A review of artificial intelligence-assisted omics techniques in plant defense: current trends and future directions

Sneha Murmu[1], Dipro Sinha[1], Himanshushekhar Chaurasia[2], Soumya Sharma[1], Ritwika Das[1], Girish Kumar Jha[1] and Sunil Archak[3]*

[1]Indian Agricultural Statistics Research Institute, Indian Council of Agricultural Research (ICAR), New Delhi, India, [2]Central Institute for Research on Cotton Technology, Indian Council of Agricultural Research (ICAR), Mumbai, India, [3]National Bureau of Plant Genetic Resources, Indian Council of Agricultural Research (ICAR), New Delhi, India

Plants intricately deploy defense systems to counter diverse biotic and abiotic stresses. Omics technologies, spanning genomics, transcriptomics, proteomics, and metabolomics, have revolutionized the exploration of plant defense mechanisms, unraveling molecular intricacies in response to various stressors. However, the complexity and scale of omics data necessitate sophisticated analytical tools for meaningful insights. This review delves into the application of artificial intelligence algorithms, particularly machine learning and deep learning, as promising approaches for deciphering complex omics data in plant defense research. The overview encompasses key omics techniques and addresses the challenges and limitations inherent in current AI-assisted omics approaches. Moreover, it contemplates potential future directions in this dynamic field. In summary, AI-assisted omics techniques present a robust toolkit, enabling a profound understanding of the molecular foundations of plant defense and paving the way for more effective crop protection strategies amidst climate change and emerging diseases.

1 Introduction

Plant defense against both biotic (living organisms like pathogens and pests) and abiotic (environmental factors such as drought, salinity, and extreme temperatures) stress is of paramount importance in ensuring global food security. With the world's population steadily growing, the demand for crops is increasing, making efficient plant protection strategies a

critical need. The ability to safeguard crops from the ravages of diseases, pests, and adverse environmental conditions is essential to maintain agricultural productivity and secure the global food supply. The substantial impact of these stressors is underscored by the extensive economic losses, as evidenced by the multi-billion-dollar reductions in crop yields (FAO 2017 report).

Biotic stressors, such as pathogens and pests, pose a constant threat to crop health. Pathogens, including fungi, bacteria, and viruses, can devastate entire plant populations, leading to substantial economic losses and food shortages. Likewise, pests, ranging from insects to nematodes, have the potential to decimate crops, leading to decreased yields and quality.

Food crops around the world face substantial yield reductions due to microbial diseases and pest infestations. These losses are significant, with rice experiencing an average loss of 30.3%, maize at 22.6%, wheat at 21.5%, soybeans at 21.4%, and potatoes at 17.2% (Savary et al., 2019). Plant diseases can be especially devastating, leading to yield reductions of up to 50% in specific regions, particularly impacting small-scale farmers and posing substantial economic challenges. Additionally, plant diseases negatively influence species diversity, increase the costs associated with disease control measures, and even have repercussions on human health (Ristaino et al., 2021). The emergence of new plant diseases and pest outbreaks carries substantial economic implications for agriculture, posing threats to food security, national stability, and public health (Anderson et al., 2004). In the coming years, it is expected that the changing distribution of pathogens due to climate variations and increased global trade will result in a higher prevalence and greater severity of emerging plant diseases (Bebber et al., 2013). A notable recent example is the outbreak of coffee rust, caused by *Hemileia vastatrix*, in Central America, which led to significant crop losses and economic crises (Avelino et al., 2015).

On the other hand, abiotic stress factors are non-living elements that challenge plant growth and survival. These include prolonged droughts, extreme temperatures, soil salinity, and heavy metal contamination. The impacts of abiotic stressors are often subtle and insidious, affecting crop yields, nutritional content, and overall plant health. Efforts to combat both biotic and abiotic stressors have traditionally relied on a combination of methods, including conventional breeding, chemical treatments, and agronomic practices. However, these approaches are often reactive and may not provide effective protection, especially in the face of emerging pathogens or rapidly changing environmental conditions.

Recently, advanced omics techniques, have revolutionized the exploration of molecular-level stress mechanisms in plants (Shen et al., 2022). These methods provide extensive information, revealing intricate networks involving genes, proteins, and metabolites during plant defense against biotic stress. However, the substantial data generated by these omics technologies poses a significant challenge in terms of analysis and interpretation, necessitating the development of highly effective computational tools.

Artificial Intelligence (AI) has emerged as a potent instrument for unraveling vast omics datasets and understanding intricate mechanisms underlying plant responses to stress. These techniques offer a deeper understanding of the genetic, molecular, physiological, and phenotypic aspects of plant defense, enabling the development of

novel strategies to bolster crop resilience and mitigate stress-induced damage. In traditional plant defense research, the complex networks of genes, proteins, and metabolites involved in stress responses posed challenges due to the high volume and complexity of biological data. However, the advent of AI-assisted omics techniques has ushered in innovative solutions to tackle and interpret this vast data landscape. These techniques encompass diverse AI methodologies, efficiently unraveling and modeling the intricate relationships between molecular components and stress responses.

In this review, we intend to provide an overview of different omics studies involving various stress factors in plants. As we navigate through the intersection of AI and plant omics, we explore cutting-edge developments in the realm of plant defense against both biotic and abiotic stressors and confront its challenges. This review critically explores the advantages of AI over traditional methods, delves into the challenges of AI in plant omics, and future directions in plant defense research, highlighting the potential for sustainable agricultural practices that enhance crop protection, stress tolerance, and global food security.

2 Different omics in plant defense research

Plant defense research encompasses various "omics" technologies, each offering unique insights into the molecular mechanisms underlying plant responses to pathogens and environmental stresses. In the following sections, we'll provide a brief overview of the primary types of data each omics approach can provide when applied to research on plant stress.

2.1 Genomics

Plants, as stationary organisms, have evolved sophisticated defense mechanisms against various stressors, including biotic and abiotic factors. The combination of genomics and AI has become a potent tool for unraveling the genetic foundations of plant defense. This review explores the current and future applications of AI-assisted genomics in understanding both biotic and abiotic stress responses in plants. It covers the identification of resistance genes, characterization of defense pathways, and improvement of stress tolerance. The challenges, ethical considerations, and potential breakthroughs in this evolving field are also discussed.

Genome-wide association studies (GWAS) play a crucial role in genomic strategies to enhance crop resilience against abiotic stress. Mangin et al. (2017) illustrate the significance of GWAS in evaluating abiotic stress impacts on sunflower oil content. Previous studies identified Quantitative Trait Loci (QTLs) associated with maize yield under heat and water stress (Millet et al., 2016). Environmental variables in GWAS investigations revealed Single Nucleotide Polymorphisms (SNPs) linked to sorghum drought stress, with 213 genomic regions associated with drought tolerance (Lasky et al., 2015; Spindel et al., 2018).

Epigenetics, involving heritable modifications beyond DNA sequences, combines with genomics in the emerging field of

epigenomics. This integration unveils genetic regulation in cellular responses to stress, with epigenomic processes responding to environmental conditions and stressors. Genome-level investigations are necessary to scrutinize these phenomena across developmental stages or assess deviations due to plant diseases (Callinan and Feinberg, 2006; Muthamilarasan et al., 2019). The revolutionary CRISPR-Cas9 technology, originating from bacteria as a defense mechanism against viruses, has transformed genome editing. CRISPR-Cas systems extensively edit eukaryotic genomes, providing opportunities to engineer crop plants for enhanced resilience against both abiotic and biotic stresses (Kumar and Jain, 2015).

2.2 Transcriptomics

Transcriptomics, the study of an organism's complete set of RNA transcripts within specific cells or tissues (transcriptome), is a dynamic field with potential for analyzing gene expression responses to various stimuli over defined timeframes (Raza et al., 2021). Transcriptome profiling, the approach used in this field, allows the investigation of gene expression differences, providing insights into the functions of specific genes.

In various crops like sorghum and rice, transcriptome studies have identified gene sets with altered expression in response to stressors such as drought, heat, osmotic stress, and hormonal treatments (Dugas et al., 2011; Jin et al., 2013; Johnson et al., 2014). These analyses are crucial for understanding gene expression changes during growth and stress responses, offering valuable insights for functional studies. Transcriptomics has proven significant in unraveling stress responses and developmental processes in crops, as demonstrated in RNA-seq studies in foxtail millet and sweet potato, revealing tissue-specific gene expression responses to abiotic and biotic stress (Li et al., 2017). The application of RNA-seq in rice, maize, and rapeseed oil research has aided in identifying genes responsive to drought stress (Bhardwaj et al., 2015). Comparative transcriptomic analysis enables exploration of distinct gene expression profiles across diverse crop species facing stress, identifying shared genes and revealing intricate cross-talk pathways (Li et al., 2013; Zhu et al., 2013). These findings emphasize the significance of regulatory networks governing stress tolerance genes, offering potential for enhancing crop traits through genetic improvement.

2.3 Proteomics

Proteomics, as a comprehensive approach to studying proteins, plays a crucial role in understanding how plants respond to both biotic and abiotic stresses. The four main aspects of proteomics—sequence, structural, functional, and expression proteomics—offer a holistic view of the complex interactions within plant cells (Aizat and Hassan, 2018). In sequence proteomics, scientists identify amino acid sequences using advanced techniques like high-performance liquid chromatography, providing insights into the building blocks of proteins (Twyman et al., 2013). Structural proteomics focuses on understanding the three-dimensional structures and functions of proteins, employing various methods like computer-based modeling, NMR, crystallization, electron microscopy, and X-ray diffraction (Woolfson et al., 2018). Functional proteomics delves into the roles of proteins, employing methodologies such as Y2H assays and protein microarray profiling to decipher the specific functions of different proteins within the cellular context. Quantitative proteomics, exemplified by the iTRAQ method, allows researchers to measure changes in protein expression levels in response to stresses, providing valuable information about how plants react to environmental challenges (Liu et al., 2015; Zhu et al., 2018; Yang et al., 2020).

In the context of plant responses to biotic stress, proteomics has proven pivotal. Studies involving Vitis species and other crops showcase the ability of proteomic analyses to identify stress-responsive proteins and uncover translational modifications like phosphorylation and ubiquitination. This information aids in understanding the intricate molecular dynamics underlying plant defense mechanisms and pathogen virulence (Mosa et al., 2017). Similarly, in the realm of abiotic stress, such as drought, proteomics reveals proteins associated with stress response in crops like wheat. The integration of proteomics and phosphoproteomics explores diverse functions in response to various stressors, contributing to the identification of both resistant and susceptible crop cultivars against these challenges. Additionally, the combination of proteomics with other omics disciplines like metabolomics and functional genomics enhances our understanding of stress biology, facilitating the identification of molecular markers for breeding programs (Margaria et al., 2013; Yang et al., 2013; Zhang et al., 2014). The various proteomic techniques, including LC-MS/MS, MALDI-TOF, SDS-PAGE, and iTRAQ, are extensively applied in different crops to investigate their responses to both biotic and abiotic stress conditions (Mohammadi et al., 2012; Ramalingam et al., 2015). The insights gained from proteomic studies significantly contribute to unraveling the molecular mechanisms by which plants adapt to environmental challenges, ultimately leading to advancements in crop yield improvement and stress resilience.

2.4 Metabolomics

Metabolomics, a study of metabolites in biological systems, is crucial for understanding the plant metabolome and revealing regulatory mechanisms under stress conditions. This field, integrated with next-generation sequencing, provides insights into molecular responses in crops, offering a broader perspective on biochemical processes influencing gene functionality.

In plant defense against stress and pathogens, metabolites play a vital role, identified through gas chromatography-mass spectrometry as biomarkers in rice varieties facing the GMB1 pathogen. Similar strategies reveal metabolite accumulation in response to other pathogens in rice and barley, showcasing the importance of metabolomics in understanding plant responses to biotic stress.

Wheat crops also exhibit the presence of phenylpropanoid and phenolic metabolites in response to biotic stress. Metabolomics is

particularly vital in plant systems due to the abundant production of metabolites. Secondary metabolites like polyamines, identified in rice crops under drought stress, highlight the relevance of environmental metabolomics in understanding plant responses to abiotic stresses. Various metabolomics techniques, including LC/GC-MS, GC/EI-TOF-MS, HPLC, and NMR, have been widely employed in crops like rice, tomato, maize, and soybean, providing valuable insights into their responses to both abiotic and biotic stress conditions.

