



# Integrating multi-omics and machine learning for disease resistance prediction in legumes

Shameela Mohamedikbal<sup>1,2</sup> · Hawlader A. Al-Mamun<sup>1,2,3</sup> · Mitchell S. Bestry<sup>1,2</sup> · Jacqueline Batley<sup>2</sup> · David Edwards<sup>1,2</sup>

Received: 11 March 2025 / Accepted: 29 May 2025 / Published online: 27 June 2025  
© The Author(s) 2025

## Abstract

**Key message** Multi-omics assisted prediction of disease resistance mechanisms using machine learning has the potential to accelerate the breeding of resistant legume varieties.

**Abstract** Grain legumes, such as soybean (*Glycine max* (L.) Merr.), chickpea (*Cicer arietinum* L.), and lentil (*Lens culinaris* Medik.) play an important role in combating micronutrient malnutrition in the growing human population. However, plant diseases significantly reduce grain yield, causing 10–40% losses in major food crops. The genetic mechanisms associated with disease resistance in legumes have been widely studied using genomic approaches. Multi-omics data encompassing various biological layers such as the transcriptome, epigenome, proteome, and metabolome, in addition to the genome, enables researchers to gain a deeper understanding of these complementary layers and their roles in complex legume-pathogen interactions. Genomic prediction, used to select the best genotypes with desirable traits for breeding, has largely relied on genome-wide markers and statistical approaches to estimate the breeding values of individuals. Integrating multi-omics data into genomic prediction can be achieved using machine learning models, which can capture nonlinear relationships prevalent in high-dimensional data better than traditional statistical methods. This integration may enable more accurate predictions and identification of resistance mechanisms for breeding resistant legumes. Despite its potential, multi-omics integration for disease resistance prediction in legumes has been largely unexplored. In this review, we explore omics studies focusing on disease resistance in legumes and discuss how machine learning models can integrate multi-omics data for disease resistance prediction. Such multi-omics assisted prediction has the potential to reduce the breeding cycle for developing disease-resistant legume varieties.

## Introduction

Micronutrient malnutrition, resulting from insufficient intake and absorption of essential vitamins and minerals, affects human health and contributes to higher maternal and child mortality rates (Black et al. 2013; Bailey et al. 2015; Kiani et al. 2022). Grain legumes, including soybean (*Glycine max* (L.) Merr.), chickpea (*Cicer arietinum* L.), and lentil

(*Lens culinaris* Medik.) are particularly valuable in combating these deficiencies due to their rich micronutrient profiles (Graham and Vance 2003; Rehman et al. 2019). As the global population is expected to reach 9.8 billion by 2050, crop production must increase by 60–110% to meet dietary demands (Tilman et al. 2002; Alexandratos et al. 2012; Ray et al. 2013; United Nations Department of Economic and Social Affairs, Population Division 2022). However, the current annual growth rates of 1.1–1.7% for crops such as wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.) and soybean are insufficient to double crop production by 2050 (Iizumi et al. 2018). This goal is further complicated by plant diseases, which cause yield losses of 10–40% in major food crops (Strange and Scott 2005; Oerke 2006; Ristaino et al. 2021). Additionally, regions with fast-growing populations and existing food shortages experience the highest yield losses in major food crops due to disease (Savary et al. 2019). Interacting climatic variables, such as elevated

Communicated by Damaris Odeny.

✉ David Edwards  
Dave.Edwards@uwa.edu.au

<sup>1</sup> Centre for Applied Bioinformatics, University of Western Australia, Perth, WA 6009, Australia

<sup>2</sup> School of Biological Sciences, University of Western Australia, Perth, WA 6009, Australia

<sup>3</sup> InterGrain Pty Ltd, Perth, WA 6163, Australia

temperatures, increased CO<sub>2</sub> levels, and changing precipitation patterns, alter host resistance and facilitate pathogen evolution, further complicating the impact of disease on grain yield (Garrett et al. 2006; Singh et al. 2023a).

Plant disease management strategies, including crop rotation and chemical use, have been successful; however, deploying host resistance using resistant varieties remains the most effective method (Stuthman et al. 2007; Deng et al. 2020). Plants defend against pathogens using two main mechanisms: qualitative and quantitative resistance. Qualitative resistance is mediated by single major resistance (R) genes, while quantitative resistance is associated with multiple genes or quantitative trait loci (QTLs) (Flor 1971; St Clair 2010). Qualitative resistance provides near-complete protection against specific pathogenic strains but can break down due to continuous pathogen evolution (Leach et al. 2001). Quantitative resistance provides partial but durable resistance against various strains (Lynch and Walsh 1998; Niks et al. 2015). Next-generation sequencing (NGS) technologies have enabled the identification of quantitative resistance loci through QTL mapping and genome-wide association studies (GWAS) (Michael and Jackson 2013; Gangurde et al. 2022). Breeding varieties with strong resistance requires combining diverse loci from both quantitative and qualitative resistance without adversely affecting other desirable traits, such as yield, due to potential growth-defence trade-offs (Nelson et al. 2018; Derbyshire et al. 2024). This necessitates intervention through several biological pathways and mechanisms in resistance breeding, as disease resistance is influenced by multiple genes and environmental factors (Keen 1990; Poland et al. 2009; van Esse et al. 2020). Additionally, the narrow genetic base in self-pollinated legumes such as soybean limits sources of genetic variation for resistance breeding (Hyten et al. 2006).

In plant breeding, genomic selection (GS) enables the selection of plants with desirable traits using a set of genome-wide DNA markers, typically single-nucleotide polymorphisms (SNPs), to estimate the total genetic value or genomic estimated breeding values (GEBVs) of individuals in a breeding population (Meuwissen et al. 2001). This process relies on genomic prediction (GP), which uses statistical models trained on a well-characterised training population that shares some level of relatedness with the breeding population (Meuwissen et al. 2001). The training population consists of individuals with known genetic information and observed phenotypic traits (labelled data) (Meuwissen et al. 2001). The resulting GEBVs serve as pseudo-phenotypes that enable the selection of best parents without further phenotypic evaluations, thereby reducing breeding cycle lengths (Feng et al. 2023). GS therefore enhances the rate of annual genetic improvements in plant breeding, enabling faster development of varieties with desirable traits (Xu et al. 2017; Alemu et al. 2024). However, in the case of polygenic

traits such as disease resistance, genomic data alone may not fully capture the dynamic changes in plant-pathogen interactions. Plant resistance to pathogens is also influenced by intermediate biological processes such as transcription, translation, and metabolic changes that occur in response to infection (Ritchie et al. 2015; Castro-Moretti et al. 2020). To associate quantitative variation in phenotypes with the effects from different biological layers, researchers are adopting multi-omics approaches to study trait variation in plants (Cao et al. 2022; Roychowdhury et al. 2023). These approaches may enable more accurate predictions by integrating interactions between various biological layers into the models (Guo et al. 2016; Wu et al. 2022).

Multi-omics integrates data from various molecular levels, including DNA sequences (genomics), gene expression (transcriptomics), epigenetic modifications (epigenomics), protein (proteomics) and metabolite levels (metabolomics) (Yang et al. 2021). It provides a comprehensive understanding of the complementary biological levels associated with traits such as plant defence, extending beyond genetic markers to include protein and metabolite expression levels, and their role in phenotypic variability within genotypes (Kliebenstein 2012; Ritchie et al. 2015). For example, Shi et al. (2021) integrated transcriptomic and metabolomic data to identify candidate genes and potential metabolites in soybean varieties resistant to soybean cyst nematode infection. Moreover, omics has provided detailed insights into the molecular mechanisms behind growth, development, and stress responses in plants (Yang et al. 2021; Derbyshire et al. 2022).

