identify novel causal genes that have not been previously reported in the literature. The use of loss-of-function mutations in target or related species can help externally validate the genes identified as important for prediction [46].

4. Understanding gene regulation

Another way to uncover the genetic mechanisms controlling resilience traits is to elucidate the regulatory network underlying gene responses to abiotic stress. Such regulatory networks consist of *cis*-regulatory elements (CREs) located in regions proximal to genes and trans-regulatory factors like TFs [47].

Gene expression is controlled by multiple genomic factors, including promoters, enhancers and TFs [48]. Integrating various omics data using ML can reveal interactions among these regulatory elements. For instance, an RF model combining putative CREs (pCREs), TF binding sites (TFBSs) and sequence data predicted gene expression in rice under heat and drought stress, with AUC–ROC scores of 0.89 and 0.88, respectively [49]. Surprisingly, coding sequences with high GC and low AT contents were more predictive of gene expression than either pCREs or TFBSs [49]. In another study, cold-responsive gene sequences were used to predict genes responsive to cold in switchgrass, achieving a median F1 score of 0.85 across multiple time points [50]. Models using the top predictive pCREs identified novel cold-responsive elements that outperformed known TFBSs in predicting cold responses, highlighting these pCREs as potential new regulatory regions in switchgrass [50]. This approach was also applied to identify pCREs associated with drought- and heat-responsive expression in *Arabidopsis* [51].

TFBS and chromatin accessibility data have been integrated to model gene expression responses under abiotic stresses. For instance, Song *et al*. [52] used an ensemble model called Condition-Specific Regulatory network inference engine (ConSReg) to combine multiple ML models to predict regulatory genes in *Arabidopsis*, achieving an AUC–ROC of 0.84 when identifying TFs that regulate the expression of genes responsive to cold, heat and drought [52]. Using a similar approach, Gupta *et al*. [53] constructed a gene regulatory network using gene expression data to identify TFs involved in gene regulation under drought stress in rice. Using the network connectivity patterns as features, a support vector machine classifier distinguished drought-responsive TFs, achieving an AUC–PR of 0.81 [53]. Furthermore, they found that the TFs classified as drought-responsive can be divided into two groups: TFs that are specific to rice and TFs that appear to play a conserved role in plant adaptation to drought. This conclusion is supported by the fact that the latter group includes orthologous genes previously reported to function as drought regulators in *Arabidopsis*, barley, maize and sorghum. These studies suggest that combining TF-related data with other genomic and epigenetic layers may reveal the complex regulatory networks underlying abiotic stress response.

In summary, these findings demonstrate that ML models can predict gene expression and identify regulatory elements. However, gene expression regulation is a complex process influenced by factors such as tissue type, cell type, epigenetic modifications and treatments. The studies presented in this section address this complexity in part by utilizing single-cell

data [52], different stress conditions [51] and time-series data to identify changes in regulatory elements over time [50,51]. To enhance our understanding of gene regulation, future studies should focus on integrating multiple independent datasets to examine how gene regulation changes across different tissues, time points and treatments. Additionally, it will be important to investigate whether regulatory elements for specific combinations of these factors are conserved across species.

5. Designing synthetic promoters

Once the regulation of stress-responsive gene expression is understood, this knowledge can be used to design promoters that drive expression under the desired conditions. Such synthetic promoters consist of specific types and numbers of CREs placed upstream of a native or minimal promoter of a transgene. These promoters have an advantage over native promoters because the strength (i.e. expression level driven by a promoter) and specificity of the promoter can be controlled to achieve the desired level of gene expression [54–56]. Another advantage of synthetic promoters is that they can be designed using randomly selected CREs from various organisms; this minimizes the chance of homologous recombination and epigenetic silencing, which can occur when the same promoter sequences are used for the expression of multiple transgenes [56]. In one noteworthy study, minimal synthetic promoters (MinSyns) with predictable strengths were constructed from CREs identified in plant and pathogen genes [57]. The strengths (i.e. expression level of a downstream reporter gene) of a library of 1000 constitutive MinSyn promoters, which consist of 3–10 CREs placed in a randomly selected order upstream of a core promoter region (TATA box), were predicted by assigning a score to each nucleotide based on the position of the CRE relative to the core promoter. This score was then converted into predicted promoter strength based on the strengths of a subset of MinSyn promoters that were determined experimentally. The predicted strengths of these MinSyns showed a good correlation with the actual expression values (R^2 = 0.71). This suggests that synthetic promoters can be created by placing CREs, regulated by either endogenous or orthogonal TFs, in a specific arrangement in the core promoter region to control the relative expression of output genes [57]. However, constructing promoters using this approach requires knowledge of the CREs and their interacting TFs, and testing all combinations of CREs experimentally is very labour intensive. Synthetic promoters can also be created by introducing mutations in the native promoter region through error-prone PCR, followed by the selection of promoters with increased strength, often referred to as directed evolution [58]. As with the creation of MinSyns, this approach relies on extensive screening of random DNA sequences, which can be labour and resource intensive.

An alternative to testing all CREs experimentally is to use ML to extract important features from the promoter region regulating gene expression; this allows the sequences subjected to experimental screens to be selected in a more efficient and scalable manner [59–61], which can significantly reduce the cost of promoter engineering. Such ML-based approaches have enabled the de novo design of promoters in bacteria [62–64]. For example, an ML-based regression algorithm, called chaos-attention net for promoter evolution (CAPE), was developed to predict the strength of native promoters and the forces driving their evolution. CAPE, which extracts evolutionary information—i.e. sequence similarity between promoters—predicted the strengths of promoters with higher accuracy compared with previous models that lacked such information [62,64]. In experimental validation, around 37.5% of promoters designed using CAPE drove higher expression than the constitutive nisin-induced PnisA promoter in *Lactococcus lactis*, suggesting that such approaches can be used to enhance the expression strength of prokaryotic promoters [62]. A different ML algorithm, a diffusion-based generative model, was developed for the de novo design of promoters in *Escherichia coli* using native promoter sequences and gene expression data [63]. The model generated synthetic promoters that closely resembled native ones and captured essential features, such as the presence of −10 and −35 motifs, GC contents and k-mer frequency, indicating that promoters can be designed by fitting sequence information to gene expression data. ML has also been used to facilitate the design of plant synthetic promoters. Jores *et al*. [65] used reporter assays to measure the activities of promoters from *Arabidopsis* and maize and found that the promoter strength was influenced by the TATA box, promoter GC content and promoter-proximal TF binding sites; synthetic promoters made using this information had activities comparable to the 35S minimal promoter. Next, they used a convulational neural network (CNN)-based modelling approach to predict promoter strength using core promoter sequences as features, yielding R^2 values of 0.71 for tobacco leaf and 0.67 for maize protoplasts. These CNN models were then used for *in silico* evolution of 150 native promoter sequences of *Arabidopsis* and maize, and an increase in promoter strength was obtained after three rounds of evolution [65].