2.5 Phenomics

Plant phenomics involves systematically acquiring and analyzing multi-dimensional traits across various crop growth stages, from cellular to field levels. This process relies on a three-step approach: trait identification, data conversion into quantifiable measurements, and computational methodologies for analysis. High-throughput phenotyping platforms are crucial in the initial phase, while computational strategies, particularly machine learning (ML) algorithms, play a pivotal role in subsequent stages. The performance of crop phenotypes is intricately linked to genetic factors and environmental conditions. The continuous evolution of sensors, imaging technologies, and analytical methodologies has led to the development of numerous dedicated infrastructure platforms for phenotyping.

Abiotic stresses such as drought, salinity, and nutrient deficiencies pose significant challenges to crop production, eliciting complex plant responses. Phenotyping for stress resistance is imperative for breeding resilient crops. Drought stress, marked by reduced water availability, can be evaluated using ground-based platforms equipped with thermometer sensors and RGB cameras. Unmanned Aerial Vehicles (UAVs) integrated with thermal cameras facilitate quicker scanning of larger plots for identifying drought-resistant genotypes. Salinity stress, impacting stomatal conductance, is observed through visible to near-infrared spectral reflectance images. Scanalyzer3D aids in
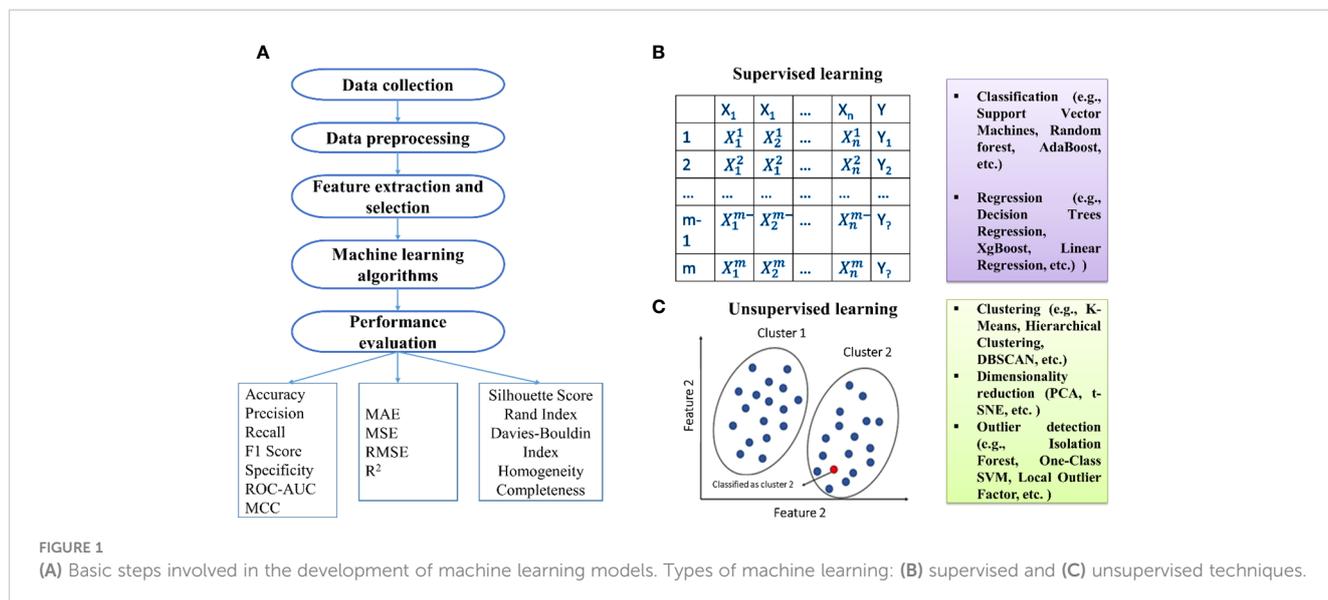
characterizing salinity tolerance mechanisms. Image-based methods, encompassing RGB and fluorescence imaging, assess tissue ion concentrations to gauge salinity tolerance. Hyperspectral imaging, coupled with ML, predicts traits associated with salinity stress. Crop nutrient deficiencies, especially nitrogen, affect chlorophyll content, growth, and disease susceptibility, with monitoring conducted through sensors like RGB, multispectral, and hyperspectral sensors. Mobile platforms incorporating these sensors estimate nitrogen content efficiently. Agriculture confronts threats from diseases and pests, and the integration of resistance genes presents a cost-effective strategy. Biotic stress induces changes in various plant characteristics, and advanced phenotyping platforms utilizing optical sensors effectively detect and manage biotic stress factors in crops.

3 Basics of AI techniques

AI involves creating computer systems to perform tasks associated with human intelligence, such as learning, problem-solving, and decision-making. ML, a subset of AI, focuses on developing algorithms and statistical models that enable computers to perform tasks without explicit programming. DL, a specialized field within ML, involves training artificial neural networks to mimic the human brain's structure, utilizing deep architectures with multiple layers for automatic hierarchical feature extraction.

The initial phase of ML includes data collection, especially in sequencing data like RNA sequencing studies (Figure 1). Denoising methods enhance expression recovery. Supervised ML uses diverse features for training data representation, including amino acid sequence information and physicochemical properties (Sperschneider et al., 2018). Feature selection is crucial, and methods fall into three categories: filter, wrapper, and embedding methods (Guyon and Elisseeff, 2006; Guyon et al., 2008; Khalid et al., 2014).

Algorithm selection is pivotal, and ML algorithms can be categorized into supervised, semi-supervised, and unsupervised. Supervised methods establish relationships between input factors

FIGURE 1
(A) Basic steps involved in the development of machine learning models. Types of machine learning: **(B)** supervised and **(C)** unsupervised techniques.

and outcomes based on training examples. Unsupervised methods, primarily clustering, identify data patterns without relying on known outcomes. Practical algorithms include SVM, DT, RF, ANN, and NB for supervised learning, while k-means, independent component analysis, and hierarchical clustering are used for unsupervised learning. Semi-supervised learning handles input data with both labeled and unlabeled information, with examples like the label propagation algorithm.

3.1 Unsupervised ML

Unsupervised ML methods fulfill two primary functions: clustering, which groups data based on similarity, and dimension reduction, generating representative features from numerous variables. A widely utilized clustering method is k-means, aiming to create non-overlapping clusters of observations. Principal Component Analysis (PCA) is a common technique for dimension reduction, transforming high-dimensional observations into a smaller set of uncorrelated principal components (PCs) to simplify subsequent analyses.

3.2 Supervised ML

Support Vector Machine (SVM) is a supervised learning method created by Vapnik (1999) for binary classification tasks. It operates in an n-dimensional space, forming a hyperplane to maximize the margin between distinct data classes. The choice of kernel functions, such as linear or nonlinear options like polynomial or radial basis, significantly influences its performance.

The k-nearest-neighbor (KNN) algorithm, introduced by Altman (1992), is another supervised learning technique that classifies data points by identifying the 'k' nearest neighbors with known labels. The classification is based on a majority vote among these neighbors. While user-friendly, traditional KNN methods may have longer computation times (Borah et al., 2020).

The Decision Tree (DT) classifier, developed by Quinlan (1986), follows a branch-test approach. It recursively partitions data based on attributes until a specified stopping condition is met, creating a tree-like structure. The classification path can be traced from the root node to each leaf node (Schietgat et al., 2010).

Random Forest (RF), introduced by Breiman in 2001, is an ensemble algorithm utilizing a group of DTs to achieve a consensus on accurate classification. Classification trees are constructed by randomly selecting from training datasets, and the predictions from each tree are combined to provide an overall prediction for each observation (Breiman, 2001).

3.3 Major deep learning architectures

ML has gained widespread popularity, and DL methods, particularly associated with Artificial Neural Network (ANN) architectures, are gaining increasing attention. DL autonomously learns from raw input data, capturing intricate patterns without extensive domain expertise. Unlike traditional ML, which relies on discrete or continuous output predictions based on data counts or measurements, DL excels in direct feature learning from input datasets, eliminating the need for conventional feature engineering (Bonetta and Valentino, 2020).

ANNs, dating back to McCulloch and Pitts (1943), mimic biological neurons and feature a learning process facilitated by synaptic connections. Deep Neural Networks (DNNs) encompass multiple hidden layers, effectively forming a complex structure. Recurrent Neural Networks (RNNs), introduced by Sperduti and Starita (1997), are suitable for supervised learning with feedback loops for cyclic data processing. Convolutional Neural Networks (CNNs), introduced by Lecun et al. (1998), excel in identifying relevant features without human supervision, while Graph Convolutional Networks (GCNs), introduced by Schaffer (1989), handle intricate problems through complex architectures (Zhou et al., 2022).

Transformers, rooted in a self-attention mechanism, find application in natural language processing tasks like text translation, improving task parallelization (Vaswani et al., 2017). Ensemble classifiers enhance decision-making by combining outputs from different models, introduced by Dietterich (2000). Clustering-based methods, like the k-means algorithm, provide an unsupervised approach for predicting protein functions by exploiting direct and indirect interactions (Hou, 2017; Yan and Wang, 2022).

Each algorithm, including ML and DL models, offers unique capabilities catering to the complexities of omics datasets. Understanding their distinctive strengths and weaknesses is crucial before exploring their applications. Table 1 provides a concise overview summarizing the key attributes defining their performance.

3.4 Validation strategies

ML predictions in plant genomics research can be validated through a variety of methods. Xavier (2021) emphasized the importance of cross-validation in comparing different algorithms for genomic prediction. K-fold cross-validation, a widely used method, involves randomly dividing training samples into k subsets, reserving one for validation, and using the others for training (Sun et al., 2020). Evaluation metrics, derived from the confusion matrix, include sensitivity, specificity, accuracy, precision, F1-score, and Matthews correlation coefficient (MCC). Sensitivity gauges correctly predicted positives, specificity assesses correctly predicted negatives, accuracy reflects overall correct predictions, and precision measures correctly predicted positives among TP and FP. The F1-score combines precision and recall, while MCC is valuable for imbalanced datasets (Jiao and Du, 2016). The receiver operating characteristic (ROC) curve, evaluated with false positive rate (FPR) and true positive rate (TPR), and the area under the ROC

TABLE 1 Strengths and weaknesses of different AI-methods.

Method	Pros	Cons	Task
k-nearest-neighbor	■ Simple implementation for handling multi-class problems. ■ Consistency in selected hyperparameters. ■ Utilizes a non-parametric algorithm.	■ Slower processing speed. ■ More suitable for datasets with a limited number of input variables. ■ Inefficiency with imbalanced data. ■ Prone to sensitivity issues with outliers.	■ Classification ■ Regression
Decision Tree	■ Easy data pre-processing without the need for scaling or normalization. ■ Capable of handling both numerical and categorical features effectively. ■ Facilitates a clear visual representation of output.	■ Longer training times, contributing to slower processing. ■ Prone to overfitting issues. ■ For complex or large datasets, may provide inadequate predictions. ■ Not ideally suited for datasets which are not balanced.	■ Classification ■ Regression
Random Forest	■ Capable of adapting to dynamic neural networks. ■ Computational power suitable for handling non-linear systems.	■ Problem of exploding and vanishing gradient for long sequences. ■ Training process tends to be slow and complex.	■ Classification ■ Regression
Support Vector Machine	■ Efficient when dealing with distinctly separable classes. ■ Well-suited for two-class classification tasks. ■ Effective for high dimensional datasets. ■ Relatively memory-efficient.	■ For noisy and large datasets, it may not be suitable. ■ Requires preprocessing of data. ■ Prone to the risk of overfitting. ■ Involves computationally costly processes. ■ Output has low interpretability.	■ Classification ■ Regression
Ensemble Classifier	■ Achieves higher predictive accuracy compared to individual models. ■ Capable of handling both linear and non-linear data. ■ Reduces overfitting and bias. ■ Produces more stable and reduced noisy predictions.	■ Computationally expensive. ■ Interpretability is limited. ■ Memory intensive.	■ Classification ■ Regression ■ Clustering
Graphical Neural Network	■ Consistency in parameter use during the training iteration. ■ Cost-effective data storage. ■ Adaptive learning of the importance of	■ Algorithm's processes can be untraceable. ■ High computational costs.	■ Classification ■ Regression ■ Clustering

(Continued)

TABLE 1 Continued

Method	Pros	Cons	Task
	neighbours in a graph-based system.		
Clustering Model	■ Efficient and requires fewer computations. ■ Intuitive without the need for pre-set clusters. ■ Identifies outliers without predetermined clusters. ■ Offers more flexibility compared to k-means. ■ Insensitivity to the choice of distance metric, allowing for hierarchy visualization.	■ Inconsistency in dealing with outliers. ■ Challenges in selecting an optimal number of clusters. ■ For high dimension datasets, there may be inconsistency in performance. ■ Utilizes all data points that are available. ■ Lower efficiency ■ Increased time complexity.	■ Clustering

curve (AUC) serve as performance measures, with higher AUC indicating superior predictor performance (Xu-Hui et al., 2009).