A challenge in incorporating multi-omics data into GP pipelines lies in the heterogeneous and often unstructured biological data generated from various omics layers (Picard et al. 2021). For instance, genomic data in the form of SNPs are encoded as numeric allele counts (0, 1, 2) for homozygous reference, homozygous alternate and heterozygous alleles, respectively. Transcriptomic data quantifies gene expression as raw counts, showing the number of RNA sequencing reads aligned to each gene in a reference genome (Wang et al. 2009; Finotello and Di Camillo 2015). These gene expression counts can vary from tens for lower-expressed genes to thousands for highly expressed genes. Such variation in the nature and distribution of multi-omics data poses challenges to traditional regression-based GP models that assume a linear relationship between genotype and phenotype (Pérez-Rodríguez et al. 2012). Additionally, higher order genetic interactions between loci, leading to epistasis in complex traits may not be adequately captured by these traditional models (Taylor and Ehrenreich 2015; Varona et al. 2018). Machine learning (ML), a branch of computer science, addresses these limitations by capturing nonlinear interactions inherent in multi-omics data without relying on statistical assumptions (Greener et al. 2022).

ML identifies and extracts hidden biological features from high-dimensional datasets to predict traits for new, unseen data (Samuel 1959; Jordan and Mitchell 2015; Reel et al. 2021). It has been shown to provide accurate genotype-to-phenotype predictions for important agronomic traits by combining high-dimensional datasets such as environmental data, images, and genomic data (Danilevich et al. 2022). ML has been successfully applied in plant disease resistance prediction using genomic information, identifying gene regulatory networks and predicting pathogen effector proteins (Sperschneider 2020; Upadhyaya et al. 2024). The integration of multi-omics data with advancements in artificial intelligence, including ML, into a crop genome design system termed “Breeding 4.0” has been proposed to enable data-driven decisions in plant breeding pipelines (Wallace et al. 2018; Jiang et al. 2020). However, despite its potential, the integration of multi-omics data with ML for disease resistance prediction in legumes remains largely limited. In legumes such as groundnut, breeding for disease resistance has been slow due to its large, tetraploid genome and limited applications of genomic tools (Zhang et al. 2017; Bangaru et al. 2023). Similarly, in faba bean, large genome size, low-density genetic maps, and insufficient development of tightly linked markers for disease resistance has made conventional breeding approaches less effective (Tiwari et al. 2023). Emerging approaches such as ML and multi-omics integration have the potential to accelerate resistance breeding in these and other underutilised legumes.

In this review, we discuss how integrating multi-omics with machine learning has the potential to improve disease resistance prediction in legumes. Multi-omics integration (MOI) in this review refers to combining more than one omics technology for analyses using machine learning models. We will briefly review various omics technologies, discuss machine learning approaches for multi-omics data integration and their applications in trait prediction.

## Overview of multi-omics approaches

Plant-pathogen interactions involve a dynamic exchange in which plants defend against pathogenic attacks, while pathogens evolve mechanisms to disrupt plant defences. These interactions are mediated by molecules produced by both plants and pathogens, including proteins, sugars, small molecules (metabolites) and lipopolysaccharides, which can be captured using individual omics approaches (Boyd et al. 2013). Genomics characterises an organism’s entire genetic material, including both genes and non-coding regions, and provides insights into genetic diversity (Weissenbach 2016). It captures the genetic blue print of plants, identifying genes and genetic variants associated with resistance or susceptibility, such as resistance genes, QTLs, regulatory elements, and structural variations (Neik et al. 2020).

Advances in NGS technologies and high-throughput phenotyping platforms (e.g. unmanned aerial vehicle-based imaging, hyperspectral sensors) have generated vast genomic and phenotypic datasets, enabling researchers to identify SNPs associated with traits and guide breeding strategies (Yang et al. 2020; Marsh et al. 2021). Plant-pathogen interactions are controlled by interconnected regulatory networks which can be captured using transcriptomics (Wani and Ashraf 2018). Transcriptomics quantifies RNA expression levels of transcripts in a cell, known as the transcriptome, including both coding messenger RNAs (mRNAs) and non-coding RNAs (Lowe et al. 2017). By using NGS technologies such as RNA sequencing (RNA-seq), transcriptomics captures gene expression changes across various tissues and cells under specific environmental conditions and developmental stages (Wilhelm et al. 2008; Wang et al. 2009). This approach is particularly useful in host–pathogen studies, as genome-wide transcriptional profiling of susceptible and resistant plants can reveal the molecular basis underlying plant defence responses at different stages of infection (Chen et al. 2013; Kamber et al. 2016; Kandel et al. 2020).

Epigenomics examines the complete set of heritable epigenetic processes such as DNA methylation, histone modifications, and alterations in chromatin structure that regulate gene expression without changing the underlying DNA sequence (Qureshi and Mehler 2011). These processes play an important role in plant defence responses by actively regulating defence-related gene expression during transcriptional reprogramming in response to pathogen attacks (Ding et al. 2012; Zhu et al. 2016; Zhi and Chang 2021). The correlation between mRNA and protein abundance varies greatly across organisms and cell types due to translational regulation (Plotkin 2010; Vogel and Marcotte 2012). This mRNA–protein correlation is often nonlinear and becomes even weaker under stress (Ponnala et al. 2014). Thus, relying solely on transcriptomics is insufficient for understanding protein functions in plant-pathogen dynamics. Proteomics provides an understanding of the expression, function and regulation of the proteins, or proteome, encoded by an organism, which helps to understand the post-translational modifications of proteins involved in plant defence responses (Zhu et al. 2003; Cembrowska-Lech et al. 2023). Metabolomics uses advanced spectroscopy techniques to identify metabolites in plants, and capture the flow of information from the genome through the transcriptome and proteome (Liu and Locasale 2017). Metabolic changes reveal interactions between small molecules and biochemical changes in response to stress, giving insights into metabolomic reprogramming during plant defence (Hong et al. 2016; Kumar et al. 2017). Comprehensive reviews covering potential applications of multi-omics approaches for crop improvement has been published previously (Yang et al. 2021; Zhang et al. 2022; Scossa et al. 2021).

## Machine learning approaches for trait prediction using multi-omics data

The availability of large and dense marker datasets in many plant species have not always translated to improvements in prediction accuracy in GS (Solberg et al. 2008; Norman et al. 2018), often requiring pre-selection of markers for improvement (Raymond et al. 2018). For instance, in sugarcane (*Saccharum sp.*), machine learning models trained on the full genome-wide markers achieved only 50% prediction accuracy for brown rust resistance; however, reducing the marker set through refinement increased the accuracy up to 95% (Aono et al. 2020). Although GWAS have been instrumental in identifying loci associated with important plant agronomic traits, their use in phenotype prediction remains limited in some cases, as QTLs often explain only a small proportion of phenotypic variability, especially in polygenic traits (de los Campos et al. 2010; Hu et al. 2019). In wheat, ML models using only QTL-targeted markers showed lower prediction accuracies for Fusarium head blight-related traits, which improved after combining with genome-wide markers for broad genomic coverage (Rutkoski et al. 2012). These findings show that simply increasing marker density or refining marker set within a single-omics layer does not always yield optimal prediction performance, highlighting the need for more integrative approaches. Such differences in prediction performance could be attributed to trait heritability, linkage disequilibrium between markers, population structure, marker density, prediction models used, and genetic architecture of the trait where the loci influencing the trait may not show purely additive effects (Kaler et al. 2022; Zhang et al. 2019). Additionally, genomic sequence information alone may not be able to capture the impact of downstream regulatory processes and gene interactions on phenotype (Zhu et al. 2012; Ritchie et al. 2015). Other single-omics layers such as transcriptome and metabolome have been explored in predictive models, with their effectiveness varying based on trait and species. Metabolic-based predictions showed low accuracy than SNPs for yield prediction in barley (*Hordeum vulgare* L.) (Gemmer et al. 2020); however, transcriptome-based models have achieved up to 92.86% accuracy in classifying disease stress-responsive genes in maize (*Zea mays* L.) (Nazari et al. 2023). In some cases integrating biological information from omics layers such as transcriptomics and metabolomics along with genomic features into a single model has been shown to improve prediction accuracy using statistical and ML models, in various plant species (Table 1). However, the relative performance of multi-omics data in prediction models may depend on the specific trait and species under study. A combination of genomic and metabolomic data showed