Although the studies published to date have been aimed at increasing promoter strength rather than increasing response to abiotic stress, they highlight the possibility of using ML approaches to uncover novel regulatory features related to abiotic stress response and to design synthetic promoters driving abiotic stress-related gene expression at the desired levels to counter the harmful effects of stress. However, to design such promoters, we still need to understand the syntax of CREs, which is their spacing, order and function within a promoter, and its correlation with gene expression in plants. ML approaches, such as transfer learning, could be used to leverage the knowledge gained from other systems for synthetic promoter design in crop plants.

6. Engineering metabolic networks

Plants produce diverse metabolites, including amino acids, sugars, organic acids, flavonoids, terpenoids, alkaloids and phenolic compounds, that are essential for survival under stress conditions [66–69]. Several metabolites, such as gamma-aminobutyric acid, which functions as an alternative energy source [67], antioxidants that protect cells from reactive oxygen species [66]

and flavonoids that protect cells from UV-light damage, accumulate in response to abiotic stress [70]. However, there is little information on the genes or pathways involved in the synthesis of such metabolites.

Identifying the genes responsible for synthesizing specific metabolites remains a significant challenge. Plants produce a vast number of specialized metabolites (SMs), and not all species produce the same metabolites. Consequently, knowledge of SM pathways from one species may not be directly transferable to others. Furthermore, the complex interplay between different metabolic pathways makes it difficult to isolate individual pathways for metabolic engineering [71,72]. Traditional approaches used to identify metabolic genes and pathways, such as gene annotation, evolutionary conservation analysis, co-expression network analysis, protein domain analysis and metabolic GWAS (reviewed in [71,73]) have been helpful but often limited in their predictive power, as they rely on a single dataset. ML offers a promising alternative: integration of multi-omics data to improve predictions and identify SM genes and pathways across various crop species [74,75].

One example of the use of ML models in SM gene prediction is the prediction of the functional annotations of unclassified plant genes. Using multi-omics data from *Arabidopsis*, tomato and maize, Bai *et al*. [74] built an ensemble model (i.e. a model that aggregates predictions from multiple models to improve accuracy and robustness) that predicted SM genes involved in the biosynthesis of terpenoids, alkaloids and phenolics within and across species using proteomics and genomics features. Their model correctly classified 92% of genes within the species from which the training data originated and achieved up to 78% accuracy when models trained on data from one species were applied to another species. In another study focused on *Arabidopsis*, multi-omics data enabled highly accurate prediction of SM genes and genes for general metabolites (i.e. metabolites important for fundamental processes; AUC–ROC = 0.87) [76], with about 50% of discrepancies between predictions and annotations potentially due to gene misannotations. ML shows potential in two areas of SM gene prediction: identifying uncharacterized SM genes across species, particularly in data-poor species, and identifying potential misannotations that prevent elucidating the molecular basis of SM.

Along with identifying SM genes, reconstructing metabolite pathways is also important. Metabolic pathway reconstruction is based on experimental evidence and computational predictions using data, such as annotated genomes and information from known pathways found in reference databases such as PlantCyc [77], KEGG [78] and BioCyc [79–81]. In one study, ML was combined with metabolite network correlation analysis to identify novel pathways in tomatoes [75]. By mapping metabolites of known tomato and non-tomato pathways as subgraphs onto metabolite correlation networks and computing network features for each subgraph, an ML model was generated to classify pathways as either a tomato or non-tomato pathway. This model showed high performance (AUC–ROC = 0.93) in predicting the presence of pathways such as ß-alanine-degradation and tryptophan-degradation through indole pyruvate pathways, which were previously not known to be present in plants, suggesting that ML can help in the identification of unknown pathways in plants [75]. In another study, Bao *et al*. [81] used the Graph Transformer and CNN model to predict plant SM pathways and classify natural products such as alkaloids and phenylpropanoids. Their model outperformed other models trained on KEGG datasets and showed an average prediction accuracy of 98% [81].

The major problem with metabolic pathway reconstruction is that the reference databases currently available for metabolic pathways are based on known gene annotations and ontology and thus can identify only known metabolic pathways. Therefore, future strategies should involve not only the identification of SM genes but also the discernment of how those genes produce a specific metabolite. This is important since the knowledge obtained from one species may not be transferable to other species [68,69], and current ML models are trained on reference databases, which may not have information from all crop species.

7. Making photosynthesis more efficient

Photosynthesis is the primary pathway that fixes CO_2 from the atmosphere and is the basis of crop production. Global warming, characterized by increased CO_2 and higher temperatures, consistently reduces photosynthesis and, thus, the yield of crop plants [82]. To address this challenge, researchers are exploring synthetic biology strategies facilitated by ML to enhance photosynthetic efficiency [83–85].

One strategy is to make the CO_2-fixation step more efficient: up to 30% of fixed CO_2 in plants is wasted through photorespiration (when Rubisco reacts with O_2 instead of CO_2), and the rate of this energetically wasteful reaction increases with increasing temperature [86]. Attempts to engineer Rubisco with lower affinity to O_2 have had limited success [87]. Thus, researchers have raised the possibility of exploiting the CO_2-fixing ability of other enzymes present in at least eight different microorganisms that may perform the carboxylation reaction [85,88,89]. Synthetic biology, coupled with ML, provides an exciting opportunity to design new CO_2-fixing pathways using enzymes with improved kinetic and thermodynamic efficiencies [85]. For instance, ML-guided protein engineering was used to improve the efficiency of a new-to-nature enzyme, glycolyl-CoA carboxylase (GCC), which catalyses the carboxylation reaction in a synthetic photorespiration bypass (the tartonyl-CoA pathway), boosting the CO_2 uptake rates by 20–60% in *in vitro* reactions [90]. In addition, using data from enzyme activity assays on randomly mutagenized versions of GCC, ML models were trained to predict promising mutations for further testing, demonstrating how ML can streamline the screening of enzyme variants for metabolic engineering applications [90].

Another example involves improving the crotonyl-coenzyme A (CoA)/ethylmalonyl-CoA/hydroxybutyryl-CoA (CETCH) cycle, which incorporates novel CO_2-fixing enzymes called enoyl-CoA carboxylases/reductases from α-proteobacteria and *Streptomyces*, to fully replace endogenous photosynthesis in *in vitro* reactions [91]. These enzymes are oxygen insensitive and more catalytically efficient than Rubisco. The CETCH cycle consists of 17 different enzymes obtained from nine different organisms and 10 cofactors, and the resulting pathway is five-fold more efficient than most of the endogenous carbon fixation

pathways. Since these enzymes are from different organisms, their activities need to be optimized. This would involve testing the synthetic CETCH cycle under 10^{25} conditions in wet lab experiments. To overcome this challenge, an ML screening strategy was developed to use iterative design–build–learn cycles to explore various reaction combinations. This significantly reduced the number of experiments required, and a 10-fold increase in CO_2-fixation efficiency compared with the original cycle was obtained after just five rounds of optimization [92]. Although the CETCH cycle has yet to be incorporated into plants, it has been assembled to work in chloroplast extracts, forming an artificial chloroplast [93].

Although these studies demonstrate how ML can significantly reduce the time and effort needed to screen thousands of enzymes and experimental conditions for increased photosynthesis efficiency, the strategies have yet to be tested in plants. The above-mentioned examples show that strategies exist to overcome the limitations of natural photosynthesis. By combining ML-based optimization and synthetic biology, it is possible to develop crop plants with enhanced photosynthetic efficiency under adverse conditions.