Individual-based models demand a context-oriented approach due to their complex and variable interaction structure (Leye et al., 2009; Kubicek et al., 2015). This approach involves separately assessing different model levels and employing various techniques, including visual inspection, statistical comparison, expert involvement, and experimental validation. In the realm of engineering and scientific models, a proposed statistical validation approach links validation experiments to the target application and considers the importance of measurements (Hills and Leslie, 2003). When applied context-dependently, these strategies enhance the support for hypotheses generated by the model. Maron et al. (2014) underscored the necessity of experiments exploring the role of abiotic factors in plant-animal interactions. Wang et al. (2018) introduced pattern-oriented modeling as an effective means to verify and validate functional-structural plant models, showcasing its predictive capabilities in plant growth. Abele et al. (2013) introduced an ontology-based approach to validate plant models, ensuring their accuracy through Semantic Web reasoning. Bylesjö et al. (2007) discussed the O2PLS method for integrating transcript and metabolite data in plant biology, providing a means to validate and interpret models. Together, these studies emphasize the significance of experimental and computational validation strategies in plant omics research.

For ML models applied to plant omic data, context specificity is crucial, as highlighted by Isewon et al. (2022), especially for improving agronomic traits and developing resilient crop varieties. Silva et al. (2019) further emphasizes the necessity of ML approaches in plant molecular biology, particularly in the analysis of pathogen-effector genes. Fukushima et al. (2009) underscores the importance of integrating multiple omics data, including metabolomics, to reconstruct complex networks in plant systems. Collectively, these studies support the use of context-specific ML models and the integration of omics data for a more comprehensive understanding of plant biology.

The effectiveness of ML prediction in plant stress omics research finds objective validation through various methods. John et al. (2022) compare classical and ML-based phenotype prediction methods, noting the varying performance of different models in real-world data. Ghosal et al. (2017) enhance the interpretability of ML models by applying a ML framework to identify and classify foliar stresses in soybean plants, isolating visual symptoms for each stress. Singh et al. (2018) emphasize the importance of standardizing visual assessments, deploying imaging techniques, and using ML tools for data assimilation and feature identification in plant stress phenotyping. Together, these studies underscore the potential of ML in accurately predicting and identifying plant stress, with a focus on interpretability and standardization.

4 Why choose machine learning for plant-omics data over traditional methods?

ML is increasingly preferred over traditional methods in plant-omics data analysis due to its adept handling of large, complex datasets. High-throughput sequencing technologies have ushered in a wealth of information, enabling biologists to explore intricate associations, decode stress responses in plants, and unravel complexities in genomic responses (Singh et al., 2016). However, challenges such as high dimensionality, uncertainty, and non-independence among variables in plant omics data have emerged. Traditional statistical models face limitations in handling this complexity (Altman and Krzywinski, 2018; Niazian and Niedbała, 2020).

ML, especially DL, has proven efficient in overcoming these challenges, providing accurate analyses of plant characteristics affected by genotype and environment interactions (Arsenovic et al., 2019). Unsupervised and semi-supervised ML algorithms have been applied to plant systems biology, facilitating big data analysis without the need for large labeled training sets (Yan and Wang, 2022). ML's application extends to improving plant agronomic traits through the integration of large omics data (Isewon et al., 2022). Studies by Farooq et al. (2022), Isewon et al. (2022), and Silva et al. (2019) highlight the superiority of ML methods, particularly decision tree-based ensemble models (Gokalp and Tasci, 2019), in genomic prediction and integrative analysis of plant omics data. ML's potential in deciphering complex interactions in plant molecular biology, including pathogen effector genes and plant immunity, is underscored (Silva et al., 2019).

In transcriptomics, ML methods stand out for enhancing the sensitivity of differential expression gene identification (Wang et al., 2018). However, the use of non-linear ML models in differential expression analysis may have limitations, leading to the recommendation of eXplainable Artificial Intelligence for model interpretation and gene set identification (Sabbatini and Calegari, 2023). The integration of ML with traditional biological information is emphasized for learning biological dynamics from

large datasets, complementing traditional modeling approaches (Xu and Jackson, 2019; Gilpin et al., 2020). Various ML tools, including tree-based methods, Bayesian models, network-based fusion methods, kernel methods, matrix factorization models, and deep neural networks, play a crucial role in connecting multi-view biological data (Li et al., 2016).

ML's proficiency in multivariate analysis is advantageous for considering numerous variables simultaneously, leading to the discovery of new biomarkers and predictive model development (Reel et al., 2021). Its application in improving agronomic traits in plant omics research is evident, although challenges in fully realizing the potential of integrating multiomics data remain, with scaling difficulties being a major obstacle (Noor et al., 2019). The complexity of high-dimensional omics data necessitates sophisticated methods for feature selection and information extraction. ML's advantage lies in its ability to discern and prioritize the most relevant features, as demonstrated by Du et al. (2019) in the analysis of RNA-sequencing data related to salt stress response in rice. The use of ML-based feature selection methods, including principal component analysis and LASSO, effectively revealed submodules associated with observed traits.

ML excels in prediction and classification tasks, allowing researchers to forecast phenotypic outcomes and identify potential biomarkers. Its scalability is crucial for large-scale omics data, setting it apart from traditional methods facing computational challenges. In conclusion, the shift towards ML in plant omics research is driven by its unique strengths in addressing data intricacies, enabling predictive modeling, and facilitating an exploratory approach to data analysis.

Pattern recognition is another strength of ML, enabling the discovery of intricate patterns and associations within complex datasets. In omics research, where uncovering subtle patterns may provide novel insights into biological mechanisms, this capability is highly valuable. Moreover, ML is adaptable to the heterogeneity often observed in omics datasets due to biological variability and technical differences. Its flexibility and generality, with less reliance on assumptions about data distribution, make it suitable for various data types and experimental designs. The exploratory nature of ML, facilitating the uncovering of hidden patterns and relationships, is crucial in omics research. This aspect allows researchers to generate hypotheses and identify novel avenues for further investigation.

In the realm of multi-omics analysis, the primary objective is constructing Gene Regulatory Networks (GRNs). While ChIP-seq experiments for profiling Transcription Factors' (TFs) binding sites are limited in plants, the inference of GRNs heavily relies on expression data (Bubb and Deal, 2020). Traditional correlation-based methods and the Mutual Information (MI) algorithm face challenges in distinguishing regulatory direction and considering temporal delays between gene expressions (Banf and Rhee, 2017; Redekar et al., 2017; Haque et al., 2019). To overcome these limitations, Probabilistic Graphical Models (PGM), such as GENIST and JRmGRN, have been introduced, though they require high spatiotemporal resolution in expression data (de Luis Balaguer et al., 2017; Deng et al., 2018).

ML has revolutionized the inference of GRNs, integrating multi-omics data to enhance accuracy (Walley et al., 2016; Ko and Brandizzi, 2020). iDREM (interactive dynamic regulatory events miner), employing a hidden Markov model, reconstructs temporal GRNs in response to biotic and abiotic stresses using transcriptomic, proteomic, and epigenomic datasets (Ding et al., 2018). With the emergence of single-cell RNA sequencing (scRNA-seq), tools like GRNBoost2, based on the GENIE3 framework, facilitate cell-specific GRN inference (Moerman et al., 2019). The SCENIC analytical pipeline, incorporating multiple tools, efficiently analyzes datasets within 2 hours comprising of 50,000 cells and 10,000 genes (Van de Sande et al., 2020).

In summary, the shift towards ML in plant omics research is driven by its unique strengths in addressing the intricacies of omics data, accommodating multiple variables, integrating diverse datasets, providing predictive modeling and classification capabilities, and facilitating an exploratory approach to data analysis.

5 Application of AI in plant omics against stress

AI-assisted omics techniques in plant defense research represent a cutting-edge approach, combining advanced molecular technologies with AI for a deeper understanding of how plants respond to stresses (Figure 2). Traditional research faced challenges in deciphering complex gene, protein, and metabolite networks, but AI-supported omics methods provide innovative solutions. These techniques rapidly identify crucial components in defense pathways, discover biomarkers, and reveal hidden patterns, enhancing our comprehension of plant defense processes. Integrating multi-omics data sources offers a holistic understanding, and as AI techniques evolve, they hold promise for developing stress-resistant crops, optimizing agricultural practices, and ensuring sustainable food production (Arabnia and Tran, 2011; Yan and Wang, 2023).

5.1 AI-assisted genomics

Machine learning (ML) algorithms play a pivotal role in identifying stress resistance genes, aiding breeders and researchers in enhancing crop production. Liang et al. (2011) utilized a variant of the Support Vector Machine (SVM) algorithm to identify key genes associated with drought resistance in A. thaliana. Shikha et al. (2017) demonstrated the superior performance of Bayes algorithms, identifying critical SNPs for drought resistance in maize. ML algorithms have been applied beyond drought resistance, with Wang et al. (2013) using an SVM-based model to predict salt resistance genes in rice. Ravari et al. (2015) assessed salt tolerance in Iranian wheat genotypes, identifying effective indices for predicting salt-tolerant varieties using artificial neural network (ANN) analysis. Schwarz et al. (2020) employed ML techniques to explore the cis-regulatory code governing the response to iron deficiency in Arabidopsis roots.

In the realm of plant disease resistance, SVM and its variants are widely employed, as demonstrated by Pal et al. (2016), achieving high accuracy in predicting disease resistance proteins. ML has also been instrumental in predicting pathogen effector proteins, with Sperschneider et al. (2016) developing EFFECTORP, the first ML classifier for fungal effectors. Despite the focus on disease resistance genes, ML algorithms hold promise in understanding genes susceptible to plant diseases, contributing significantly to agricultural practices (Yang and Guo, 2017). The application of ML in exploring plant single-cell genomic data offers opportunities to unravel cellular heterogeneity, decode regulatory networks, and identify novel cell types. Recent studies (Silva et al., 2019; Raimundo et al., 2021) highlight ML's potential in tasks such as generating low-dimensional representations, classifying cell types, inferring trajectories, deducing gene regulatory networks, and integrating multimodal data. Challenges related to low sequencing coverage and amplified artifacts in single-cell RNA (scRNA) sequencing are addressed by ML approaches, such as the SIMLR
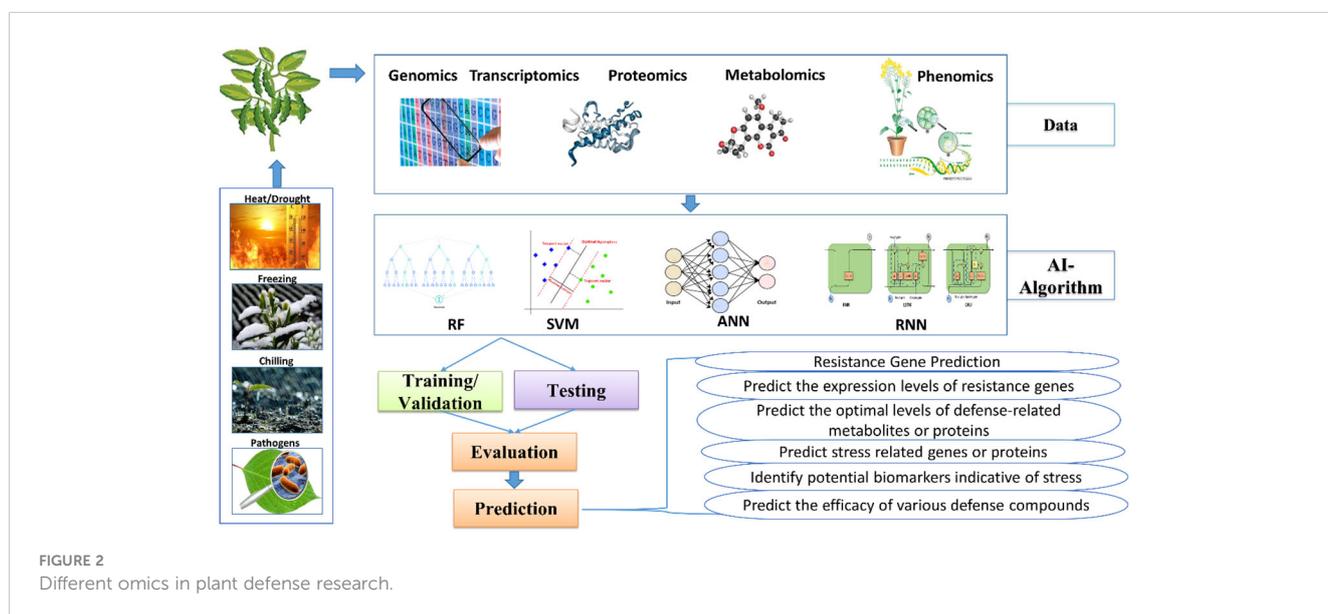