higher prediction power in rice (Wang et al. 2019), while combining mRNA and genomic data improved prediction in maize (Schrag et al. 2018).

The different ML approaches that can be used for MOI are supervised and unsupervised learning (Reel et al. 2021). The goal of supervised learning is to build predictive models for new data using labelled data with known genotype–phenotype relationships (e.g. resistant vs. susceptible) (Danilevicz et al. 2022). Commonly used supervised learning algorithms include Support Vector Machines (SVM) (Cortes and Vapnik 1995), Random Forest (RF) (Breiman 2001), Gradient tree boosting (Friedman 2001) and neural networks (McCulloch and Pitts 1943). SVM can identify nonlinear patterns present in multi-omics data through the use of kernel functions, and they are particularly effective when the number of features such as genes or proteins exceeds the number of samples. Ensemble learning methods such as RF generate multiple decision trees based on random subsets of features and averages their prediction to improve accuracy (Breiman 2001). RF ranks the importance of each feature within a model, making it useful for identifying genes or metabolites contributing to the prediction outcomes. RF captures nonlinear relationships between genotypes and traits, and is resistant to overfitting, making it well suited for high-dimensional multi-omics data analysis (Mukherjee et al. 2024). Unsupervised learning methods identify hidden patterns and structures within multi-omics input data using unlabelled data, without relying on pre-defined trait labels. This approach is helpful to reveal novel associations or groupings related to disease resistance that may not be apparent through traditional analysis methods (Shomorony et al. 2020). Unsupervised methods based on regression, clustering, or network-based have been developed for multi-omics data integration (Vahabi and Michailidis 2022).

Deep learning (DL), a subset of ML, is based on the information processing patterns present in the human brain and uses neural networks to learn from data (Alzubaidi et al. 2021). DL models, especially deep neural networks have been used for MOI due to their ability to identify nonlinear patterns and reduce dimensionality (Krassowski et al. 2020; Kang et al. 2022). Among DL methods, graph neural networks (GNNs) are well suited for MOI as GNNs use a graph architecture for representing data, with nodes representing molecules (e.g. genes, proteins), and edges representing interactions between these molecules (Valous et al. 2024). This architecture allows different omics data to be integrated into a single graph structure, facilitating comprehensive analysis of biological systems (Valous et al. 2024). Gene network prediction model (NetGP) is a DL approach that integrates transcriptomic, genomic and multi-omics data for plant phenotypic prediction (Zhao et al. 2025).

Accessible web-based deep learning tools, such as G2PDeep-v2, enable phenotype prediction in plants using

**Table 1** Recent studies integrating multi-omics data for trait prediction in plants

| Crop                                          | Trait                                                   | Multi-omics data | Model                                   | Performance                                                                                                                                               | References          |
|-----------------------------------------------|---------------------------------------------------------|------------------|-----------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|
| Rice ( <i>Oryza sativa</i> L.)                | Yield                                                   | G, T, M          | MLASSO                                  | G + T + M improved predictive ability to 0.2451 from 0.1588 with G alone                                                                                  | (Hu et al. 2019)    |
| Rice                                          | Yield, 1000-grain weight, grains/panicle, tillers/plant | G, M, T          | BLUP and other seven statistical models | G + M achieved the best prediction for all four traits. BLUP was the most efficient prediction method                                                     | (Wang et al. 2019)  |
| Oat ( <i>Avena sativa</i> L.)                 | Agronomic traits (17 traits)                            | G, T, M          | GBLUP, multi-trait models               | G + T + M improved prediction accuracy for all traits in single environment. Multi-trait omics models outperformed multi-trait GBLUP in multi-environment | (Hu et al. 2021)    |
| Barley ( <i>Hordeum vulgare</i> L.)           | Leaf angle, Heading time, Plant height                  | G, T, M          | GBLUP                                   | T + M improved prediction ability over single predictors, with trait-specific optimal weights                                                             | (Wu et al. 2022)    |
| Arabidopsis ( <i>Arabidopsis thaliana</i> L.) | Complex traits including flowering time                 | G, T, Me         | rrBLUP, RF                              | G + T and G + T + Me improved RF model performance and revealed additional gene interactions                                                              | (Wang et al. 2024a) |
| Maize ( <i>Zea mays</i> L.)                   | Yield                                                   | G, M, Ph         | Linear regression-based and ML models   | G + M + Ph improved prediction accuracy from 0.32 to 0.43 in RF                                                                                           | (Wu et al. 2024)    |

Abbreviations: G, Genomic data; T, Transcriptomic data; M, Metabolomic data; Me, Methylomic data; E, Epigenomic data; P, Proteomic data; Ph, image-based phenomic data; MLASSO, Multilayered Least Absolute Shrinkage and Selection Operator; BLUP, Best Linear Unbiased Prediction; GBLUP, Genomic Best Linear Unbiased Prediction; rrBLUP, ridge regression Genomic Best Linear Unbiased Prediction; RF, Random Forest; AUC-ROC, Area Under the Curve of the Receiver Operating Characteristic

multi-omics data, including copy number variation, gene expression and genetic markers, through algorithms including SVM, RF and multi-convolutional neural networks (Zeng et al. 2024). Multimodal deep learning approaches can be used to integrate additional information such as environmental data or infection-stage images into multi-omics data. Multimodal deep learning enables the fusion of data from different sources and types including images, text and high-dimensional data into a single model (Gao et al. 2020). This approach has the potential to integrate multi-omics data with environmental factors to provide robust prediction accuracy for implementing GS in breeding programmes (Montesinos-López et al. 2024).