8. Predicting plant responses under stress

Plant responses to stress are manifested as changes in phenotypes, such as leaf colouration and emission of volatile compounds or even sounds [94–96]. High-throughput plant image data documenting changes in response to diverse stress conditions can be analysed using ML to quantify the severity of stress before any visible symptoms appear in field settings [97–99]. For example, ML algorithms have been developed to classify multiple crop species as stressed or unstressed under drought conditions [100–104]. Kaneda *et al.* developed a model integrating environmental data (e.g. transpiration rate) and plant images to predict water stress levels in plants [101]. This model utilized a deep neural network with an image feature extractor to identify wilting based on plant motion. The multi-modal model significantly outperformed models that rely solely on image data [101]. In another study, Das Choudhury and colleagues [102] used a deep neural network called HyperStress Propagate Net using hyperspectral imagery data as features to predict the onset of drought and classify cotton plants as stressed or unstressed (F1 score = 0.98). This algorithm was able to detect the early effects of drought stress, within 3 days of onset, by comparing the reflectance spectra at different wavelengths generated from hyperspectral imagery of stressed and control plants, indicating that stress can be detected before visible symptoms appear [102]. In a recent study, ultrasonic airborne sounds emitted from tomato and tobacco plants were found to reflect the physiological status of the plants [96]. ML models trained on these acoustic emissions could distinguish between drought-stressed, injured and control plants. For example, ML models could distinguish between high and low levels of dehydration stress in tomato plants under greenhouse conditions with an accuracy of 81%, suggesting that sounds emitted from plants can be utilized to monitor their drought stress levels [96].

These examples demonstrate that ML models can effectively identify patterns in high-throughput plant phenotyping data and reveal signs of stress, even before any visible symptoms appear. This application is important because it enables timely intervention to maintain plant health under adverse conditions, which is a crucial part of improving crop resilience. However, to design resilient crops the focus should be on studying multiple abiotic stresses. Recently, studies have focused on incorporating multi-omics data from multiple abiotic stresses, such as a recent study in potato [105] where high-throughput phenotyping and multi-omics data analysis were carried out under heat, drought and water-logging conditions. These data will be useful to include in future ML training models.

9. Approaches for designing resilient crops and challenges

The integration of extensive multi-omics data with interpretable ML models holds immense potential to unravel the biological mechanisms underlying resilient crop traits. Examples in the previous sections demonstrated the successful application of ML models in predicting molecular functions and complex phenotypic–environmental interactions, advancing our understanding of the genetic and mechanistic bases of crop resilience. However, a critical challenge remains: translating this knowledge into the development of climate-resilient crops.

Introducing a single gene conferring tolerance to abiotic stresses has yielded promising results in some cases (reviewed in [106,107]), such as increased drought resistance through the expression of a vacuolar H^+pyrophosphatase in maize [108] and increased thermotolerance obtained by expressing *Thermo-tolerance 1* in rice [109]. However, plants are exposed to multiple abiotic stresses simultaneously in the field, and the responses to these stresses might involve different pathways that either enhance or compromise stress resistance. In addition, we should also consider the effect of increasing stress resilience on yield. Trade-offs between resilience and yield arise because plants have finite resources that must be allocated among various physiological processes, including growth, reproduction and defence [110]. To maintain yield under stress, plants must balance resource use between processes that enhance resilience and those that support yield. Consequently, when designing stress-resistant crops, accounting for this plasticity of resource allocation is crucial to ensure stable yields under varying environmental conditions [110]. Examples in the previous sections provide evidence that ML models can identify genes and regulatory elements associated with resilience; this information can be combined with genes associated with yield in multiple species [111–114] to support the development of resilient crop cultivars with minimal yield penalties. One way to balance resilience and yield is by ensuring that stress-responsive traits are expressed only when needed, thereby preserving resources for yield under normal conditions and mitigating potential penalties associated with constant trait expression. Creating crops that express stress-responsive traits precisely when needed requires a comprehensive understanding of gene expression, metabolic pathways and regulatory modules. This integrated approach, as illustrated in figure 2, involves the identification

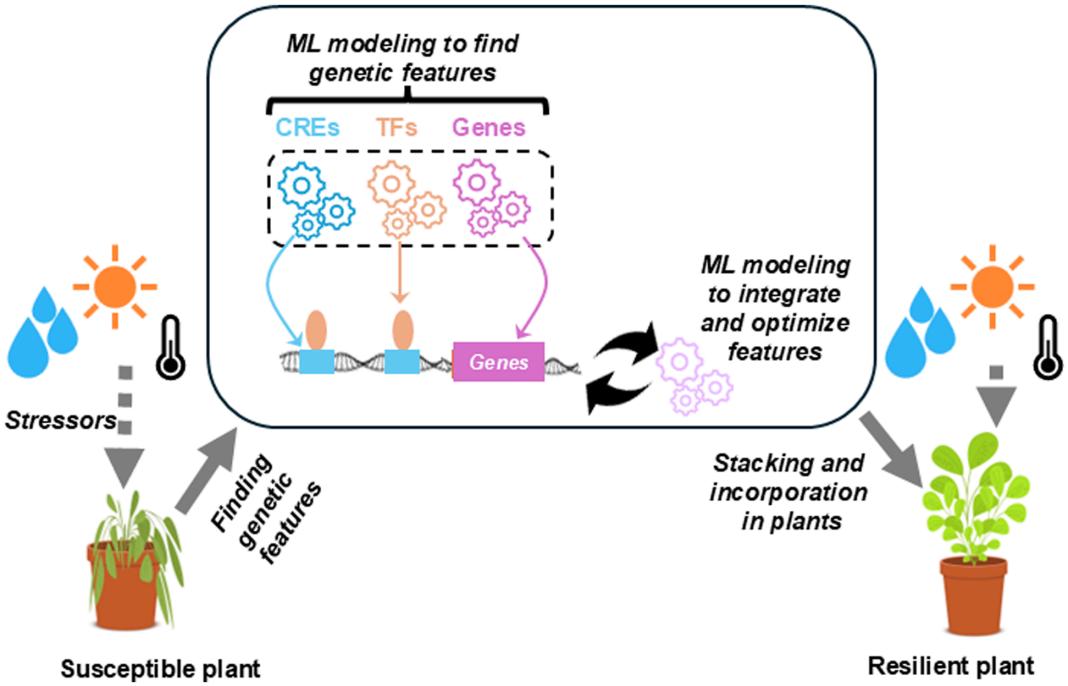

Figure 2. Designing a resilient plant. A plant susceptible to climatic conditions can be engineered to have genetic features that increase resilience. Such features include CREs, TFs and genes related to abiotic stress tolerance that are identified with the help of ML approaches. ML modelling can help determine the optimal combination of traits that can be stacked together in plants. These traits can then be incorporated into plants either through transformation or the use of genome editing tools, leading to the development of a resilient plant.

of crucial genes, such as those conferring stress resistance, and introgressing them into susceptible germplasm. Such genes can be regulated by a synthetic genetic circuit composed of CREs and corresponding TFs that activates or represses expression in response to environmental cues. Multiple genes, each controlled by a genetic circuit, can be stacked to create a robust stress response against multiple abiotic and biotic stresses. Genome editing tools, such as CRISPR/Cas9, can be used to create and transfer these modules into plants. ML models can greatly accelerate these tasks by helping to identify optimal combinations of features, such as genes and regulatory elements, that need to be incorporated in a plant for specific stress conditions.