FIGURE 2
Different omics in plant defense research.

algorithm (Wang et al., 2017) and neural network models (Lin et al., 2017), providing more reliable insights into the intricate landscape of single-cell genomics. Despite these advancements, further research is needed to fully unlock the potential of ML in plant single-cell genomics. The current application of ML in identifying stress resistance genes is limited to a few plant species, urging the extension of ML utilization to other economically significant plants for a comprehensive understanding of stress resistance mechanisms and accelerated breeding efforts.

5.2 AI-assisted transcriptomics

In a comprehensive exploration of miRNAs and their roles in plant stress responses, Asefpour Vakilian (2020) conducted a study focusing on both biotic and abiotic stresses. The research employed feature selection algorithms to delve into the contributions of individual miRNAs in Arabidopsis thaliana responses to various abiotic stresses, including drought, salinity, cold, and heat. Utilizing information theory-based feature selection, key miRNAs, such as miRNA-169, miRNA-159, miRNA-396, and miRNA-393, were identified as significant contributors to the plant's reactions to these stressors. The study harnessed regression models, including DT, SVMs, and NB, revealing the exceptional predictive capabilities of SVM with a Gaussian kernel, achieving a high coefficient of determination ($R^2 = 0.96$) for plant stress based on miRNA concentrations.

Contrary to the traditional belief in separate signaling pathways for abiotic and biotic stresses in plants, a study on rice by Shaik and Ramakrishna (2014) shed light on the intricate relationship between these stress responses. Through a meta-analysis of microarray studies, the researchers identified shared stress-responsive genes in rice, revealing conserved expression patterns across both types of stresses for approximately 70% of the common differentially expressed genes. Advanced data analysis techniques and ML models, including recursive-support vector machine and random forests decision tree, effectively distinguished between abiotic and biotic stress responses based on gene expression profiles. The recursive-support vector machine achieved a perfect 100% accuracy in classifying these stress types, identifying 196 genes that significantly contributed to the accurate classification.

In a study led by Meng et al. (2021), supervised classification models were employed to identify genes responding transcriptionally to cold stress. Surprisingly, models trained solely with features derived from genome assemblies displayed modest reductions in performance compared to those incorporating a wider range of data. Notably, models trained with data from one plant species demonstrated remarkable success in predicting gene responses to cold stress in related species, even when transferring predictions between cold-sensitive and cold-tolerant species. Multi-species models, trained using data from multiple species, outperformed single-species models when it came to cross-species prediction accuracy. This approach, driven by ML, shows promise in accelerating the understanding of gene expression responses to environmental stresses across diverse plant species.

In response to abiotic stress, such as heat or cold, plants undergo significant changes in gene expression to adapt and survive. In a study conducted by Zhou et al. (2022), transcriptome profiling of maize genotypes exposed to heat or cold stress revealed extensive alterations in transcript abundance. Motifs near the transcription start sites (TSSs) of genes responsive to thermal stress were found to be enriched. Predictive models developed using these motifs could forecast gene expression responses to stress, with enhanced accuracy focusing on motifs within unmethylated regions near the TSSs. However, challenges emerged when applying these models across different maize genotypes, indicating reduced performance when transferred between genotypes.

In a recent study, Pradhan et al. (2023a) employed artificial intelligence, specifically ML, to tackle the challenge of identifying long non-coding RNAs (lncRNAs) associated with abiotic stress responses in plants. Abiotic stresses significantly impact crop yields, emphasizing the importance of developing stress-resistant crop cultivars. The researchers devised a novel computational model capable of predicting abiotic stress-responsive lncRNAs. They utilized a dataset comprising stress-responsive and non-stress-responsive lncRNA sequences for binary classification. Various ML algorithms, including SVM, were applied, and the representation of lncRNAs was numeric based on Kmer features. Through effective feature selection strategies, the SVM model demonstrated impressive cross-validation accuracy at 68.84%. Further validating its robustness, the model exhibited an accuracy of 76.23% on an independent test dataset. To enhance accessibility, the researchers also introduced an online prediction tool called ASLncR.

In a parallel study, Pradhan et al. (2023a) directed their focus towards predicting microRNAs (miRNAs) associated with specific abiotic stresses, such as cold, drought, heat, and salt. Given the vital role of miRNAs in plant responses to these stresses, their identification holds significance for breeding stress-resistant crops. Leveraging ML, specifically SVM, the researchers developed a computational model for predicting stress-responsive miRNAs. They utilized pseudo-K-tuple nucleotide compositional features to numerically represent miRNAs. The SVM model achieved high cross-validation prediction accuracies ranging from 87.71% to 90.15% across different stress conditions. To facilitate the utility of this computational tool, an online prediction server named ASmiR was established.

Similarly, Meher et al. (2022a) contributed to the field by developing a ML-based method for predicting miRNAs responsive to abiotic stresses. They worked with three types of datasets: miRNA, pre-miRNA, and pre-miRNA + miRNA. Using pseudo-K-tuple nucleotide compositional features, sequence data was transformed into numeric feature vectors. SVM was employed for prediction, and the model achieved respectable results. The area under the receiver operating characteristics curve (auROC) and area under the precision-recall curve (auPRC) percentages ranged from 65.64% to 77.94%. Overall prediction accuracies for the independent test set ranged from 62.33% to 69.21%. To facilitate the application of this approach, the researchers provided an online prediction server named ASRmiRNA. The method shows promise in advancing the identification of abiotic stress-responsive pre-miRNAs and miRNAs.

5.3 AI-assisted proteomics

Meher et al. (2022a) employed computational methods and machine learning (ML) to streamline the identification of abiotic stress-responsive genes (SRGs) across various stress conditions, achieving accuracy levels of 60% to 78% with the SVM model. They introduced an online prediction application, ASRpro, for broader accessibility. In the realm of plant-pathogen protein-protein interactions (PPIs), Yang et al. (2019) utilized Random Forest to predict known plant-pathogen PPIs, showcasing enhanced accuracy by incorporating sequence data and network attributes. The InterSPPI web server was introduced to support ongoing research. Karan et al. (2023) focused on plant-microbe interactions, predicting PPIs for rice and blast fungus interactions with ML models achieving up to 95% accuracy on experimental datasets. The specificity of the model to *O. sativa* and *M. grisea* was confirmed through assessments against other pathogen-host datasets. Ahmed et al. (2023) introduced a novel activation function, Gaussian Error Linear Unit with Sigmoid (SIELU), in a deep learning model for classifying unknown abiotic stress protein sequences, outperforming other models with high accuracies ranging from 80.78% to 95.11%.

5.4 AI-assisted metabolomics

Liu et al. (2017) conducted a study focusing on the classification of 216 plants based on their incomplete metabolite content. Their research employed a network clustering algorithm to group metabolites with similar structures. Plants were represented as binary vectors, and hierarchical clustering was used for classification. Despite working with incomplete data, the approach successfully clustered plants in accordance with known evolutionary relationships, underscoring the significance of metabolite content as a taxonomic marker. Furthermore, the study discussed how metabolite content could serve as a predictor for nutritional and medicinal properties in plants, revealing previously unknown species-metabolite relationships.

In Fürtauer et al.'s study (2018), the emphasis was on understanding how abiotic stress influences the metabolic regulation of plants. The researchers utilized Arabidopsis wild-type plants and mutant lines with deficiencies in sucrose or starch metabolism, subjecting them to cold and high-light stress conditions. Through quantifying changes in the primary metabolome and proteome, they trained a machine-learning algorithm to classify mutant lines under control and stress conditions. This innovative approach identified a core module consisting of 23 proteins that reliably predicted combined temperature and high-light stress conditions. Importantly, 18 of these proteins were associated with protein-protein interactions, providing insights into the intricate biochemical regulation occurring in response to changing environmental conditions.

5.5 AI-assisted phenomics

DL techniques have proven remarkably effective in the realm of crop management, spanning various crops like rice, wheat, tomato, and potato. Li et al. (2020) pioneered a DL-based video detection system aimed at addressing plant diseases and pests in crops. Their primary goal was the swift identification of plant diseases and pests through comprehensive video analysis, utilizing advanced models such as Faster R-CNN and YOLO v3 for real-time video detection systems. The approach involved transforming videos into individual frames, analyzed using a Faster R-CNN framework for detection. Video-based evaluation metrics were introduced to assess detection quality, demonstrating that their custom backbone system outperformed existing systems in detecting untrained rice videos. Focusing on wheat stripe rust, a prevalent disease affecting wheat yields, Mi et al. (2020) introduced a novel DL network called C-DenseNet. This network incorporated the Convolutional Block Attention Module (CBAM) into a densely connected convolutional network (DenseNet), surpassing classical DenseNet and ResNet models in wheat stripe rust severity grading with a test accuracy of 97.99%.

Wang and Liu (2021) presented an early recognition method for tomato leaf spot using the MobileNetv2-YOLOv3 model. They enhanced recognition accuracy by introducing the GIoU bounding box regression loss function. This lightweight model demonstrated significant improvements in recognition performance compared to other models, achieving an F1 score of 94.13% under specific conditions. Addressing virus diseases in seed potatoes, Polder et al. (2019) proposed a hyperspectral imaging approach for field detection. They designed an imaging setup with a hyperspectral line-scan camera, training a convolutional neural network (CNN) on field data. The method achieved high precision and recall, showcasing its potential for real-world disease detection in potato crops. Chen and Yuan (2019) developed a deep-learning pipeline for localizing and counting agricultural pests in images. Their method integrated a convolutional neural network (CNN) and a region proposal network (RPN) with Non-Maximum Suppression (NMS) to remove overlapping detections. The model demonstrated high precision (0.93) with a low miss rate (0.10), showcasing its effectiveness in pest detection.

Feng et al. (2020) addressed plant defense against salinity stress using image processing and DL algorithms. They utilized high-throughput plant phenotyping technologies for non-destructive monitoring of plant traits. Employing hyperspectral imaging (HSI), the researchers assessed the phenotypes of 13 okra genotypes following salt treatment. Advanced plant and leaf segmentation techniques, coupled with DL algorithms, achieved outstanding results in accurately delineating plant and leaf structures. Salinity stress was found to have deleterious effects on okra's physiological and biochemical processes, leading to significant alterations in spectral information. Leveraging this data, the study constructed predictive models for various traits, yielding promising results with correlation coefficients ranging from 0.588 to 0.835.

An overview of dedicated ML-based tools designed for addressing both abiotic and biotic stresses in plants are enlisted in Table 2. These specialized tools cater to specific types of stressors, offering a comprehensive resource for researchers and practitioners in the field of plant defense.