### Characteristics of multi-omics data and integration strategies

Biological data from different omics layers typically show high dimensionality, where the number of features is much

larger than the number of samples (large  $p$ , small  $n$ ) (Feldner-Busztin et al. 2023). This high dimensionality can lead to overfitting or spurious associations between genotypes and phenotypes, reducing the generalisability of the data (Athieniti and Spyrou 2023). Omics data vary in their data types (e.g. continuous, numerical, categorical) and statistical distributions, often with high noise levels, leading to data heterogeneity (Athieniti and Spyrou 2023). Biological data also contains highly correlated features due to strong underlying relationships, leading to redundancy when overlapping information is captured across layers (Li et al. 2018). Additionally, data for complex traits such as disease resistance suffer from class imbalance with uneven distribution of phenotypic classes (e.g. unequal representation of resistant vs. susceptible individuals) (Picard et al. 2021). Moreover, certain omics data, such as gene expression, protein abundance, or metabolite levels are inherently non-negative, which needs to be preserved throughout data analysis. These challenges require advanced ML

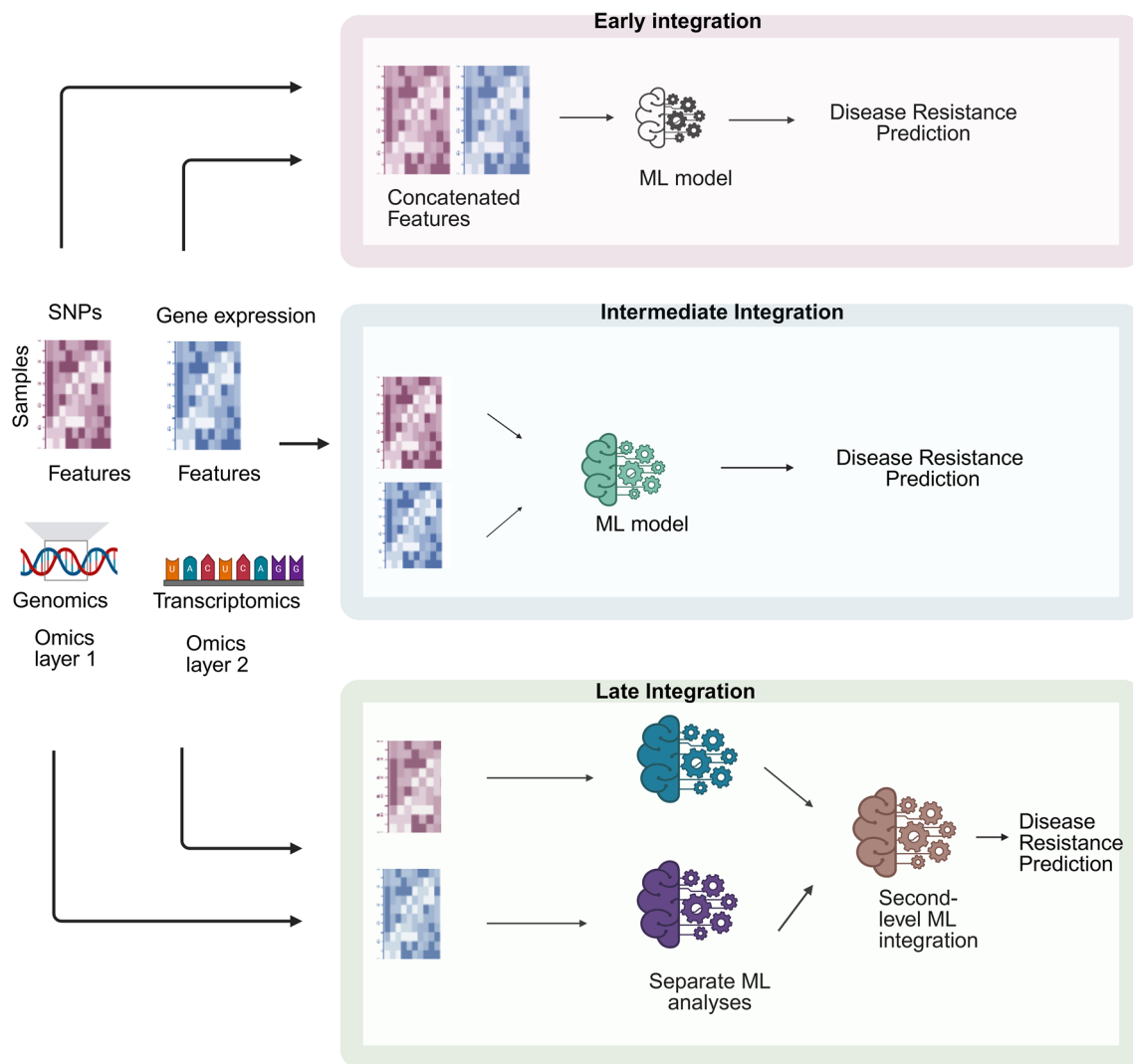


**Fig. 1** Illustration of different integration strategies for multi-omics data. In early integration, features from different omics are concatenated into a single matrix before being fed into the ML model. In intermediate integration, ML algorithms learn a joint representation of the dataset during the analysis stage to develop predictions, preserving the unique characteristics of each dataset. In late integration,

outputs from the separate analysis of each omics dataset are used to train a second-level model, which is used to make the final prediction (Zitnik et al. 2019). Figure created with BioRender.com (Ikbal 2025). Abbreviations: ML, machine learning; SNPs, single-nucleotide polymorphisms

models and specialised MOI strategies to ensure robust analyses (Fig. 1).

Depending on when the multi-omics data are integrated into a machine learning process, MOI strategies can be classified as early, intermediate (joint), or late integration (Fig. 1) (Zitnik et al. 2019). Early integration combines features, which are individual measurable properties associated with a phenotype, from different omics datasets into a single vector before feeding it into a ML model (Zitnik et al. 2019). This approach treats all features as independent without capturing the relationships between the features from different omics layers (Zitnik et al. 2019). In intermediate integration, the algorithm learns a common representation of the data

during the analysis stage while maintaining the unique characteristics of each dataset (Zitnik et al. 2019). This approach does not combine raw input data as in early integration or develop separate models for each dataset but uses algorithms such as multiple kernel learning or deep learning to form a unified representation of the data (Zitnik et al. 2019). Intermediate integration often gives superior performance and allows the model to leverage the relationships across omics datasets, making it particularly effective for predicting complex traits such as disease resistance (Zitnik et al. 2019). The most straightforward strategy is late integration, in which models are initially built separately for each omics data, and then combined by training a second-level model (Zitnik et al.

2019). However, late integration fails to capture interactions among different omics layers (Picard et al. 2021). Comprehensive reviews of MOI for machine learning in various systems have been published previously (Subramanian et al. 2020; Reel et al. 2021; Picard et al. 2021).

### Feature selection and dimensionality reduction

Regardless of the integration strategy, multi-omics datasets require dimensionality reduction to manage the large number of variables, reduce noise, and improve the accuracy of analyses (Meng et al. 2016). For instance, in high-dimensional gene expression data, a number of features or genes can be highly correlated (Clarke et al. 2008). Similarly, significant SNPs identified through GWAS models that do not consider linkage disequilibrium between SNPs may often be linked, causing redundancy (Pudjihartono et al. 2022). As correlated features explain the same information about the data, retaining one of the features is sufficient for phenotype prediction (Chandrashekar and Sahin 2014). In multi-omics datasets, correlated features across omics layers may add more complexity. The extra variables describing redundant information, serve as noise in the dataset, often reducing predictor performance (Kubus 2019). Feature selection (variable elimination) is a dimensionality reduction technique that addresses this issue. It involves selecting a subset of relevant features that explain most of the variance in the dataset, thereby reducing redundant variables (Guyon and Elisseeff 2003). This subset is then used to train the model. Feature selection reduces the risk of overfitting, as models with fewer, more relevant features are less likely to fit noise into the training data, leading to improved prediction on unseen data (Guyon and Elisseeff 2003).