Although ML models have recently been useful in elucidating various aspects of stress-tolerance mechanisms, several shortcomings need to be addressed for making efficient use of ML. One major limitation is the scarcity of experimentally validated causal genes for model validation. In such cases, semi-supervised learning can be used, for example, to mine functional genes in data-scarce species [115]. Some studies reviewed here have overcome this problem by using loss-of-function mutants or information from related species to validate the effects of genes on target traits. Another significant challenge is the limited sample size in most studies, which reduces an ML model's ability to discover generalizable patterns in data that can reveal the genetic mechanisms of resilience traits. An effective approach to increase sample size involves integrating multiple independent datasets, which can enhance the generalizability of the models. A second critical challenge is the lack of genomic annotations and omics data for many crops, which reduces the performance of ML models. Transfer learning, in which models pre-trained on large, well-annotated datasets, such as those from *Arabidopsis*, and fine-tuned for crops with smaller datasets, such as rice or maize, is a potential solution for elucidating abiotic stress-tolerance mechanisms in data-scarce species [116]. However, even in the model plant *Arabidopsis*, only 74% of proteins are associated with a gene ontology term, and only half of these proteins (31–38%) have terms supported by experimental evidence [117]. This knowledge gap must be addressed, and ML can assist by identifying functional annotations from the literature [117–119]. A third challenge is understanding cell-specific responses to environmental stresses, which can be masked when analysing data from multiple tissue types. In this regard, single-cell data can provide high-resolution insights into how individual cells respond to abiotic stresses. However, these datasets are often sparse and complex owing to high dimensionality (i.e. a large number of features, such as gene expression data, relative to the number of samples) and variability across cell types [120]. Such challenges can be addressed by using advanced ML models, such as deep-learning techniques, to identify underlying cell patterns and relationships [121]. Furthermore, integrating single-cell data with multi-omics data, such as epigenomics data, could enhance our understanding of how abiotic stress tolerance is regulated at the cellular level. Integration of such datasets and clustering of single-cell omics data can be performed using unsupervised learning [122]. A fourth challenge lies in understanding interactions across different omics datasets. One approach is to use graph-based ML models, which offer a framework for identifying genetic mechanisms associated with abiotic stresses by representing biological data as networks in which, for example, entities such as genes or proteins are represented as nodes and their interactions or regulatory relationships are depicted as edges [123]. By capturing relationships within and across data layers, graph-based approaches are well-suited for studying the interconnected nature of plant stress responses [124], the elucidation of which will further advance our understanding of abiotic stress tolerance.

Together, ML approaches, including transfer learning, single-cell modelling and graph-based methods, address key challenges in understanding abiotic stress tolerance. They provide powerful tools to uncover the genetic, molecular and cellular mechanisms underlying stress resilience, offering a clear pathway towards more effective crop improvement strategies.

10. Conclusion

Designing stress-resilient crops will require a comprehensive understanding of the various processes affecting plant growth, such as gene expression, metabolic pathways and regulatory modules. ML, a data-driven approach that uses algorithms to analyse large datasets to identify patterns and trends, offers a promising avenue to enhance our understanding of mechanisms controlling abiotic stress tolerance in plants. By integrating high-dimensional data, ML aids in identifying candidate genes, regulatory elements and pathways associated with resilience traits. Moreover, its ability to analyse spectral and image data allows for precisely detecting stress indicators, enhancing the efficiency and accuracy of large-scale phenotyping in experimental settings. Importantly, insights from ML models trained on species with extensive data can be transferred to related species with limited data. This underscores ML's potential to contribute significantly to the breeding and development of stress-tolerant varieties of crops for which there are less data available.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. Supplementary material is available online [125].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. R.S.: conceptualization, supervision, writing—original draft, writing—review and editing; P.I.: conceptualization, writing—original draft, writing—review and editing; T.R.: writing—original draft; K.S.A.: writing—original draft; B.N.I.B.: writing—original draft; M.D.L.-S.: conceptualization, funding acquisition, visualization, writing—review and editing; S.-H.S.: conceptualization, funding acquisition, supervision, visualization, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was supported by the U.S. Department of Energy Great Lakes Bioenergy Research Center (BER DE-SC0018409 to S.-H.S), the National Science Foundation (DGE-1828149 and IOS- 2218206 to S.-H.S. and K.S.A, and IOS-2107215 and MCB-2210431 to M.D.L.-S. and S.-H.S.).