Supervised ML and DL have been extensively applied in various plant biology studies, but there are situations where unsupervised and semi-supervised approaches are crucial. In the field of plant systems biology, unsupervised and semi-supervised learning algorithms play essential roles in diverse areas such as data clustering, dimensionality reduction (DR), visualization, gene regulatory network inference, cross-species prediction, and single-cell omics data analysis (Rai et al., 2019). PCA is widely used for DR and visualization of genotypic and multi-omics data (Yan et al., 2020). Hierarchical clustering is extensively employed for clustering genes with similar expression patterns in transcriptomic and proteomic research (Xu et al., 2012; Klepikova et al., 2016). Algorithms like t-sne and optics contribute significantly to the analysis of genotypic data, enhancing visualization of large-scale maize hybrid populations' structures (Yan et al., 2021). Non-negative matrix factorization (NMF) proves valuable in breaking down expression matrices with thousands of genes into a small number of metagenes in Arabidopsis and maize (Wilson et al., 2012; Ma et al., 2022). The multifactor dimensionality reduction (MDR) algorithm is employed for identifying multiple pairwise epistatic effects and gene–environment interactions affecting agronomic and quality traits in rice and barley (Xu et al., 2015; Xu et al., 2018).

Semi-supervised and transfer learning strategies have emerged to overcome the scarcity of annotated genes and pathways in plants. Transfer learning was employed to predict specialized/general metabolism-related genes in *Solanum lycopersicum* (tomato) by leveraging well-annotated Arabidopsis genes (Moore et al., 2020). Another innovative approach, 'evolutionarily informed machine learning,' used an xgboost model trained on transcriptomic data in Arabidopsis to predict nitrogen-use efficiency (NUE) and related genes in maize (Cheng et al., 2021). With the advent of single-cell sequencing technology, challenges arise due to higher dimensionality and complexity (Bobrovskikh et al., 2021). Advanced algorithms like t-sne, umap, magic, phate, Saucie, and Beeline have been proposed to address these challenges and are prevalent in both human and plant studies (Van Dijk et al., 2018; Amodio et al., 2019; Becht et al., 2019; Pratapa et al., 2020; Wu and Zhang, 2020; Marand et al., 2021).

6 Challenges

The rapid advancements in biological data generation and ML development have opened up significant possibilities for unraveling complex biological information. However, integrating ML into plant molecular studies poses notable challenges. ML approaches, similar to traditional plant molecular methods, are highly context-specific, underscoring the importance of meticulous experimental design. It's essential to recognize that while ML aims to create predictive models, each ML algorithm comes with distinct strengths and weaknesses, influencing predictive efficiency under specific conditions. Consequently, an ML model crafted for one dataset may struggle to generalize well to others due to inherent biological and technical variations.

The abundance of omics datasets provides a treasure trove of information. However, a notable portion of these datasets is marked by characteristics such as noise and sparsity. This poses a substantial challenge when it comes to accurately identifying biological features, especially during the integration of various omics data sources (Joyce and Palsson, 2006). The challenge of imbalanced datasets, where sample sizes vary across categories, is pervasive in ML. Researchers address this through resampling strategies such as oversampling and undersampling (Maimon and Rokach, 2005). For instance, Li et al. (2021) employed the synthetic minority oversampling technique (SMOTE) to bolster the representation of minority cases, a crucial step in the identification of effector proteins. The presence of noise and sparsity in these datasets introduces uncertainty and complexity, potentially hindering the identification of meaningful patterns or features within the biological data. Researchers must grapple with the task of distinguishing genuine biological signals from the background noise, emphasizing the need for robust analytical approaches to ensure the reliability of findings. Additionally, the issue of overfitting looms prominently, particularly in the domain of DL. Overfitting occurs when a model becomes overly tailored to the intricacies of a specific dataset to the extent that it struggles to generalize well to new, unseen data. This phenomenon can compromise the model's predictive capabilities and hinder its applicability to real-world scenarios. In addressing this concern, techniques like dropout have been employed (Scholz et al., 2004).

Dropout is a regularization technique that involves randomly "dropping out" or deactivating a subset of neurons during the training of a neural network. By doing so, dropout helps prevent the neural network from becoming overly reliant on specific features or relationships present in the training data, thereby enhancing its ability to generalize to new and unseen data. This technique acts as a safeguard against overfitting, promoting a more robust and adaptable model. Various factors, including data preprocessing, user-defined parameters, and domain knowledge, significantly influence the effectiveness of ML models. ML practitioners play a pivotal role in decision-making throughout the process, underscoring the importance of incorporating prior knowledge and domain expertise to unveil meaningful patterns.

Dealing with big data characteristics in plant system biology studies, encompassing volume, variety, veracity, value, and velocity, presents its own set of challenges. ML methods must adapt to handle multi-omics data, considering the unique insights each omics layer provides. Challenges include addressing high-dimensional data issues such as sparsity, multicollinearity, and overfitting, necessitating tailored methods and collaborative efforts in data integration.

Interpreting complex models, particularly in advanced ML approaches like DL, remains challenging due to their 'black box' nature. Researchers often prioritize understanding the biological significance of a predictive model over its accuracy, requiring careful processing and correlation with existing biological knowledge.

TABLE 2 AI-based tools for plant defense against abiotic and biotic stress.

Tool	Stressors	Features	Algorithm	Description	Website (Accessed on November 27, 2023)
ASRpro (Meher et al., 2022b)	Cold, drought, heat, light, oxidative and salt	autocross covariance & and K-mer composition	SVM	Identification of proteins associated with multiple abiotic stress in plants	https://iasri-sg.icar.gov.in/asrpro/
AsmiR (Pradhan et al., 2023b)	Drought, cold, salt and heat	pseudo-K-tuple nucleotide composition	SVM	Prediction of abiotic stress–specific miRNAs in plants	https://iasri-sg.icar.gov.in/asmir/
PLncPRO (Singh et al., 2017)	Abiotic	71 features were extracted using Framefinder (Slater, 2000) and BlastX (Altschul et al., 1997)	RF	Prediction of long abiotic stress-responsive long non-coding RNAs	http://ccbb.jnu.ac.in/plncpro/
ASRmiRNA (Meher et al., 2022a)	Abiotic	pseudo K-tuple nucleotide composition	SVM	Prediction of abiotic Stress-Responsive miRNA	http://cabgrid.res.in:8080/asrmirna
ASLncR (Pradhan et al., 2023a)	Abiotic	Kmer	SVM	Prediction of abiotic stress-responsive long non-coding RNAs	https://iasri-sg.icar.gov.in/aslncr/
DeepAProt (Ahmed et al., 2023)	Drought, cold, salinity and heat	46 features were extracted using bio-python package (Cock et al., 2009)	Long-short term memory	Identification and classification of abiotic stress protein sequences in cereals	http://login1.cabgrid.res.in:5500/
PredHSP (Kumar et al., 2016)	Heat	Dipeptide composition	SVM	Sequence-based prediction and classification of heat shock protein	http://14.139.227.92/mkumar/predhsp/index.html
iHSP-PseRAAC (Feng et al., 2013)	Heat	Spaced dipeptide composition	SVM	Identification of the heat shock protein families	http://lin.uestc.edu.cn/server/iHSP-PseRAAAC
ir-HSP (Meher et al., 2018)	Heat	pseudo amino acid composition	SVM	Classification of heat shock proteins sequences into one of the six heat shock proteins families	http://cabgrid.res.in:8080/ir-hsp
DeeperHSP (Min et al., 2021)	Heat	CNN	CNN	Identification of heat shock proteins	https://github.com/mswzeus/DeeperHSP
afpCOOL (Eslami et al., 2018)	Cold	Amino acid composition; evolutionary features	SVM	Prediction of anti-freeze proteins	–
AFP-Pred (Kandaswamy et al., 2011)	Cold	Overall composition of helix, stand and coil; Physicochemical properties, etc.	RF	Predicts antifreeze proteins from sequence-derived properties	https://www3.ntu.edu.sg/home/EPNSugan/index_files/AFP-Pred.htm
DRPPP (Pal et al., 2016)	Biotic	Protein sequence	SVM	Prediction of disease resistance proteins in plants	http://14.139.240.55/NGS/download.php
prPred (Wang et al., 2022)	Biotic	bidirectional long short-term memory	light gradient boosting	Prediction of plant resistance proteins	http://lab.malab.cn/soft/prPred-DRLF/
StackRPred (Chen et al., 2022)	Biotic	pairwise energy content of residues	Ensemble learning	Prediction of plant resistance proteins	–
ResCap (Kushwaha et al., 2021)	Biotic	Sequence compositional properties	SVM	Prediction of plant resistance gene	http://rescap.ltj.slu.se/ResCap/
EffectorP 3.0 (Sperschneider and Dodds, 2022)	Pathogen	Protein sequence features	Ensemble learning	Prediction of fungal and oomycetes effector proteins	https://effectorp.csiro.au/
InterSPPI (Yang et al., 2019)	Pathogens	Protein sequence features; network encoding	RF	Prediction of PPI between Arabidopsis and pathogens	http://systbio.cau.edu.cn/intersppi/index.php

Despite these challenges, the studies discussed in this context represent success stories in the application of AI in plant omics. To fully harness the potential of AI, robust and scalable algorithms that uncover meaningful biological insights are crucial. Ensuring accuracy and reliability through experimental validation is essential for translating computational findings into practical applications.

Bridging collaboration gaps between omics researchers, data scientists, and agricultural experts is vital for realizing the full potential of AI in plant defense and practical applications.

7 Future directions

Integration of AI and Omics Data in Early Disease Detection: One promising future direction is the seamless integration of AI with various omics data (genomics, transcriptomics, proteomics, and metabolomics) to enable the early detection of plant diseases. Advanced AI algorithms can analyze multi-dimensional omics datasets to identify subtle changes in plant molecular profiles associated with disease onset, even before visible symptoms emerge.

High-Throughput Phenotyping with AI-Omics Fusion: Combining AI-assisted omics techniques with high-throughput phenotyping methods offers an exciting avenue. This fusion can enable real-time monitoring of plant health by linking molecular responses to observable phenotypic traits. The integration of phenomics data into AI-driven analysis pipelines enhances our understanding of plant defense mechanisms.

Predictive Modeling of Disease Dynamics: Leveraging AI and omics data, predictive modeling can be developed to forecast disease dynamics in plant populations. ML and DL models can factor in genetic, molecular, and environmental variables to predict disease outbreaks and assess the impact of preventive measures.

Customized Crop Breeding: Future research may focus on using AI-assisted omics techniques to tailor crop breeding programs for enhanced disease resistance. By pinpointing specific genetic markers and pathways associated with resistance, breeders can design crops with improved defense mechanisms.

AI-Guided Sustainable Disease Management: AI can assist in optimizing disease management strategies. Integrating AI-powered recommendations with omics data allows for precision application of pesticides, reducing environmental impact and lowering costs while effectively controlling plant pathogens.

Addressing Combined Stressors: With climate change and evolving agricultural practices, plants often face the challenge of multiple stressors simultaneously. Future research should focus on understanding how plants respond to combined biotic and abiotic stress, as this represents a significant real-world scenario. AI-assisted omics techniques can play a pivotal role in unraveling the intricate interactions between different stress factors and their cumulative effects on plant defense mechanisms. This knowledge is essential for developing holistic and resilient strategies to protect crops in complex stress environments, ensuring sustainable food production.

Integration of Remote Sensing Data with AI-Omics Fusion: The integration of remote sensing technology with AI-assisted omics techniques offers a powerful approach to monitor and mitigate plant stress. Remote sensing provides valuable spatial and temporal data on plant health, stress factors, and environmental conditions. By merging remote sensing data with omics information, researchers can gain a comprehensive understanding of the interplay between genetic responses and environmental stressors. This integrated approach enables more precise and timely interventions to enhance plant defense and reduce crop losses.