For multi-omics data, selected features might include genomic data such as SNPs associated with disease resistance, and gene expression profiles to identify genes activated during pathogen attack (Liu et al. 2024; Michel et al. 2021). Epigenetic features such as histone modifications, global DNA methylation levels, and genome-wide chromatin accessibility can provide insights into the epigenomic regulatory landscape during plant-pathogen interactions (Ramirez-Prado et al. 2018). Proteome features, including protein abundance levels, post-translational modifications, and metabolomic features such as labelled plant metabolite levels involved in defence pathways can further contribute to building a comprehensive multi-omics model for predicting disease resistance (Castro-Moretti et al. 2020; Elmore et al. 2021). However, in the case of disease resistance prediction in legumes, selecting features based only on their known associations with resistance through supervised feature selection may introduce bias (Ambroise and McLachlan 2002). This may lead to overfitting, as the ML model might only learn from these specific features resulting in reduced

performance for new unseen data (Ambroise and McLachlan 2002). Approaches such as using a variance filter through unsupervised feature selection to select features that are highly variable across all the samples without relying on their disease resistance status have been used previously to reduce bias and improve generalisability (Bommert et al. 2022).

Before training the model, the selected features from various omics layers must be pre-processed and normalised to prevent biases that could arise due to difference in data scale or distributions (Picard et al. 2021). There are numerous feature selection methods available varying in their underlying principles (Effrosynidis and Arampatzis 2021). These methods are classified into filter, wrapper, and embedded methods depending on how the feature selection process interacts with the learning algorithm (Effrosynidis and Arampatzis 2021). The method to be chosen for multi-omics datasets depends on several factors including performance of the method, computation time, dataset characteristics and size (Li et al. 2022). Hybrid methods integrating multiple feature selection techniques through ensemble approaches may give better performance for multi-omics datasets (Claude et al. 2024).

### Machine learning applications in multi-omics integration for predicting disease resistance

#### Integrating single-nucleotide polymorphisms, and genomic features for machine learning-based trait prediction

Advances in genomic technologies and research have enabled characterisation of genetic diversity and resistance loci associated with numerous diseases in a number of legume species (Gangurde et al. 2022). Trait-associated regions identified through GWAS can be refined using local haplotyping by detecting haplotype structures and the linkage landscape surrounding the causal variants within specific genomic regions (Marsh et al. 2023). The stacking of beneficial haplotypes associated with multiple diseases has been shown, through simulation studies, to improve disease resistance (Tong et al. 2024). The development of pangenomes has further enhanced disease resistance studies. Pangenomes reduce reference bias to improve trait prediction accuracy (Danilevicz et al. 2020). They capture the genetic diversity within a species beyond a single reference genome, by identifying both the core conserved genome present in all individuals, and the dispensable genome absent in some individuals (Danilevicz et al. 2020; Bayer et al. 2020). This is important for crop improvement, as rare beneficial alleles present in a few individuals or wild crop relatives can be

introduced to domesticated varieties that lack these alleles (Tay Fernandez et al. 2022). Pangenomes can facilitate the transfer of knowledge from major crops to underutilised crops (Hu et al. 2025) and uncover novel resistance alleles absent from reference genomes, providing breeders with new genetic diversity to improve disease resistance (Golicz et al. 2016; Bayer et al. 2019). The construction of a chickpea super pangenome enabled the identification of a catalogue of R genes, providing a valuable resource for breeding disease-resistant varieties (Khan et al. 2024). Insights from genomics have facilitated the breeding of resistant legumes, such as fusarium wilt-resistant chickpea (Mannur et al. 2019), improved late leaf spot and rust resistant groundnut (*Arachis hypogaea* L.) (Ramakrishnan et al. 2020), and generated elite common bean (*Phaseolus vulgaris* L.) varieties with multiple disease resistance and high seed iron (Portilla et al. 2022), and forms a foundation for ML-based predictive modelling.

The growing availability of genomic resources have enabled their use in ML models for plant disease resistance prediction using features such as pre-selected SNPs, haplotypes, image-derived traits and environmental covariates (Upadhyaya et al. 2024). For bacterial wilt resistance in common bean, whole genome-wide SNPs ( $n = 4568$ ) and preselected SNPs ( $n = 14$ ) were used for genomic prediction with seven models, including RF and SVM (Zia et al. 2022). Prediction accuracy ranged from 0.37 to 0.58 with the full SNP set and 0.30–0.53 with the preselected SNPs in different models. However, only for one bacterial wilt isolate, the accuracy was higher with the selected SNP set in RF model (Zia et al. 2022). Haplotype blocks have the potential to improve prediction accuracy by capturing local epistasis between markers, though the extent of improvement is often trait-, species-, and model-dependent (Weber et al. 2023; Lin et al. 2024). In wheat, LD-based haplotype-tagged SNPs improved prediction accuracy for Fusarium head blight resistance showing correlation higher than models using non-pre-selected SNPs (Alemu et al. 2023). Moreover, genotype-environment interactions that impact phenotypic variability can be effectively captured in ML models through feature engineering that combines genomic and environmental data, thereby enhancing prediction accuracy as shown in maize and adaptable for disease resistance studies in legumes (Fernandes et al. 2024). Beyond genomic information, transcriptomic data provides dynamic changes in gene expression in response to a disease and has the potential to improve predictive power (Michel et al. 2021).

### Transcriptomic and genomic feature integration for disease resistance prediction

RNA-seq studies have identified differentially expressed genes and pathways involved in various legume-pathogen

interactions. For instance, comparative transcriptome profiling in susceptible and resistant chickpea infected with *Fusarium oxysporum* f.sp. *ciceri* (Foc) revealed that Foc genes necessary for pathogen survival are present only in susceptible cultivars, providing potential targets for pathogen intervention and disease management strategies (Upasani et al. 2017). Similarly in common bean, transcriptome profiling identified upregulated sugar transporter genes in susceptible lines during early stages of *Colletotrichum lindemuthianum* (Sacc. & Magnus)-Briosi & Cavara. infection (Padder et al. 2016). Moreover, integrating publicly available data from GWAS and transcriptomics have led to identifying candidate genes associated with immunity-related cellular processes against five fungal pathogens in soybean, underscoring the potential of multi-omics data integration to provide insights into disease resistance mechanisms (Almeida-Silva and Venancio 2021).

ML methods have the potential to improve the accuracy of transcriptomic analyses by identifying novel defence-related gene expression patterns and regulatory interactions, compared to conventional methods (Panahi et al. 2025). For instance, RNA-seq, in vitro transcription factor-DNA profiling and DL were integrated to study the soybean defence response to *Phytophthora sojae* Kaufmann and Gerdemann infections (Hale et al. 2023). Two defence-related transcription factors (TFs) (WRKY and RAV families) were selected from the differentially expressed genes and their binding sites were identified using DAP-seq. These binding sites were used to train Convolutional Recurrent Neural Networks (CRNNs), which predicted novel TF binding sites across the soybean genome with 89–90% test accuracy (Hale et al. 2023). The study further leveraged the homology between Arabidopsis and soybean to build CRNNs for other defence-related TF families with high performance accuracy (> 90% accuracy) in cross-family predictions and moderate accuracy of 60% for Arabidopsis-to-soybean predictions (Hale et al. 2023). Another application of ML on transcriptomic data was demonstrated by Sia et al. (2025) who used transcriptional patterns derived from early stages of infection to predict final disease outcomes across pathosystems. Using a dual-species RNA-seq data for Arabidopsis-*Botrytis cinera*, Sia et al. (2025) developed multiclass classification of disease severity using SVM, RF, XGBoost and Deep Neural Network which were trained on a subset of the transcriptome using 12 feature selection techniques including domain knowledge and co-expression network measures. XGBoost achieved the highest accuracy (72.3%), when trained on dual-species transcriptome data to predict final disease outcome (Sia et al. 2025). They further show that pretrained models on reduced feature selection gene sets from Arabidopsis-*B.cinera* interaction can predict disease outcomes for Arabidopsis-*Sclerotinia sclerotiorum* with RF showing the highest 1-class error accuracy (Sia et al. 2025).