References

1. Ranganathan J, Waite R, Searchinger T, Hanson C. 2018 *How to sustainably feed 10 billion people by 2050, in 21 charts*. See https://www.wri.org/insights/how-sustainably-feed-10-billion-people-2050-21-charts.
2. van Dijk M, Morley T, Rau ML, Saghai Y. 2021 A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. *Nat. Food* **2**, 494–501. (doi:10.1038/s43016-021-00322-9)
3. McCulloch MT, Winter A, Sherman CE, Trotter JA. 2024 300 years of sclerosponge thermometry shows global warming has exceeded 1.5°C. *Nat. Clim. Chang*. **14**, 171–177. (doi:10.1038/s41558-023-01919-7)
4. Mulhern O. *A graphical history of atmospheric CO2 levels over time*. Earth. Org. See https://earth.org/data_visualization/a-brief-history-of-co2/.
5. NASA. *Interactives – Climate Change: Vital Signs of the Planet*. Climate Time Machine. See https://climate.nasa.gov/interactives/climate-time-machine.
6. Counts TW. *CO2 concentration*. See https://www.theworldcounts.com/challenges/global-warming/CO2-concentration.
7. Zandalinas SI, Peláez-Vico MÁ, Sinha R, Pascual LS, Mittler R. 2024 The impact of multifactorial stress combination on plants, crops, and ecosystems: how should we prepare for what comes next? *Plant J*. **117**, 1800–1814. (doi:10.1111/tpj.16557)
8. Sinha R *et al*. 2024 The effects of multifactorial stress combination on rice and maize. *Plant Physiol*. **194**, 1358–1369. (doi:10.1093/plphys/kiad557)
9. Rezaei EE, Webber H, Asseng S, Boote K, Durand JL, Ewert F, Martre P, MacCarthy DS. 2023 Climate change impacts on crop yields. *Nat. Rev. Earth Environ.* **4**, 831–846. (doi:10.1038/s43017-023-00491-0)
10. Khoury CK *et al*. 2022 Crop genetic erosion: understanding and responding to loss of crop diversity. *New Phytol*. **233**, 84–118. (doi:10.1111/nph.17733)
11. Gao L, Lee JS, Hübner S, Hulke BS, Qu Y, Rieseberg LH. 2019 Genetic and phenotypic analyses indicate that resistance to flooding stress is uncoupled from performance in cultivated sunflower. *New Phytol*. **223**, 1657–1670. (doi:10.1111/nph.15894)
12. Bailey-Serres J, Parker JE, Ainsworth EA, Oldroyd GED, Schroeder JI. 2019 Genetic strategies for improving crop yields. *Nature* **575**, 109–118. (doi:10.1038/s41586-019-1679-0)
13. Raj SRG, Nadarajah K. 2022 QTL and candidate genes: techniques and advancement in abiotic stress resistance breeding of major cereals. *Int. J. Mol. Sci*. **24**, 6. (doi:10.3390/ijms24010006)
14. Das G, Rao GJN. 2015 Molecular marker assisted gene stacking for biotic and abiotic stress resistance genes in an elite rice cultivar. *Front. Plant Sci*. **6**, 698. (doi:10.3389/fpls.2015.00698)
15. Gantait S, Sarkar S, Verma SK. 2019 Marker-assisted Selection for Abiotic StressTolerance in Crop Plants. In *Molecular plant abiotic stress* (eds A Roychoudhury, D Tripathi), pp. 335–368. Hoboken, NJ: John Wiley & Sons, Ltd. (doi:10.1002/9781119463665)
16. Devi EL *et al*. 2017 Marker assisted selection (MAS) towards generating stress tolerant crop plants. *Plant Gene* **11**, 205–218. (doi:10.1016/j.plgene.2017.05.014)
17. Ma C, Wang H, Macnish AJ, Estrada-Melo AC, Lin J, Chang Y, Reid MS, Jiang CZ. 2015 Transcriptomic analysis reveals numerous diverse protein kinases and transcription factors involved in desiccation tolerance in the resurrection plant *Myrothamnus flabellifolia*. *Hortic. Res*. **2**, 15034. (doi:10.1038/hortres.2015.34)
18. Jia X-mei, Zhu Y-fang, Hu Y, Zhang R, Cheng L, Zhu Z-lei, Zhao T, Zhang X, Wang Y-xiu. 2019 Integrated physiologic, proteomic, and metabolomic analyses of *Malus halliana* adaptation to saline–alkali stress. *Hortic. Res*. **6**, 19. (doi:10.1038/s41438-019-0172-0)
19. Xu Y, Yuan Y, Du N, Wang Y, Shu S, Sun J, Guo S. 2018 Proteomic analysis of heat stress resistance of cucumber leaves when grafted onto *Momordica* rootstock. *Hortic. Res*. **5**, 18. (doi:10.1038/s41438-018-0060-z)
20. Zeng R, Li Z, Shi Y, Fu D, Yin P, Cheng J, Jiang C, Yang S. 2021 Natural variation in a type-A response regulator confers maize chilling tolerance. *Nat. Commun*. **12**, 4713. (doi:10.1038/s41467-021-25001-y)
21. Mitchell TM. 1997 Introduction to machine learning. In *Machine learning*. New York, NY: McGraw-Hill Education. See https://www.cs.cmu.edu/~tom/mlbook.html.
22. Géron A. 2022 *Hands-on machine learning with scikit-learn, keras, and tensorflow: concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media, Inc.
23. Attri I, Awasthi LK, Sharma TP, Rathee P. 2023 A review of deep learning techniques used in agriculture. *Ecol. Informatics* **77**, 102217. (doi:10.1016/j.ecoinf.2023.102217)