8 Conclusion

In conclusion, this comprehensive review explores the landscape of omics studies in plant defense against biotic and abiotic stress, and the transformative role of ML techniques in various omics domains. From genomics to metabolomics, AI-assisted techniques showcase their prowess in extracting meaningful insights from expansive datasets, surpassing traditional methods. While emphasizing the advantages of ML, the review also addresses the challenges associated with its implementation in plant omics, paving the way for future developments. Progress in computational frameworks facilitates the seamless application of modern methods. With the increasing volume of plant sequencing data, ML emerges as a catalyst in accelerating various facets of plant genomic research. This includes pinpointing genes associated with resistance against biotic and abiotic stress, as well as enhancing our comprehension of gene regulation mechanisms. These strides are poised to aid agricultural researchers in enhancing crop yield and quality, fostering improved resilience to biotic and abiotic stressors.

Author contributions

SM: Writing – original draft. DS: Writing – original draft. HC: Writing – original draft. SS: Writing – review & editing. RD: Writing – review & editing. GJ: Writing – review & editing. SA: Funding acquisition, Writing – review & editing.

Funding

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abele, L., Legat, C., Grimm, S., and Müller, A. W. (2013). "Ontology-based validation of plant models," in *2013 11th IEEE International Conference on Industrial Informatics (INDIN)*, Bochum, Germany. 236–241. doi: 10.1109/INDIN.2013.6622888

Ahmed, B., Haque, M. A., Iquebal, M. A., Jaiswal, S., Angadi, U. B., Kumar, D., et al. (2023). DeepAProt: Deep learning based abiotic stress protein sequence classification and identification tool in cereals. *Front. Plant Sci.* 13, 1008756. doi: 10.3389/fpls.2022.1008756

Aizat, W. M., and Hassan, M. (2018). Proteomics in systems biology. *Omics applications for systems biology*, 31–49.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Statistician* 46 (3), 175–185. doi: 10.1080/00031305.1992.10475879

Altman, N., and Krzywinski, M. (2018). The curse (s) of dimensionality. *Nat. Methods* 15 (6), 399–400. doi: 10.1038/s41592-018-0019-x

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25 (17), 3389–3402. doi: 10.1093/nar/25.17.3389

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. *Nat. Methods* 16 (11), 1139–1145. doi: 10.1038/s41592-019-0576-7

Anderson, P. K., Cunningham, A. A., Patel, N. G., Morales, F. J., Epstein, P. R., and Daszak, P. (2004). Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends Ecol. Evol.* 19 (10), 535–544. doi: 10.1016/j.tree.2004.07.021

Arabnia, H., and Tran, Q. N. (2011). *Software tools and algorithms for biological systems* (New York, NY: Springer).

Arsenovic, M., Karanovic, M., Sladojevic, S., Anderla, A., and Stefanovic, D. (2019). Solving current limitations of deep learning based approaches for plant disease detection. *Symmetry* 11 (7), 939.

Asefpour Vakilian, K. (2020). Machine learning improves our knowledge about miRNA functions towards plant abiotic stresses. *Sci. Rep.* 10 (1), 3041. doi: 10.1038/s41598-020-59981-6

Avelino, J., Cristancho, M., Georgiou, S., Imbach, P., Aguilar, L., Bornemann, G., et al. (2015). The coffee rust crises in Colombia and Central America, (2008–2013): impacts, plausible causes and proposed solutions. *Food Secur.* 7, 303–321. doi: 10.1007/s12571-015-0446-9

Banf, M., and Rhee, S. Y. (2017). Computational inference of gene regulatory networks: approaches, limitations and opportunities. *Biochim. Biophys. Acta (BBA)- Gene Regul. Mech.* 1860 (1), 41–52. doi: 10.1016/j.bbagrm.2016.09.003

Bebber, D. P., Ramotowski, M. A., and Gurr, S. J. (2013). Crop pests and pathogens move polewards in a warming world. *Nat. Climate Change* 3 (11), 985–988. doi: 10.1038/nclimate1990

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. *Nat. Biotechnol.* 37 (1), 38–44. doi: 10.1038/nbt.4314

Bhardwaj, A. R., Joshi, G., Kukreja, B., Malik, V., Arora, P., Pandey, R., et al. (2015). Global insights into high temperature and drought stress regulated genes by RNA-Seq in economically important oilseed crop Brassica juncea. *BMC Plant Biol.* 15 (1), 1–15.

Bobrovskikh, A., Doroshkov, A., Mazzoleni, S., Cartenì, F., Giannino, F., and Zubairova, U. (2021). A sight on single-cell transcriptomics in plants through the prism of cell-based computational modeling approaches: benefits and challenges for data analysis. *Front. Genet.* 12, 652974. doi: 10.3389/fgene.2021.652974

Bonetta, R., and Valentino, G. (2020). Machine learning techniques for protein function prediction. *Proteins: Structure Function Bioinf.* 88 (3), 397–413. doi: 10.1002/prot.25832

Borah, P., Teja, A., Jha, S. A., and Bhattacharyya, D. K. (2020). "TUKNN: a parallel KNN algorithm to handle large data," in *Big Data, Machine Learning, and Applications: First International Conference, BigDML 2019, Silchar, India*, December 16–19, 2019, Revised Selected Papers 1. (Springer International Publishing), 1–13.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Bubb, K. L., and Deal, R. B. (2020). Considerations in the analysis of plant chromatin accessibility data. *Curr. Opin. Plant Biol.* 54, 69–78. doi: 10.1016/j.pbi.2020.01.003

Bylesjö, M., Eriksson, D., Kusano, M., Moritz, T., and Trygg, J. (2007). Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. *Plant J.* 52 (6), 1181–1191. doi: 10.1111/j.1365-313X.2007.03293.x

Callinan, P. A., and Feinberg, A. P. (2006). The emerging science of epigenomics. *Hum. Mol. Genet.* 15 (suppl_1), R95–R101.

Chen, L., and Yuan, Y. (2019). "Agricultural disease image dataset for disease identification based on machine learning," in *Big Scientific Data Management: First International Conference, BigSDM 2018*, Beijing, China, November 30–December 1, 2018, Revised Selected Papers 1. (Springer International Publishing), 263–274.

Chen, Y., Li, Z., and Li, Z. (2022). Prediction of plant resistance proteins based on pairwise energy content and stacking framework. *Front. Plant Sci.* 13, 912599. doi: 10.3389/fpls.2022.912599

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. *Nat. Commun.* 12 (1), 5627. doi: 10.1038/s41467-021-25893-w

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25 (11), 1422. doi: 10.1093/bioinformatics/btp163

de Luis Balaguer, M. A., Fisher, A. P., Clark, N. M., Fernandez-Espinosa, M. G., Möller, B. K., Weijers, D., et al. (2017). Predicting gene regulatory networks by combining spatial and temporal gene expression data in Arabidopsis root stem cells. *Proc. Natl. Acad. Sci.* 114 (36), E7632–E7640. doi: 10.1073/pnas.1707566114

Deng, W., Zhang, K., Liu, S., Zhao, P. X., Xu, S., and Wei, H. (2018). JRmGRN: joint reconstruction of multiple gene regulatory networks with common hub genes using data from multiple tissues or conditions. *Bioinformatics* 34 (20), 3470–3478. doi: 10.1093/bioinformatics/bty354

Dietterich, T. G. (2000). Multiple classifier systems. *Lecture Notes Comput. Sci.* 1857, 1–15. doi: 10.1007/3-540-45014-9_1

Ding, J., Hagood, J. S., Ambalavanan, N., Kaminski, N., and Bar-Joseph, Z. (2018). iDREM: Interactive visualization of dynamic regulatory networks. *PloS Comput. Biol.* 14 (3), e1006019. doi: 10.1371/journal.pcbi.1006019

Du, Q., Campbell, M., Yu, H., Liu, K., Walia, H., Zhang, Q., et al. (2019). Network-based feature selection reveals substructures of gene modules responding to salt stress in rice. *Plant Direct* 3 (8), e00154. doi: 10.1002/pld3.154

Dugas, D. V., Monaco, M. K., Olson, A., Klein, R. R., Kumari, S., Ware, D., et al. (2011). Functional annotation of the transcriptome of Sorghum bicolor in response to osmotic stress and abscisic acid. *BMC Genomics* 12, 1–21.

Eslami, M., Zade, R. S. H., Takalloo, Z., Mahdevar, G., Emamjomeh, A., Sajedi, R. H., et al. (2018). afpCOOL: A tool for antifreeze protein prediction. *Heliyon* 4 (7). doi: 10.1016/j.heliyon.2018.e00705

Farooq, A., Mir, R. A., Sharma, V., Pakhtoon, M. M., Bhat, K. A., Shah, A. A., et al. (2022). "Crop proteomics: Towards systemic analysis of abiotic stress responses," in *Advancements in developing abiotic stress-resilient plants*. (CRC Press), 265–285.

Feng, P. M., Chen, W., Lin, H., and Chou, K. C. (2013). iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Analytical Biochem.* 442 (1), 118–125. doi: 10.1016/j.ab.2013.05.024

Feng, X., Zhan, Y., Wang, Q., Yang, X., Yu, C., Wang, H., et al. (2020). Hyperspectral imaging combined with machine learning as a tool to obtain high-throughput plant salt-stress phenotyping. *Plant J.* 101 (6), 1448–1461. doi: 10.1111/tpj.14597

Fukushima, A., Kusano, M., Redestig, H., Arita, M., and Saito, K. (2009). Integrated omics approaches in plant systems biology. *Curr. Opin. Chem. Biol.* 13 (5-6), 532–538. doi: 10.1016/j.cbpa.2009.09.022

Fürtauer, L., Pschenitschnigg, A., Scharkosi, H., Weckwerth, W., and Nägele, T. (2018). Combined multivariate analysis and machine learning reveals a predictive module of metabolic stress response in Arabidopsis thaliana. *Mol. Omics* 14 (6), 437–449. doi: 10.1039/C8MO00095F

Ghosal, S., Blystone, D., Singh, A. K., Ganapathysubramanian, B., Singh, A., and Sarkar, S. (2017). Interpretable deep learning applied to plant stress phenotyping. *arXiv preprint arXiv:1710.08619*. doi: 10.48550/arXiv.1710.08619

Gilpin, W., Huang, Y., and Forger, D. B. (2020). Learning dynamics from large biological data sets: machine learning meets systems biology. *Curr. Opin. Syst. Biol.* 22, 1–7. doi: 10.1016/j.coisb.2020.07.009

Gokalp, O., and Tasci, E. (2019). "Weighted voting based ensemble classification with hyper-parameter optimization," in *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*. 1–4 (IEEE).

Guyon, I., and Elisseeff, A. (2006). "An introduction to feature extraction," in *Feature extraction: foundations and applications* (Berlin, Heidelberg: Springer), 1–25.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A. (2008). *Feature extraction: foundations and applications* Vol. 207 (Springer).

Haque, S., Ahmad, J. S., Clark, N. M., Williams, C. M., and Sozzani, R. (2019). Computational prediction of gene regulatory networks in plant growth and development. *Curr. Opin. Plant Biol.* 47, 96–105. doi: 10.1016/j.pbi.2018.10.005

Hills, R. G., and Leslie, I. H. (2003). *Statistical validation of engineering and scientific models: validation experiments to application (No. SAND2003-0706)*, Sandia National Lab. (SNL-NM), Albuquerque, NM (United States); Sandia National Lab. (SNL-CA), Livermore, CA (United States).

Hou, J. (2017). *New approaches of protein function prediction from protein interaction networks* (Academic Press).

Isewon, I., Apata, O., Oluwamuyiwa, F., Aromolaran, O., and Oyelade, J. (2022). Machine learning algorithms: their applications in plant omics and agronomic traits' improvement. *F1000Research* 11, 1256. doi: 10.12688/f1000research.125425.1

Jiao, Y., and Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biol.* 4, 320–330. doi: 10.1007/s40484-016-0081-2

Jin, H., Huang, F., Cheng, H., Song, H., and Yu, D. (2013). Overexpression of the GmNAC2 gene, an NAC transcription factor, reduces abiotic stress tolerance in tobacco. *Plant Mol. Biol. Rep.* 31, 435–442.

John, M., Haselbeck, F., Dass, R., Malisi, C., Ricca, P., Dreischer, C., et al. (2022). A comparison of classical and machine learning-based phenotype prediction methods on simulated data and three plant species. *Front. Plant Sci.* 13, 932512. doi: 10.3389/fpls.2022.932512

Johnson, J. M., Alex, T., and Oelmüller, R. (2014). Piriformospora indica: The versatile and multifunctional root endophytic fungus for enhanced yield and tolerance to biotic and abiotic stress in crop plants.