The study further demonstrated transfer learning, applying the pretrained models with the fungal interaction to predict *Arabidopsis-Pseudomonas syringae* bacterial infections.

Integrating genomic data with RNA-seq has been shown to further improve prediction accuracy. A machine learning algorithm, Integrated Multi-Omics Analysis and Machine Learning for Target Gene Prediction (iMAP), was developed to predict calcium signalling genes within resistance-associated loci influencing *Sclerotinia sclerotiorum* resistance in oilseed rape (*Brassica napus* L.) (Wang et al. 2024b). The iMAP framework is a prime example of how multi-omics data including genomic, transcriptomic, and functional annotations can be pre-processed, integrated, and used in ML algorithms to improve resistance gene prediction compared to single-omics approaches. The authors first pre-processed genomic data using PLINK v1.9 (Purcell et al. 2007), and conducted GWAS using three statistical models in rMVP v1.0.0 R-package (Yin et al. 2021). This was followed by haplotype-based GWAS in RAINBOW v0.1.36 R-package (Hamazaki and Iwata 2020) and weighted gene co-expression network analysis (WGCNA) using RNA-seq data. They also compiled gene function annotations for the top significant SNPs and differentially expressed genes from public databases (Wang et al. 2024b). Features were extracted from each data layer including single-SNP GWAS statistics, haplotype GWAS results, WGCNA, gene ontology functional annotations, and combined to a high-dimensional matrix using an early integration strategy. Principal component analysis was used for dimensionality reduction, retaining 99.5% variance. ML models including RF, SVM, XGBoost and logistic regression were trained on these features using scikit-learn library in Python (Wang et al. 2024b). When using only single-SNP GWAS features, RF model showed an accuracy of 0.78 and F1 score of 0.69. However, the RF model using the combined features consistently outperformed single-omics features, with a highest accuracy of 0.95 (Wang et al. 2024b). The iMAP algorithm using the RF model and the combined features also predicted seven calcium signalling genes within three resistance-associated loci on chromosome A06 (Wang et al. 2024b), showing the potential of MOI for improving the prediction of quantitative disease resistance mechanisms and suggests that similar strategies could be effective for legumes.

### Epigenomic data for trait prediction using machine learning

Advancements in third generation NGS technologies, such as nanopore and PacBio DNA sequencing, have enabled genome-wide epigenetic profiling by detecting DNA base modifications without the need for chemical treatments that could degrade DNA (Liu et al. 2019; Tse et al. 2021). As part of their defence strategy, plants can actively demethylate

promoters and regions around defence-related genes, leading to increased gene expression to enhance protection (Annacondia et al. 2021). Such differentially methylated regions have been identified in *Arabidopsis thaliana* (L.) Heynh in response to *Pseudomonas syringae* pv. *tomato* DC3000 infection using whole-genome bisulfite sequencing (Downen et al. 2012). Genome-wide epigenomic and transcriptomic changes in common bean-rust (*Uromyces appendiculatus*) interactions revealed differential acetylation and methylation patterns modulating gene expression in defence-related genes (Ayyappan et al. 2015). Similarly, in soybean, a combined transcriptomic and epigenomic study identified differentially expressed genes associated with defence pathways in resistant lines following infection by *Phytophthora sansomeana*. The study also found that increased CHH methylation levels (where cytosine is followed by non-guanine bases (A, T, or C)) in long intergenic non-coding RNAs suggest that specific methylation patterns may play a role in soybean's defence against *Phytophthora sansomeana* (Lee et al. 2024). Thus, understanding epigenomic variations, such as DNA methylation patterns, in plant-pathogen interaction provides insights into disease resistance mechanisms in crops (Tirnaz and Batley 2019).

However, despite this growing understanding of the role of epigenetic mechanisms in plant defence response, ML studies integrating genome-wide epigenomic data for disease resistance prediction remains largely limited. For other complex traits including flowering time, integrating genomic, transcriptomic, and methylomic data has improved prediction accuracy using RF model revealing additional gene interactions (Wang et al. 2024a, Table 1). Similarly, in barley, including genomic, DNA methylation, and gene expression data in ML model enhanced complex trait prediction including grain yield and nitrogen uptake, explaining a greater proportion of phenotypic variance (0.72–0.92) than genomic models alone (0.55–0.86) (Hansen et al. 2022). Potential of epigenomic data for plant trait prediction has been shown in balsam poplar trees (*Populus balsamifera* L.) where artificial neural networks explained 57.5% and 40.9% phenotypic variance for biomass and wood density respectively (Champigny et al. 2020). Moreover, several ML and DL frameworks for predicting DNA methylation sites have been developed such as Deep6mAPred (Tang et al. 2022), MuLan-Methyl (Zeng et al. 2022), Methyl-GP (Xie et al. 2025) and AMPS (Sereshki et al. 2023) that uses genomic annotations to improve methylation prediction accuracy. Recently, a cross-species hybrid ML model showed that gene expression could be predicted from DNA methylation patterns in fungi, with the authors proposing its adaptation across species including plants (Weinstock et al. 2024). Future efforts are needed to understand the impact of epigenomic data in ML pipelines for predicting novel insights in legume-pathogen interactions.

## Proteomics for predictive trait modelling

Proteomics often complements transcriptomics, using high-throughput technologies such as mass spectrometry, providing insights into molecular mechanisms associated with plant stress response (Lan et al. 2012). For example, a combined transcriptomic and proteomic approach showed that yellow mosaic virus infection in soybean leads to differential expression of genes associated with cell wall biosynthesis, photosynthesis, stress and metabolism pathways (Pavan Kumar et al. 2017). A similar combined approach revealed that inoculating soybean with arbuscular mycorrhizal fungi promotes symbiosis, increasing resistance to root rot. This approach provided insights into the molecular mechanisms by which arbuscular mycorrhizal fungi promotes resistance through the upregulation of proteins such as phenylalanine ammonia-lyase, calcium-dependent protein kinase, and other defence-related proteins (Zhang et al. 2020b). Proteomics provides insights into molecular features that can be targeted

for crop improvement, by aiding in the identification of protein biomarkers and cellular mechanisms involved in plant stress responses (Mustafa and Komatsu 2021).

Proteomic, genomic, transcriptomic, and metabolomic features were extracted and used for prediction of genes encoding enzymes in plant specialised metabolite synthesis in Arabidopsis using an automated ML framework, AutoGluonTabular (Bai et al. 2024) (Table 2). In this study, models trained only on proteomic and genomic features outperformed (AUC-ROC > 0.881) models that were trained using all the multi-omics features. Bai et al. (2024) also evaluated cross-species prediction by training models on Arabidopsis, maize, and tomato (*Solanum lycopersicum* L.) using genomic and proteomic features. A three-species model outperformed single-species prediction with an accuracy of 0.844, showing the value of shared molecular features across species (Bai et al. 2024). A graph-based DL technique, Weighted Gene Autoencoder Multi-Omics Relationship Prediction was used to analyse and predict regulatory