24. Jia J, Wang W. 2020 Review of reinforcement learning research. In *2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 186–191. (doi:10.1109/YAC51587.2020.9337653)
25. Wu X, Liu X, Zhou Y. 2022 Review of unsupervised learning techniques. In *Proceedings of 2021 Chinese Intelligent Systems Conference*. Lecture Notes in Electrical Engineering, vol. **804**, pp. 576–590, Singapore: Springer. (doi:10.1007/978-981-16-6324-6_59)
26. Hesami M, Alizadeh M, Jones AMP, Torkamaneh D. 2022 Machine learning: its challenges and opportunities in plant system biology. *Appl. Microbiol. Biotechnol*. **106**, 3507–3530. (doi:10.1007/s00253-022-11963-6)
27. Gautron R, Maillard OA, Preux P, Corbeels M, Sabbadin R. 2022 Reinforcement learning for crop management support: review, prospects and challenges. *Comput. Electron. Agric*. **200**, 107182. (doi:10.1016/j.compag.2022.107182)
28. Yan J, Wang X. 2022 Unsupervised and semi-supervised learning: the next frontier in machine learning for plant systems biology. *Plant J*. **111**, 1527–1538. (doi:10.1111/tpj.15905)
29. Azodi CB, Tang J, Shiu SH. 2020 Opening the black box: interpretable machine learning for geneticists. *Trends Genet*. **36**, 442–455. (doi:10.1016/j.tig.2020.03.005)
30. Allen GI, Gan L, Zheng L. 2024 Interpretable machine learning for discovery: statistical challenges and opportunities. *Annu. Rev. Stat. Its Appl*. **11** 97–121 . (doi:10.1146/annurev-statistics-040120-030919)
31. Altmann A, Toloşi L, Sander O, Lengauer T. 2010 Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**, 1340–1347. (doi:10.1093/bioinformatics/btq134)
32. Montesinos López OA, Montesinos López A, Crossa J. 2022 Random forest for genomic prediction. In *Multivariate statistical machine learning methods for genomic prediction* (eds OA Montesinos López, A Montesinos López), pp. 633–681. Cham, Switzerland: Springer International Publishing. (doi:10.1007/978-3-030-89010-0_15)
33. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In *Advances in neural information processing systems 30* (eds I Guyon, U Von Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett), pp. 4765–4774. Newry, UK: Curran Associates,Inc.
34. Cheng Q, Wang X. 2024 Machine Learning for AI Breeding in Plants. *Genomics, Proteomics Bioinformatics* **22**, qzae051. (doi:10.1093/gpbjnl/qzae051)
35. Baier F, Gauye F, Perez-Carrasco R, Payne JL, Schaerli Y. 2023 Environment-dependent epistasis increases phenotypic diversity in gene regulatory networks. *Sci. Adv.* **9**, eadf1773. (doi:10.1126/sciadv.adf1773)
36. Lin F, Fan J, Rhee SY. 2019 QTG-finder: a machine-learning based algorithm to prioritize causal genes of quantitative trait loci in *Arabidopsis* and rice. *G3* **9**, 3129–3138. (doi:10.1534/g3.119.400319)
37. Zhang M, Deng Y, Shi W, Wang L, Zhou N, Wang H, Zhang Z, Guan X, Zhao T. 2025 Predicting cold-stress responsive genes in cotton with machine learning models. *Crop Des*. **4**, 100085. (doi:10.1016/j.cropd.2024.100085)
38. Mandrekar JN. 2010 Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol*. **5**, 1315–1316. (doi:10.1097/jto.0b013e3181ec173d)
39. Zhang Y. 2023 Plants' response to abiotic stress: mechanisms and strategies. *Int. J. Mol. Sci* **24**, 10915. (doi:10.3390/ijms241310915)
40. Nazari L, Ghotbi V, Nadimi M, Paliwal J. 2023 A novel machine-learning approach to predict stress-responsive genes in *Arabidopsis*. *Algorithms* **16**, 407. (doi:10.3390/a16090407)
41. Tiwari M, Sharma D, Singh M, Tripathi RD, Trivedi PK. 2014 Expression of OsMATE1 and OsMATE2 alters development, stress responses and pathogen susceptibility in *Arabidopsis*. *Sci. Rep*. **4**, 3964. (doi:10.1038/srep03964)
42. Ye H, Du H, Tang N, Li X, Xiong L. 2009 Identification and expression profiling analysis of TIFY family genes involved in stress and phytohormone responses in rice. *Plant Mol. Biol*. **71**, 291–305. (doi:10.1007/s11103-009-9524-8)
43. Kahraman N, Pehlivan N. 2022 Harboured cation/proton antiporters modulate stress response to integrated heat and salt via up-regulating. *Funct. Plant Biol*. **49**, 1070–1084. (doi:10.1071/fp21334)
44. Shaik R, Ramakrishna W. 2014 Machine learning approaches distinguish multiple stress conditions using stress-responsive genes and identify candidate genes for broad resistance in rice. *Plant Physiol.* **164**, 481–495. (doi:10.1104/pp.113.225862)
45. Cheng CY *et al*. 2021 Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships. *Nat. Commun*. **12**, 5627. (doi:10.1038/s41467-021-25893-w)
46. Wang P, Lehti-Shiu MD, Lotreck S, Segura Abá K, Krysan PJ, Shiu SH. 2024 Prediction of plant complex traits via integration of multi-omics data. *Nat. Commun*. **15**, 6856. (doi:10.1038/s41467-024-50701-6)
47. de Jongh RPH, van Dijk ADJ, Julsing MK, Schaap PJ, de Ridder D. 2020 Designing eukaryotic gene expression regulation using machine learning. *Trends Biotechnol*. **38**, 191–201. (doi:10.1016/j.tibtech.2019.07.007)
48. Wang M, Li Q, Liu L. 2023 Factors and methods for the detection of gene expression regulation. *Biomolecules* **13**, 304. (doi:10.3390/biom13020304)
49. Smet D, Opdebeeck H, Vandepoele K. 2023 Predicting transcriptional responses to heat and drought stress from genomic features using a machine learning approach in rice. *Front. Plant Sci*. **14**, 1212073. (doi:10.3389/fpls.2023.1212073)
50. Ranaweera T, Brown BNI, Wang P, Shiu SH. 2022 Temporal regulation of cold transcriptional response in switchgrass. *Front. Plant Sci*. **13**, 998400. (doi:10.3389/fpls.2022.998400)
51. Azodi CB, Lloyd JP, Shiu SH. 2020 The *cis*-regulatory codes of response to combined heat and drought stress in *Arabidopsis thaliana*. *NAR Genom. Bioinforma* **2**, lqaa049. (doi:10.1093/nargab/lqaa049)
52. Song Q, Lee J, Akter S, Rogers M, Grene R, Li S. 2020 Prediction of condition-specific regulatory genes using machine learning. *Nucleic Acids Res* **48**, e62. (doi:10.1093/nar/gkaa264)
53. Gupta C, Ramegowda V, Basu S, Pereira A. 2021 Using network-based machine learning to predict transcription factors involved in drought resistance. *Front. Genet*. **12**, 652189. (doi:10.3389/fgene.2021.652189)
54. Rushton PJ, Reinstädler A, Lipka V, Lippok B, Somssich IE. 2002 Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen- and wound-induced signaling. *Plant Cell* **14**, 749–762. (doi:10.1105/tpc.010412)
55. Liu W, Stewart CN. 2016 Plant synthetic promoters and transcription factors. *Curr. Opin. Biotechnol.* **37**, 36–44. (doi:10.1016/j.copbio.2015.10.001)
56. Yasmeen E, Wang J, Riaz M, Zhang L, Zuo K. 2023 Designing artificial synthetic promoters for accurate, smart, and versatile gene expression in plants. *Plant Commun*. **4**, 100558. (doi:10.1016/j.xplc.2023.100558)
57. Cai YM, Kallam K, Tidd H, Gendarini G, Salzman A, Patron NJ. 2020 Rational design of minimal synthetic promoters for plants. *Nucleic Acids Res*. **48**, 11845–11856. (doi:10.1093/nar/gkaa682)
58. Alper H, Fischer C, Nevoigt E, Stephanopoulos G. 2005 Tuning genetic control through promoter engineering. *Proc. Natl Acad. Sci. USA* **102**, 12678–12683. (doi:10.1073/pnas.0504604102)
59. Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015 Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol*. **33**, 831–838. (doi:10.1038/nbt.3300)
60. Zrimec J, Buric F, Kokina M, Garcia V, Zelezniak A. 2021 Learning the regulatory code of gene expression. *Front. Mol. Biosci*. **8**, 673363. (doi:10.3389/fmolb.2021.673363)