Joyce, A. R., and Palsson, B.Ø. (2006). The model organism as a system: integrating'omics' data sets. *Nat. Rev. Mol. Cell Biol.* 7 (3), 198–210. doi: 10.1038/nrm1857

Kandaswamy, K. K., Chou, K. C., Martinetz, T., Möller, S., Suganthan, P. N., Sridharan, S., et al. (2011). AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *J. Theor. Biol.* 270 (1), 56–62. doi: 10.1016/j.jtbi.2010.10.037

Karan, B., Mahapatra, S., Sahu, S. S., Pandey, D. M., and Chakravarty, S. (2023). Computational models for prediction of protein–protein interaction in rice and Magnaporthe grisea. *Front. Plant Sci.* 13, 1046209. doi: 10.3389/fpls.2022.1046209

Khalid, S., Khalil, T., and Nasreen, S. (2014). "A survey of feature selection and feature extraction techniques in machine learning," in *2014 science and information conference*. 372–378 (IEEE).

Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D., and Penin, A. A. (2016). A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant J.* 88 (6), 1058–1070. doi: 10.1111/tpj.13312

Ko, D. K., and Brandizzi, F. (2020). Network-based approaches for understanding gene regulation and function in plants. *Plant J.* 104 (2), 302–317. doi: 10.1111/tpj.14940

Kubicek, A., Jopp, F., Breckling, B., Lange, C., and Reuter, H. (2015). Context-oriented model validation of individual-based models in ecology: A hierarchically structured approach to validate qualitative, compositional and quantitative characteristics. *Ecol. Complexity* 22, 178–191. doi: 10.1016/j.ecocom.2015.03.005

Kumar, V., and Jain, M. (2015). The CRISPR–Cas system for plant genome editing: advances and opportunities. *J. Exp. Bot.* 66 (1), 47–57. doi: 10.1093/jxb/eru429

Kumar, R., Kumari, B., and Kumar, M. (2016). PredHSP: sequence based proteome-wide heat shock protein prediction and classification tool to unlock the stress biology. *PloS One* 11 (5), e0155872. doi: 10.1371/journal.pone.0155872

Kushwaha, S. K., Åhman, I., and Bengtsson, T. (2021). ResCap: plant resistance gene prediction and probe generation pipeline for resistance gene sequence capture. *Bioinf. Adv.* 1 (1), vbab033. doi: 10.1093/bioadv/vbab033

Lasky, J. R., Upadhyaya, H. D., Ramu, P., Deshpande, S., Hash, C. T., Bonnette, J., et al. (2015). Genome-environment associations in sorghum landraces predict adaptive traits. *Sci. Adv.* 1 (6), e1400218. doi: 10.1126/sciadv.1400218

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86 (11), 2278–2324. doi: 10.1109/5.726791

Leye, S., Himmelspach, J., and Uhrmacher, A. M. (2009). "A discussion on experimental model validation," in *2009 11th International Conference on Computer Modelling and Simulation*. 161–167 (IEEE).

Li, J., Wei, L., Guo, F., and Zou, Q. (2021). EP3: an ensemble predictor that accurately identifies type III secreted effectors. *Briefings Bioinf.* 22 (2), 1918–1928. doi: 10.1093/bib/bbaa008

Li, P., Cao, W., Fang, H., Xu, S., Yin, S., Zhang, Y., et al. (2017). Transcriptomic profiling of the maize (Zea mays L.) leaf response to abiotic stresses at the seedling stage. *Front. Plant Sci.* 8, 290.

Li, L., Snyder, J. C., Pelaschier, I. M., Huang, J., Niranjan, U. N., Duncan, P., et al. (2016). Understanding machine-learned density functionals. *Int. J. Quantum Chem.* 116 (11), 819–833.

Li, D., Wang, R., Xie, C., Liu, L., Zhang, J., Li, R., et al. (2020). A recognition method for rice plant diseases and pests video detection based on deep convolutional neural network. *Sensors* 20 (3), 578.

Li, Y. F., Wu, Y., Hernandez-Espinosa, N., and Peña, R. J. (2013). Heat and drought stress on durum wheat: Responses of genotypes, yield, and quality parameters. *J. Cereal Sci.* 57 (3), 398–404.

Liang, Y., Zhang, F., Wang, J., Joshi, T., Wang, Y., and Xu, D. (2011). Prediction of drought-resistant genes in Arabidopsis thaliana using SVM-RFE. *PloS One* 6 (7), e21750. doi: 10.1371/journal.pone.0021750

Lin, C., Jain, S., Kim, H., and Bar-Joseph, Z. (2017). Using neural networks for reducing the dimensions of single-cell RNA-Seq data. *Nucleic Acids Res.* 45 (17), e156–e156. doi: 10.1093/nar/gkx585

Liu, K., Abdullah, A. A., Huang, M., Nishioka, T., Altaf-Ul-Amin, M., and Kanaya, S. (2017). Novel approach to classify plants based on metabolite-content similarity. *BioMed. Res. Int.* 2017. doi: 10.1155/2017/5296729

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., and Pan, T. (2015). N 6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature* 518 (7540), 560–564.

Ma, W., Chen, S., Qi, Y., Song, M., Zhai, J., Zhang, T., et al. (2022). easyMF: A web platform for matrix factorization-based gene discovery from large-scale transcriptome data. *Interdiscip. Sciences: Comput. Life Sci.* 14 (3), 746–758. doi: 10.1007/s12539-022-00522-2

Maimon, O., and Rokach, L. (2005). *Data Mining and Knowledge Discovery Handbook* Vol. 2 (New York: Springer).

Mangin, B., Casadebaig, P., Cadic, E., Blanchet, N., Boniface, M. C., Carrère, S., et al. (2017). Genetic control of plasticity of oil yield for combined abiotic stresses using a joint approach of crop modelling and genome-wide association. *Plant Cell Environ.* 40 (10), 2276–2291. doi: 10.1111/pce.12961

Marand, A. P., Chen, Z., Gallavotti, A., and Schmitz, R. J. (2021). A cis-regulatory atlas in maize at single-cell resolution. *Cell* 184 (11), 3041–3055. doi: 10.1016/j.cell.2021.04.014

Margaria, P., Abbà, S., and Palmano, S. (2013). Novel aspects of grapevine response to phytoplasma infection investigated by a proteomic and phospho-proteomic approach with data integration into functional networks. *BMC Genomics* 14 (1), 1–15.

Maron, J. L., Baer, K. C., and Angert, A. L. (2014). Disentangling the drivers of context-dependent plant–animal interactions. *J. Ecol.* 102 (6), 1485–1496. doi: 10.1111/1365-2745.12305

McCulloch, W. S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. biophysics* 5, 115–133. doi: 10.1007/BF02478259

Meher, P. K., Begam, S., Sahu, T. K., Gupta, A., Kumar, A., Kumar, U., et al. (2022a). ASRmiRNA: abiotic stress-responsive miRNA prediction in plants by using machine learning algorithms with pseudo K-tuple nucleotide compositional features. *Int. J. Mol. Sci.* 23 (3), 1612. doi: 10.3390/ijms23031612

Meher, P. K., Sahu, T. K., Gahoi, S., and Rao, A. R. (2018). ir-HSP: improved recognition of heat shock proteins, their families and sub-types based on g-spaced di-peptide features and support vector machine. *Front. Genet.* 8, 235. doi: 10.3389/fgene.2017.00235

Meher, P. K., Sahu, T. K., Gupta, A., Kumar, A., and Rustgi, S. (2022b). ASRpro: A machine-learning computational model for identifying proteins associated with multiple abiotic stress in plants. *Plant Genome*, e20259. doi: 10.1002/tpg2.20259

Meng, X., Liang, Z., Dai, X., Zhang, Y., Mahboub, S., Ngu, D. W., et al. (2021). Predicting transcriptional responses to cold stress across plant species. *Proc. Natl. Acad. Sci.* 118 (10), e2026330118. doi: 10.1073/pnas.2026330118

Mi, Z., Zhang, X., Su, J., Han, D., and Su, B. (2020). Wheat stripe rust grading by deep learning with attention mechanism and images from mobile devices. *Front. Plant Sci.* 11, 558126.

Millet, E. J., Welcker, C., Kruijer, W., Negro, S., Coupel-Ledru, A., Nicolas, S. D., et al. (2016). Genome-wide analysis of yield in Europe: allelic effects vary with drought and heat scenarios. *Plant Physiol.* 172 (2), 749–764. doi: 10.1104/pp.16.00621

Min, S., Kim, H., Lee, B., and Yoon, S. (2021). Protein transfer learning improves identification of heat shock protein families. *PloS One* 16 (5), e0251865. doi: 10.1371/journal.pone.0251865

Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., et al. (2019). GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* 35 (12), 2159–2161. doi: 10.1093/bioinformatics/bty916

Mohammadi, P. P., Moieni, A., and Komatsu, S. (2012). Comparative proteome analysis of drought-sensitive and drought-tolerant rapeseed roots and their hybrid F1 line under drought stress. *Amino Acids* 43, 2137–2152.

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. *In silico Plants* 2 (1), diaa005. doi: 10.1093/insilicoplants/diaa005

Mosa, K. A., Ismail, A., Helmy, M., Mosa, K. A., Ismail, A., and Helmy, M. (2017). Omics and system biology approaches in plant stress research. *Plant stress tolerance: An integrated omics approach*, 21–34.

Muthamilarasan, M., Singh, N. K., and Prasad, M. (2019). Multi-omics approaches for strategic improvement of stress tolerance in underutilized crop species: a climate change perspective. *Adv. Genet.* 103, 1–38.

Niazian, M., and Niedbała, G. (2020). Machine learning for plant breeding and biotechnology. *Agriculture* 10 (10), 436.

Noor, E., Cherkaoui, S., and Sauer, U. (2019). Biological insights through omics data integration. *Curr. Opin. Syst. Biol.* 15, 39–47.

Pal, T., Jaiswal, V., and Chauhan, R. S. (2016). DRPPP: A machine learning based tool for prediction of disease resistance proteins in plants. *Comput. Biol. Med.* 78, 42–48. doi: 10.1016/j.compbiomed.2016.09.008

Polder, G., Blok, P. M., De Villiers, H. A., van der Wolf, J. M., and Kamp, J. (2019). Potato virus Y detection in seed potatoes using deep learning on hyperspectral images. *Front. Plant Sci.* 10, 209.

Pradhan, U. K., Meher, P. K., Naha, S., Rao, A. R., and Gupta, A. (2023a). ASLncR: a novel computational tool for prediction of abiotic stress-responsive long non-coding RNAs in plants. *Funct. Integr. Genomics* 23 (2), 113. doi: 10.1007/s10142-023-01040-0

Pradhan, U. K., Meher, P. K., Naha, S., Rao, A. R., Kumar, U., Pal, S., et al. (2023b). ASmiR: a machine learning framework for prediction of abiotic stress–specific miRNAs in plants. *Funct. Integr. Genomics* 23 (2), 92. doi: 10.1007/s10142-023-01014-2

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* 17 (2), 147–154. doi: 10.1038/s41592-019-0690-6

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.* 1, 81–106. doi: 10.1007/BF00116251

Rai, A., Yamazaki, M., and Saito, K. (2019). A new era in plant functional genomics. *Curr. Opin. Syst. Biol.* 15, 58–67. doi: 10.1016/j.coisb.2019.03.005

Raimundo, F., Meng-Papaxanthos, L., Vallot, C., and Vert, J. P. (2021). Machine learning for single-cell genomics data analysis. *Curr. Opin. Syst. Biol.* 26, 64–71. doi: 10.1016/j.coisb.2021.04.006

Ramalingam, A., Kudapa, H., Pazhamala, L. T., Weckwerth, W., and Varshney, R. K. (2015). Proteomics and metabolomics: two emerging areas for legume improvement. *Front. Plant Sci.* 6, 1116.

Ravari, S. Z., Dehghani, H., and Naghavi, H. (2015). Evaluation of relationship between salinity stress tolerance indices and some physiological traits in bread wheat. *Iran. J. Field Crop Sci.* 46 (3), 423–432.