**Table 2** Studies integrating multi-omics data using machine learning for trait prediction in plants

| Crop/pathosystem                                                                                                 | Omics layer                 | ML method                      | Trait predicted                                                  | Accuracy/metrics                                                 | References           |
|------------------------------------------------------------------------------------------------------------------|-----------------------------|--------------------------------|------------------------------------------------------------------|------------------------------------------------------------------|----------------------|
| Common bean/ <i>Curto-bacterium flaccumfaciens</i> pv. <i>flaccumfaciens</i> (Cff)                               | G                           | Seven models including RF, SVM | Bacterial wilt resistance                                        | Up to 0.58 (full SNPs), up to 0.53 (preselected SNPs)            | (Zia et al. 2022)    |
| Soybean/ <i>Phytophthora sojae</i>                                                                               | T, DAP-seq                  | CRNN (DL)                      | TF binding site prediction                                       | 89–90% test accuracy                                             | (Hale et al. 2023)   |
| Arabidopsis/ <i>Botrytis cinera</i> , Arabidopsis/ <i>Sclerotinia sclerotiorum</i> & <i>Pseudomonas syringae</i> | Dual RNA-seq                | SVM, RF, XGBoost, DNN          | Disease class prediction based on early transcriptional response | All models outperformed whole transcriptome-only baseline        | (Sia et al. 2025)    |
| Canola/ <i>Sclerotinia sclerotiorum</i>                                                                          | G, T, functional annotation | RF, SVM, XGBoost, LR           | Resistance to <i>S. sclerotiorum</i>                             | RF with combined features: accuracy 0.95; SNP-only: 0.78         | (Wang et al. 2024b)  |
| Barley                                                                                                           | G, T, E                     | BGAM                           | Complex traits                                                   | G + T + E 0.72–0.92 PVE, G only -0.55–0.86 PVE                   | (Hansen et al. 2022) |
| Multispecies                                                                                                     | G, T, E, P                  | AutoGluon-Tabular              | Metabolite biosynthesis                                          | P and G features achieved the highest accuracy (AUC-ROC > 0.881) | (Bai et al. 2024)    |
| Rice/ <i>Magnaporthe oryzae</i> <i>Oryzae</i>                                                                    | G, T, P                     | WGCNA and graph autoencoder    | Relationship prediction for rice and pathogen                    | Predicted for multi-omics relationship data                      | (Zhao et al. 2023)   |
| Citrus/ <i>Candidatus Liberibacter asiaticus</i>                                                                 | M, microbiome sequencing    | RF, Gradient Boost, Adaboost   | Differentially expressed metabolites & bacteria                  | RF accuracy up to 97.22% ± 2.27%                                 | (Li et al. 2023)     |

Abbreviations: AUC-ROC, area under the curve of the receiver operating characteristic; BGAM, Bayesian generalised additive models; CRNN, convoluted recurrent neural network; DAP-seq, DNA affinity purification sequencing; DL, deep learning; E, epigenomic data; G, genomic data; LR, logistic regression; M, metabolomic data; ML, machine learning; P, proteomic data; PVE, phenotypic variance explained; RF, random forest; RNA-seq, RNA sequencing; SNP, single-nucleotide polymorphisms; SVM, support vector machines; T, transcriptomic data; TF, transcription factor; WGCNA, weighted gene co-expression network analysis

interactions between the fungal pathogen, *Magnaporthe oryzae* *Oryzae* and rice, by integrating genomic, transcriptomic and proteomic data (Zhao et al. 2023). Despite the growing availability of proteomic data, their use in ML frameworks to uncover novel post-translational mechanisms associated with disease resistance remains largely unexplored in plants, unlike human disease studies (Geddes-McAlister and Uhrig 2025). A study in human medicine has shown that deep learning can classify disease states directly from proteomic data obtained through data-independent acquisition mass spectrometry, without peptide or protein identification (Zhang et al. 2020a; Meyer 2021). This approach of converting data matrices to image-like tensors and using convolutional neural networks for classification of healthy and diseased states achieved superior performance compared to models based on identified proteins (Zhang et al. 2020a) and similar strategies can be adapted in legume systems for disease classification.

### Integrating metabolomics with other omics layers for trait prediction

Metabolites are involved in enzyme regulation and can enhance the plant's ability to respond to pathogen attack. For example, metabolomic analysis identified specific sugars and secondary metabolites in soybean resistant to *Phytophthora sojae* Kauffmann and Gerdemann, indicating their potential roles in defence (Zhu et al. 2018). Metabolomics often complements data from other omics layers to reveal pathogen infection mechanisms (Peyraud et al. 2017; Castro-Moretti et al. 2020). A combined transcriptomic and metabolomic analysis showed that *Fusarium solani* f. sp. *phaseoli* infection in common bean leads to cell wall modification and reactive oxygen species generation, highlighting their roles in defence (Chen et al. 2020). Metabolomic approaches have also identified biomarkers including palatinitol and L-proline, which contribute to *Fusarium oxysporum* f. sp. *lentis* resistance in lentils (Foti et al. 2024). Large-scale metabolomic data can be integrated with genomic information to perform metabolite GWAS, enabling the identification of candidate genes associated with specific metabolite profiles (Fang and Luo 2019). In legumes, a metabolic and lipidomics atlas has been developed for multiple species, including chickpea and lentil that can facilitate the study of metabolites involved in disease resistance through metabolite-based GWAS (Bulut et al. 2023).

Metabolomics, when integrated with other multi-omics data has the potential to enhance the understanding of gene regulatory networks, protein interactions and pathways involved to link genotypes to phenotypes (Hao et al. 2025). In a study on huanglongbing-affected citrus cultivars, eight supervised ML classifiers including RF, Gradient Boost and Adaboost, were used for feature selection on metabolomics

and microbiomics sequencing data to identify differentially expressed metabolites and bacteria (Li et al. 2023). RF showed a testing accuracy up to  $97.22\% \pm 2.27\%$  in this study. The ML results were further integrated with the Kyoto Encyclopedia of Genes and Genomes pathway enrichment methods and correlation analysis to identify potential interactions between plant metabolic pathways, such as ABC transporters, which play a major role in plant defence (Li et al. 2023). Similarly, for complex traits such as grain yield, integrating metabolomic, genomic data and phenotyping images from different developmental stages, increased the prediction accuracy in maize from 0.32 to 0.43 using an RF model (Wu et al. 2024). When evaluating feature importance, RF models ranked the metabolite markers higher than the genomic markers, suggesting that continuous quantitative traits provided by metabolites contain more information for predicting yield than binary markers (Wu et al. 2024). These findings show that RF is particularly effective at handling complex interactions within multi-omics data with higher prediction accuracy (Wu et al. 2024).

### Integration of phenotypic imaging with multi-omics data for disease resistance prediction

Recent advancement in phenotyping include automated systems, high-throughput methodologies for large-scale testing, and fine-scale phenotyping of specific organs or tissues at high resolution (Zhao et al. 2019). These high-throughput and automated approaches enable efficient, non-destructive assessment of plant stress response, capturing the dynamic expression of traits and generating temporal and spatial data, including images (Gill et al. 2022). Imaging-based phenotyping has improved disease evaluations, and can be integrated into ML and DL frameworks, to enable models to learn features distinguishing healthy from diseased phenotypes (Dolatabadian et al. 2025). For example, visible and thermal imaging were used to train ML models to predict wilt severity in chickpea under field conditions (Sing et al. 2023b). Visible RGB imaging detects visible disease symptoms such as leaf discolouration, lesions, and wilting, while thermal imaging captures changes in surface temperature, stomatal closure and water stress responses (Zhao et al. 2019). In the study by Singh et al. (2023b), twelve ML models including SVM, RF, Cubist, and k-nearest neighbours were trained on image-derived features, with model combination techniques which further improved wilt severity prediction. Additionally, Bai et al. (2023) used time-series remote sensing data from unmanned aerial vehicles along with accumulated temperature data to train ML models for bacterial blight prediction model in rice. Support vector regression using both spectral and auxiliary traits showed the highest predictive accuracy ( $R_p^2 = 0.85$ ) across geographical locations (Bai et al. 2023). Integration of multi-omics data

with image-derived traits has also been applied to predict complex traits such as yield (Table 1). For instance, Wu et al. (2024) combined genomic, metabolomic, and high-throughput imaging data with ML approaches for maize yield prediction. An explainable deep learning framework (EG-CNN) that integrates multi-omics data with hyperspectral imaging has also been developed for plant disease classification (Shoaib et al. 2023). The model trained on gene expression profiles, metabolite abundances, and hyperspectral images for four diseases (powdery mildew, rust, leaf spot, blight), achieving 95.5% accuracy in classifying disease types (Shoaib et al. 2023). Notably, the results of the EG-CNN model were explained using saliency maps, which shows the regions of the input data that contribute most to the model's output and decision-making. Extending DL models such EG-CNN to legumes could further enhance our understanding of underlying biological mechanisms in diseases.