61. Zrimec J, Zelezniak A, Gruden K. 2022 Toward learning the principles of plant gene regulation. *Trends Plant Sci*. **27**, 1206–1208. (doi:10.1016/j.tplants.2022.08.010)
62. Ren R, Yu H, Teng J, Mao S, Bian Z, Tao Y, Yau SST. 2024 CAPE: a deep learning framework with Chaos-Attention net for promoter evolution. *Brief. Bioinform.* **25**, bbae398. (doi:10.1093/bib/bbae398)
63. Lin J, Wang X, Liu T, Teng Y, Cui W. 2024 Diffusion-based generative network for *de novo* synthetic promoter design. *ACS Synth. Biol*. **13**, 1513–1522. (doi:10.1021/acssynbio.4c00041)
64. Zhang P, Wang H, Xu H, Wei L, Liu L, Hu Z, Wang X. 2023 Deep flanking sequence engineering for efficient promoter design using DeepSEED. *Nat. Commun*. **14**, 6309. (doi:10.1038/s41467-023-41899-y)
65. Jores T, Tonnies J, Wrightsman T, Buckler ES, Cuperus JT, Fields S, Queitsch C. 2021 Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nat. Plants* **7**, 842–855. (doi:10.1038/s41477-021-00932-y)
66. Mittler R. 2002 Oxidative stress, antioxidants and stress tolerance. *Trends Plant Sci*. **7**, 405–410. (doi:10.1016/s1360-1385(02)02312-9)
67. Kavi Kishor PB, Sreenivasulu N. 2014 Is proline accumulation *per se* correlated with stress tolerance or is proline homeostasis a more critical issue? *Plant Cell Environ*. **37**, 300–311. (doi:10.1111/pce.12157)
68. Yang L, Wen KS, Ruan X, Zhao YX, Wei F, Wang Q. 2018 Response of plant secondary metabolites to environmental factors. *Molecules* **23**, 762. (doi:10.3390/molecules23040762)
69. Weng JK, Lynch JH, Matos JO, Dudareva N. 2021 Adaptive mechanisms of plant specialized metabolism connecting chemistry to function. *Nat. Chem. Biol*. **17**, 1037–1045. (doi:10.1038/s41589-021-00822-6)
70. Fait A, Batushansky A, Shrestha V, Yobi A, Angelovici R. 2020 Can metabolic tightening and expansion of co-expression network play a role in stress response and tolerance? *Plant Sci*. **293**, 110409. (doi:10.1016/j.plantsci.2020.110409)
71. Wang P, Schumacher AM, Shiu SH. 2022 Computational prediction of plant metabolic pathways. *Curr. Opin. Plant Biol*. **66**, 102171. (doi:10.1016/j.pbi.2021.102171)
72. Küken A, Nikoloski Z. 2019 Computational approaches to design and test plant synthetic metabolic pathways. *Plant Physiol*. **179**, 894–906. (doi:10.1104/pp.18.01273)
73. Mutwil M. 2020 Computational approaches to unravel the pathways and evolution of specialized metabolism. *Curr. Opin. Plant Biol*. **55**, 38–46. (doi:10.1016/j.pbi.2020.01.007)
74. Bai W, Li C, Li W, Wang H, Han X, Wang P, Wang L. 2024 Machine learning assists prediction of genes responsible for plant specialized metabolite biosynthesis by integrating multi-omics data. *BMC Genom*. **25**, 418. (doi:10.1186/s12864-024-10258-6)
75. Toubiana D *et al*. 2019 Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. *Commun. Biol*. **2**, 214. (doi:10.1038/s42003-019-0440-4)
76. Moore BM *et al*. 2019 Robust predictions of specialized metabolism genes through machine learning. *Proc. Natl Acad. Sci. USA* **116**, 2344–2353. (doi:10.1073/pnas.1817074116)
77. Hawkins C, Xue B, Yasmin F, Wyatt G, Zerbe P, Rhee SY. 2025 Plant Metabolic Network 16: expansion of underrepresented plant groups and experimentally supported enzyme data. *Nucleic Acids Res*. **53**, D1606–D1613. (doi:10.1093/nar/gkae991)
78. Kanehisa M, Furumichi M, Sato Y, Matsuura Y, Ishiguro-Watanabe M. 2025 KEGG: biological systems database as a model of the real world. *Nucleic Acids Res*. **53**, D672–D677. (doi:10.1093/nar/gkae909)
79. Karp PD *et al*. 2019 The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinformatics* **20**, 1085–1093. (doi:10.1093/bib/bbx085)
80. Dale JM, Popescu L, Karp PD. 2010 Machine learning methods for metabolic pathway prediction. *BMC Bioinform*. **11**, 15. (doi:10.1186/1471-2105-11-15)
81. Bao H, Zhao J, Zhao X, Zhao C, Lu X, Xu G. 2023 Prediction of plant secondary metabolic pathways using deep transfer learning. *BMC Bioinform*. **24**, 348. (doi:10.1186/s12859-023-05485-9)
82. Bernacchi CJ, Ruiz-Vera UM, Siebers MH, DeLucia NJ, Ort DR. 2023 Short- and long-term warming events on photosynthetic physiology, growth, and yields of field grown crops. *Biochem. J*. **480**, 999–1014. (doi:10.1042/bcj20220433)
83. Ort DR *et al*. 2015 Redesigning photosynthesis to sustainably meet global food and bioenergy demand. *Proc. Natl Acad. Sci. USA* **112**, 8529–8536. (doi:10.1073/pnas.1424031112)
84. Naseem M, Osmanoglu Ö, Dandekar T. 2020 Synthetic rewiring of plant CO_2 sequestration galvanizes plant biomass production. *Trends Biotechnol*. **38**, 354–359. (doi:10.1016/j.tibtech.2019.12.019)
85. Erb TJ. 2024 Photosynthesis 2.0: realizing new-to-nature CO_2-fixation to overcome the limits of natural metabolism. *Cold Spring Harb. Perspect. Biol.* **16**, a041669. (doi:10.1101/cshperspect.a041669)
86. Walker BJ, VanLoocke A, Bernacchi CJ, Ort DR. 2016 The costs of photorespiration to food production now and in the future. *Annu. Rev. Plant Biol*. **67**, 107–129. (doi:10.1146/annurev-arplant-043015-111709)
87. Zhao L, Cai Z, Li Y, Zhang Y. 2024 Engineering Rubisco to enhance CO_2 utilization. *Synth. Syst. Biotechnol*. **9**, 55–68. (doi:10.1016/j.synbio.2023.12.006)
88. Santos Correa S, Schultz J, Lauersen KJ, Soares Rosado A. 2023 Natural carbon fixation and advances in synthetic engineering for redesigning and creating new fixation pathways. *J. Adv. Res*. **47**, 75–92. (doi:10.1016/j.jare.2022.07.011)
89. Bierbaumer S, Nattermann M, Schulz L, Zschoche R, Erb TJ, Winkler CK, Tinzl M, Glueck SM. 2023 Enzymatic conversion of CO_2: from natural to artificial utilization. *Chem. Rev*. **123**, 5702–5754. (doi:10.1021/acs.chemrev.2c00581)
90. Marchal DG, Schulz L, Schuster I, Ivanovska J, Paczia N, Prinz S, Zarzycki J, Erb TJ. 2023 Machine learning-supported enzyme engineering toward improved CO_2-fixation of glycolyl-CoA carboxylase. *ACS Synth. Biol.* **12**, 3521–3530. (doi:10.1021/acssynbio.3c00403)
91. Schwander T, Schada von Borzyskowski L, Burgener S, Cortina NS, Erb TJ. 2016 A synthetic pathway for the fixation of carbon dioxide *in vitro*. *Science* **354**, 900–904. (doi:10.1126/science.aah5237)
92. Pandi A *et al*. 2022 A versatile active learning workflow for optimization of genetic and metabolic networks. *Nat. Commun*. **13**, 3876. (doi:10.1038/s41467-022-31245-z)
93. Miller TE *et al*. 2020 Light-powered CO_2 fixation in a chloroplast mimic with natural and synthetic parts. *Science* **368**, 649–654. (doi:10.1126/science.aaz6802)
94. Potters G, Pasternak TP, Guisez Y, Palme KJ, Jansen MAK. 2007 Stress-induced morphogenic responses: growing out of trouble? *Trends Plant Sci*. **12**, 98–105. (doi:10.1016/j.tplants.2007.01.004)
95. Midzi J, Jeffery DW, Baumann U, Rogiers S, Tyerman SD, Pagay V. 2022 Stress-induced volatile emissions and signalling in inter-plant communication. *Plants* **11**, 2566. (doi:10.3390/plants11192566)
96. Khait I *et al*. 2023 Sounds emitted by plants under stress are airborne and informative. *Cell* **186**, 1328–1336.(doi:10.1016/j.cell.2023.03.009)
97. Singh AK, Ganapathysubramanian B, Sarkar S, Singh A. 2018 Deep learning for plant stress phenotyping: trends and future perspectives. *Trends Plant Sci*. **23**, 883–898. (doi:10.1016/j.tplants.2018.07.004)
98. Singh A, Jones S, Ganapathysubramanian B, Sarkar S, Mueller D, Sandhu K, Nagasubramanian K. 2021 Challenges and opportunities in machine-augmented plant stress phenotyping. *Trends Plant Sci*. **26**, 53–69. (doi:10.1016/j.tplants.2020.07.010)