Raza, A., Su, W., Hussain, M. A., Mehmood, S. S., Zhang, X., Cheng, Y., et al. (2021). Integrated analysis of metabolome and transcriptome reveals insights for cold tolerance in rapeseed (Brassica napus L.). *Front. Plant Sci.* 12, 721681.

Redekar, N., Pilot, G., Raboy, V., Li, S., and Saghai Maroof, M. A. (2017). Inference of transcription regulatory network in low phytic acid soybean seeds. *Front. Plant Sci.* 8, 2029. doi: 10.3389/fpls.2017.02029

Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.* 49, 107739. doi: 10.1016/j.biotechadv.2021.107739

Ristaino, J. B., Anderson, P. K., Bebber, D. P., Brauman, K. A., Cunniffe, N. J., Fedoroff, N. V., et al. (2021). The persistent threat of emerging plant disease pandemics to global food security. *Proc. Natl. Acad. Sci.* 118 (23), e2022239118. doi: 10.1073/pnas.2022239118

Sabbatini, F., and Calegari, R. (2023). "Explainable clustering with CREAM," in *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*. 19 (1), 593–603.

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., and Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nat. Ecol. Evol.* 3 (3), 430–439. doi: 10.1038/s41559-018-0793-y

Schaffer, J. D. (1989). *Proceedings of the 3rd International Conference on Genetic Algorithms* (Morgan Kaufmann Publishers Inc).

Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D., and Džeroski, S. (2010). Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinf.* 11, 1–14. doi: 10.1186/1471-2105-11-2

Scholz, M., Gatzek, S., Sterling, A., Fiehn, O., and Selbig, J. (2004). Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics* 20 (15), 2447–2454. doi: 10.1093/bioinformatics/bth270

Schwarz, B., Azodi, C. B., Shiu, S. H., and Bauer, P. (2020). Putative cis-regulatory elements predict iron deficiency responses in Arabidopsis roots. *Plant Physiol.* 182 (3), 1420–1439. doi: 10.1104/pp.19.00760

Shaik, R., and 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 (1), 481–495. doi: 10.1104/pp.113.225862

Shen, D., Wang, Y., Chen, X., Srivastava, V., and Toffolatti, S. L. (2022). Advances in multi-omics study of filamentous plant pathogens. *Front. Microbiol.*, 3222. doi: 10.3389/978-2-83250-034-7

Shikha, M., Kanika, A., Rao, A. R., Mallikarjuna, M. G., Gupta, H. S., and Nepolean, T. (2017). Genomic selection for drought tolerance using genome-wide SNPs in maize. *Front. Plant Sci.* 8, 550. doi: 10.3389/fpls.2017.00550

Silva, J. C. F., Teixeira, R. M., Silva, F. F., Brommonschenkel, S. H., and Fontes, E. P. (2019). Machine learning approaches and their current application in plant molecular biology: A systematic review. *Plant Sci.* 284, 37–47. doi: 10.1016/j.plantsci.2019.03.020

Singh, A. K., Ganapathysubramanian, B., Sarkar, S., and Singh, A. (2018). Deep learning for plant stress phenotyping: trends and future perspectives. *Trends Plant Sci.* 23 (10), 883–898. doi: 10.1016/j.tplants.2018.07.004

Singh, A., Ganapathysubramanian, B., Singh, A. K., and Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends Plant Sci.* 21 (2), 110–124. doi: 10.1016/j.tplants.2015.10.015

Singh, U., Khemka, N., Rajkumar, M. S., Garg, R., and Jain, M. (2017). PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea. *Nucleic Acids Res.* 45 (22), e183–e183.

Slater, G. (2000). Algorithms for analysis of ESTs. PhD thesis (UK: University of Cambridge).

Spindel, J. E., Dahlberg, J., Colgan, M., Hollingsworth, J., Sievert, J., Staggenborg, S. H., et al. (2018). Association mapping by aerial drone reveals 213 genetic associations for Sorghum bicolor biomass traits under drought. *BMC Genomics* 19 (1), 1–18.

Sperduti, A., and Starita, A. (1997). Supervised neural networks for the classification of structures. *IEEE Trans. Neural Networks* 8 (3), 714–735. doi: 10.1109/72.572108

Sperschneider, J., and Dodds, P. N. (2022). EffectorP 3.0: prediction of apoplastic and cytoplasmic effectors in fungi and oomycetes. *Mol. Plant-Microbe Interact.* 35 (2), 146–156. doi: 10.1094/MPMI-08-21-0201-R

Sperschneider, J., Dodds, P. N., Gardiner, D. M., Singh, K. B., and Taylor, J. M. (2018). Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. *Mol. Plant Pathol.* 19 (9), 2094–2110. doi: 10.1111/mpp.12682

Sperschneider, J., Gardiner, D. M., Dodds, P. N., Tini, F., Covarelli, L., Singh, K. B., et al. (2016). EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytol.* 210 (2), 743–761. doi: 10.1111/nph.13794

Sun, S., Wang, C., Ding, H., and Zou, Q. (2020). Machine learning and its applications in plant molecular studies. *Briefings Funct. Genomics* 19 (1), 40–48. doi: 10.1093/bfgp/elz036

Twyman, R. M., Schillberg, S., and Fischer, R. (2013). Optimizing the yield of recombinant pharmaceutical proteins in plants. *Curr. Pharm. Des.* 19 (31), 5486–5494.

Van de Sande, B., Flerin, C., Davie, K., De Waegeneer, M., Hulselmans, G., Aibar, S., et al. (2020). A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* 15 (7), 2247–2276. doi: 10.1038/s41596-020-0336-2

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 (3), 716–729. 10.1016/j.cell.2018.05.061

Vapnik, V. (1999). *The nature of statistical learning theory* (Springer science & business media).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Proceedings of the 31st international conference on neural information processing systems.

Walley, J. W., Sartor, R. C., Shen, Z., Schmitz, R. J., Wu, K. J., Urich, M. A., et al. (2016). Integration of omic networks in a developmental atlas of maize. *Science* 353 (6301), 814–818. doi: 10.1126/science.aag1125

Wang, J., Chen, L., Wang, Y., Zhang, J., Liang, Y., and Xu, D. (2013). A computational systems biology study for understanding salt tolerance mechanism in rice. *PloS One* 8 (6), e64929. doi: 10.1371/journal.pone.0064929

Wang, M., White, N., Grimm, V., Hofman, H., Doley, D., Thorp, G., et al. (2018). Pattern-oriented modelling as a novel way to verify and validate functional–structural plant models: a demonstration with the annual growth module of avocado. *Ann. Bot.* 121 (5), 941–959. doi: 10.1093/aob/mcx187

Wang, X., and Liu, J. (2021). Multiscale parallel algorithm for early detection of tomato gray mold in a complex natural environment. *Front. Plant Sci.* 12, 620273.

Wang, Y., Xu, L., Zou, Q., and Lin, C. (2022). prPred-DRLF: plant R protein predictor using deep representation learning features. *Proteomics* 22 (1-2), 2100161. doi: 10.1002/pmic.202100161

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* 14 (4), 414–416. doi: 10.1038/nmeth.4207

Wilson, T. J., Lai, L., Ban, Y., and Ge, S. X. (2012). Identification of metagenes and their interactions through large-scale analysis of Arabidopsis gene expression data. *BMC Genomics* 13 (1), 1–14. doi: 10.1186/1471-2164-13-237

Woolfson, K. N., Haggitt, M. L., Zhang, Y., Kachura, A., Bjelica, A., Rey Rincon, M. A., et al. (2018). Differential induction of polar and non-polar metabolism during wound-induced suberization in potato (Solanum tuberosum L.) tubers. *Plant J.* 93 (5), 931–942.

Wu, Y., and Zhang, K. (2020). Tools for the analysis of high-dimensional single-cell RNA sequencing data. *Nat. Rev. Nephrol.* 16 (7), 408–421. doi: 10.1038/s41581-020-0262-0

Xavier, A. (2021). Technical nuances of machine learning: implementation and validation of supervised methods for genomic prediction in plant breeding. *Crop Breed. Appl. Biotechnol.* 21, e381421S2. doi: 10.1590/1984-70332021v21sa15

Xu, C., and Jackson, S. A. (2019). Machine learning and complex biological data. *Genome Biol.* 20, 1–4. doi: 10.1186/s13059-019-1689-0

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. Res. Int.* 2015. doi: 10.1155/2015/135782

Xu, Y., Wu, Y., and Wu, J. (2018). Capturing pair-wise epistatic effects associated with three agronomic traits in barley. *Genetica* 146, 161–170. doi: 10.1007/s10709-018-0008-0

Xu, H., Zhang, W., Gao, Y., Zhao, Y., Guo, L., and Wang, J. (2012). Proteomic analysis of embryo development in rice (Oryza sativa). *Planta* 235, 687–701. doi: 10.1007/s00425-011-1535-4

Xu-Hui, W., Ping, S., Li, C., and Ye, W. (2009). "A ROC curve method for performance evaluation of support vector machine with optimization strategy," in *2009 International Forum on Computer Science-Technology and Applications*, Vol. 2. 117–120 (IEEE).

Yan, J., and Wang, X. (2022). Unsupervised and semi-supervised learning: the next frontier in machine learning for plant systems biology. *Plant J.* 111 (6), 1527–1538. doi: 10.1111/tpj.15905

Yan, J., and Wang, X. (2023). Machine learning bridges omics sciences and plant breeding. *Trends Plant Sci.* 28 (2), 199–210. doi: 10.1016/j.tplants.2022.08.018

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 Biol.* 22, 1–24. doi: 10.1186/s13059-021-02492-y

Yan, J., Zou, D., Li, C., Zhang, Z., Song, S., and Wang, X. (2020). SR4R: an integrative SNP resource for genomic breeding and population research in rice. *Genomics Proteomics Bioinf.* 18 (2), 173–185. doi: 10.1016/j.gpb.2020.03.002

Yang, D. L., Yang, Y., and He, Z. (2013). Roles of plant hormones and their interplay in rice immunity. *Mol. Plant* 6 (3), 675–685.

Yang, Z., Li, J. L., Liu, L. N., Xie, Q., and Sui, N. (2020). Photosynthetic regulation under salt stress and salt-tolerance mechanism of sweet sorghum. *Front. Plant Sci.* 10, 1722.

Yang, X., and Guo, T. (2017). Machine learning in plant disease research. doi: 10.18088/ejbmr.3.1.2017.pp6-9

Yang, S., Li, H., He, H., Zhou, Y., and Zhang, Z. (2019). Critical assessment and performance improvement of plant–pathogen protein–protein interaction prediction methods. *Briefings Bioinf.* 20 (1), 274–287. doi: 10.1093/bib/bbx123

Zhang, Q., Chen, Q., Wang, S., Hong, Y., and Wang, Z. (2014). Rice and cold stress: methods for its evaluation and summary of cold tolerance-related quantitative trait loci. *Rice* 7 (1), 1–12.

Zhou, P., Enders, T. A., Myers, Z. A., Magnusson, E., Crisp, P. A., Noshay, J. M., et al. (2022). Prediction of conserved and variable heat and cold stress response in maize using cis-regulatory information. *Plant Cell* 34 (1), 514–534. doi: 10.1093/plcell/koab267

Zhou, Y., Zheng, H., Huang, X., Hao, S., Li, D., and Zhao, J. (2022). Graph neural networks: Taxonomy, advances, and trends. *ACM Trans. Intelligent Syst. Technol. (TIST)* 13 (1), 1–54. doi: 10.1145/3495161

Zhu, Y. N., Shi, D. Q., Ruan, M. B., Zhang, L. L., Meng, Z. H., Liu, J., et al. (2013). Transcriptome analysis reveals crosstalk of responsive genes to multiple abiotic stresses in cotton (Gossypium hirsutum L.). *PloS One* 8 (11), e80218.

Zhu, G., Xiao, H., Guo, Q., Zhang, Z., Zhao, J., and Yang, D. (2018). Effects of cadmium stress on growth and amino acid metabolism in two Compositae plants. *Ecotoxicol. Environ. Saf.* 158, 300–308.