99. Anshori MF, Dirpan A, Sitaresmi T, Rossi R, Farid M, Hairmansis A, Purwoko B, Suwarno WB, Nugraha Y. 2023 An overview of image-based phenotyping as an adaptive 4.0 technology for studying plant abiotic stress: a bibliometric and literature review. *Heliyon* **9**, e21650. (doi:10.1016/j.heliyon.2023.e21650)
100. Römer C *et al*. 2012 Early drought stress detection in cereals: simplex volume maximisation for hyperspectral image analysis. *Funct. Plant Biol* **39**, 878–890. (doi:10.1071/fp12060)
101. Kaneda Y, Shibata S, Mineno H. 2017 Multi-modal sliding window-based support vector regression for predicting plant water stress. *Knowl. Based Syst*. **134**, 135–148. (doi:10.1016/j.knosys.2017.07.028)
102. Das Choudhury S, Saha S, Samal A, Mazis A, Awada T. 2023 Drought stress prediction and propagation using time series modeling on multimodal plant image sequences. *Front. Plant Sci*. **14**, 1003150. (doi:10.3389/fpls.2023.1003150)
103. Das Choudhury S, Guadagno CR, Bashyam S, Mazis A, Ewers BE, Samal A, Awada T. 2024 Stress phenotyping analysis leveraging autofluorescence image sequences with machine learning. *Front. Plant Sci*. **15**, 1353110. (doi:10.3389/fpls.2024.1353110)
104. Arya S, Sahoo RN, Sehgal VK, Bandyopadhyay K, Rejith RG, Chinnusamy V, Kumar S, Kumar S, Manjaiah KM. 2024 High-throughput chlorophyll fluorescence image-based phenotyping for water deficit stress tolerance in wheat. *Plant Physiol. Rep*. **29**, 278–293. (doi:10.1007/s40502-024-00783-7)
105. Zagorščak M. 2024 Integration of multi-omics and deep phenotyping provides novel insights into multiple abiotic stress responses in potato. *BioRxiv* 2024.07.18.604140. (doi:10.1101/2024.07.18.604140)
106. Zhang H, Zhu J, Gong Z, Zhu JK. 2022 Abiotic stress responses in plants. *Nat. Rev. Genet*. **23**, 104–119. (doi:10.1038/s41576-021-00413-0)
107. Villalobos-López MA, Arroyo-Becerra A, Quintero-Jiménez A, Iturriaga G. 2022 Biotechnological advances to improve abiotic stress tolerance in crops. *Int. J. Mol. Sci*. **23**, 12053. (doi:10.3390/ijms231912053)
108. Wang X, Wang H, Liu S, Ferjani A, Li J, Yan J, Yang X, Qin F. 2016 Genetic variation in *ZmVPP1* contributes to drought tolerance in maize seedlings. *Nat. Genet*. **48**, 1233–1241. (doi:10.1038/ng.3636)
109. Li XM *et al*. 2015 Natural alleles of a proteasome α2 subunit gene contribute to thermotolerance and adaptation of African rice. *Nat. Genet*. **47**, 827–833. (doi:10.1038/ng.3305)
110. Monson RK, Trowbridge AM, Lindroth RL, Lerdau MT. 2022 Coordinated resource allocation to plant growth–defense tradeoffs. *New Phytol*. **233**, 1051–1066. (doi:10.1111/nph.17773)
111. Khahani B, Tavakol E, Shariati V, Fornara F. 2020 Genome wide screening and comparative genome analysis for meta-QTLs, ortho-MQTLs and candidate genes controlling yield and yield-related traits in rice. *BMC Genom*. **21**, 294. (doi:10.1186/s12864-020-6702-1)
112. Saini DK, Srivastava P, Pal N, Gupta PK. 2022 Meta-QTLs, ortho-meta-QTLs and candidate genes for grain yield and associated traits in wheat (*Triticum aestivum* L.). *Theor. Appl. Genet*. **135**, 1049–1081. (doi:10.1007/s00122-021-04018-3)
113. Yang Y *et al*. 2021 Large-scale integration of meta-QTL and genome-wide association study discovers the genomic regions and candidate genes for yield and yield-related traits in bread wheat. *Theor. Appl. Genet*. **134**, 3083–3109. (doi:10.1007/s00122-021-03881-4)
114. Izquierdo P, Kelly JD, Beebe SE, Cichy K. 2023 Combination of meta-analysis of QTL and GWAS to uncover the genetic architecture of seed yield and seed yield components in common bean. *Plant Genome* **16**, e20328. (doi:10.1002/tpg2.20328)
115. Shen K, Bunescu R, Wyatt SE. 2020 Mining functionally related genes with semi-supervised learning. *arXiv* 2011.03089. (doi:10.48550/arXiv.2011.03089)
116. Zhuang F. 2020 A comprehensive survey on transfer learning. *arXiv* 1911.02685. (doi:10.48550/arXiv.1911.02685)
117. Reiser L, Bakker E, Subramaniam S, Chen X, Sawant S, Khosa K, Prithvi T, Berardini TZ. 2024 The *Arabidopsis* information resource in 2024. *Genetics* **227**, iyae027. (doi:10.1093/genetics/iyae027)
118. Müller HM, Van Auken KM, Li Y, Sternberg PW. 2018 Textpresso Central: a customizable platform for searching, text mining, viewing, and curating biomedical literature. *BMC Bioinform*. **19**, 94. (doi:10.1186/s12859-018-2103-8)
119. Kishore R *et al*. 2020 Automated generation of gene summaries at the Alliance of Genome Resources. *Database* **2020**, a037. (doi:10.1093/database/baaa037)
120. Lähnemann D *et al*. 2020 Eleven grand challenges in single-cell data science. *Genome Biol*. **21**, 31. (doi:10.1186/s13059-020-1926-6)
121. Erfanian N *et al*. 2023 Deep learning applications in single-cell genomics and transcriptomics data analysis. *Biomed. Pharmacother*. **165**, 115077. (doi:10.1016/j.biopha.2023.115077)
122. Petegrosso R, Li Z, Kuang R. 2020 Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Brief. Bioinformatics* **21**, 1209–1223. (doi:10.1093/bib/bbz063)
123. Wu Z. 2021 A comprehensive survey on graph neural networks. *IEEE Trans. Neural Networks Learn. Syst* **32**, 4–24. (doi:10.1109/TNNLS.2020.2978386)
124. Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, Wang L, Li C, Sun M. 2020 Graph neural networks: a review of methods and applications. *AI Open* **1**, 57–81. (doi:10.1016/j.aiopen.2021.01.001)
125. Singhal R, Izquierdo P, Ranaweera T, Segura Abá K, Brown, Brianna B, Lehti-Shiu MD *et al*. 2025 Supplementary material from: Using supervised machine-learning approaches to understand abiotic stress tolerance and design resilient crops. Figshare (doi:10.6084/m9.figshare.c.7758939)