These studies collectively show the potential of multi-omics and machine learning approaches to understand the complex biological mechanisms behind legume-pathogen interactions for resistance improvement, as well as public resources available for such analyses (Table 3). However, not all machine learning models, integration strategies, or omics data type perform equally well across different datasets or prediction tasks (Picard et al. 2021). Adding more omics layers may introduce noise and redundancy, and optimal integration requires combining biological knowledge along with machine learning approaches (Picard et al. 2021). Prediction accuracy can vary widely depending on data type, integration strategies, specific biological questions, necessitating careful benchmarking and model selection for multi-omics data integration (Flores et al. 2023; Picard et al. 2021).

## Challenges in using machine learning with high-dimensional multi-omics data and future perspectives

Multi-omics data pose several challenges in ML models, due to differences in normalisation, scaling, and data distribution across different layers. For example, metabolomics data may be assigned null values when measurements fall below detection levels (Antonelli et al. 2019). This can be overcome through imputation for missing data including for genotype, epigenomic and proteomic datasets (Song et al. 2020). In disease resistance studies, class imbalance in phenotype values can lead to overfitting in ML models trained using such imbalanced datasets. This can be overcome by over-sampling the minority class or under-sampling the majority class, generating synthetic data using methods such as the synthetic minority oversampling technique, adaptive synthetic sampling, or by measuring the performance of the

model based on weighted or normalised metrics instead of relying solely on the accuracy of the model (He et al. 2008; Jeni et al. 2013; Blagus and Lusa 2013; Reel et al. 2021).

Additionally, ML and DL models require a large volume of data to learn from as their predictive performance decreases with smaller datasets, and this data volume may not be available for all omics layers (Dou et al. 2023). Strategies such as transfer learning, where a model is trained on a source task with abundant data to learn features, and then applied to a target task with less data for prediction can be employed in such cases (Pan and Yang 2010; Weiss et al. 2016). For example, TrG2P, a transfer learning framework has been used for yield prediction by pre-training the model with non-yield traits using convolutional neural networks, and validated in rice, maize, and wheat (Li et al. 2024). The success of ML for MOI also depends on access to high-quality, well-annotated omics datasets that can include features such as histone modifications, chromosomal interactions, and metabolic pathways. Databases such as SoyMD and SoyOmics provide omics datasets for studies in soybean, but expanding such resources to underutilised legumes will help to advance MOI for disease resistance prediction (Liu et al. 2023; Yang et al. 2024). Deploying ML and DL models requires high-performance computing, cloud storage, and efficient data-sharing platforms (Luo et al. 2024). This poses challenges to sustainable deployment and necessitates collaboration and resource sharing among the various sectors for efficient plant disease management (Jeger et al. 2024).

While extensive multi-omics databases are being developed, legumes, particularly underutilised crops such as pigeon pea and cowpea, lack the comprehensive resources available for other major crops. Mining relevant information from existing databases to guide disease control strategies requires accessible bioinformatics tools, ML models and researchers trained to interpret these data. Additionally, improving the accuracy of plant disease assessments by adopting advanced methods, including multispectral and chlorophyll fluorescence imaging, facilitated by recent advancements in artificial intelligence is required (Bock et al. 2022). To improve legume disease resistance prediction, it is important to capture, label and connect large amounts of clean data from various sources including plant breeding companies and research institutions. This will support the training of advanced, complex ML models in agriculture (Bayer and Edwards 2021). Interdisciplinary collaborative efforts for sharing data following FAIR principles such as findable, accessible, interoperable and reuse will accelerate progress in this field (Wilkinson et al. 2016).

Translating research findings into practical breeding programmes requires user-friendly tools tailored for legume breeders. These tools would help to incorporate ML predictions into their decision-making process, and support the challenges of legume breeding, such as balancing disease

**Table 3** Publicly available tools and pipelines for multi-omics integration and trait prediction

| Name                                                                                                                                  | Description                                                                                                          | Focus   | Github link                                                                                                                             |
|---------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|---------|-----------------------------------------------------------------------------------------------------------------------------------------|
| awesome-multi-omics                                                                                                                   | Curated list of multi-omics integration software, including many ML-based prediction methods                         | General | <a href="https://github.com/mikelove/awesome-multi-omics">https://github.com/mikelove/awesome-multi-omics</a>                           |
| IntegratedLearner                                                                                                                     | R-package for multi-omics prediction and classification                                                              | General | <a href="https://github.com/himelmallik/IntegratedLearner">https://github.com/himelmallik/IntegratedLearner</a> (Mallick et al. 2024)   |
| DEM                                                                                                                                   | A multimodal deep learning architecture for phenotypic prediction and functional gene mining of complex traits       | Plant   | <a href="https://github.com/cma2015/DEM/">https://github.com/cma2015/DEM/</a> (Ren et al. 2024)                                         |
| DNNGP                                                                                                                                 | Deep neural network-based method for genomic prediction using multi-omics data in plants                             | Plant   | <a href="https://github.com/AIBreeding/DNNGP">https://github.com/AIBreeding/DNNGP</a> (Wang et al. 2023)                                |
| NetGP                                                                                                                                 | Gene network-based multi-omics integration for genomic prediction in plants                                          | Plant   | <a href="https://github.com/pingT-researcher/NetGP">https://github.com/pingT-researcher/NetGP</a> (Zhao et al. 2025)                    |
| panomiX                                                                                                                               | Platform for multi-omics feature analysis and prediction in plants                                                   | Plant   | <a href="https://github.com/NAMlab/panomiX-tool">https://github.com/NAMlab/panomiX-tool</a> (Sahu et al. 2025)                          |
| Prediction of plant complex traits via integration of multi-omics data                                                                | A study integrating genomic, transcriptomic, and methylation data for complex trait prediction                       | Plant   | <a href="https://github.com/ShiuLab/2024_Ath_GP">https://github.com/ShiuLab/2024_Ath_GP</a> (Wang et al. 2024a)                         |
| AttentionMOI                                                                                                                          | Deep learning algorithm for multi-omics integration to predict cancer subtypes and survival                          | General | <a href="https://github.com/BioAI-kits/AttentionMOI">https://github.com/BioAI-kits/AttentionMOI</a>                                     |
| Integration of multi-omics data and deep phenotyping provides insights into responses to single and combined abiotic stress in potato | A study on integrative multi-omics data analysis using a bioinformatics pipeline based on ML and knowledge networks  | Plant   | <a href="https://github.com/NIB-SI/multiOmics-integration">https://github.com/NIB-SI/multiOmics-integration</a> (Zagorščak et al. 2025) |
| MultiOmicsIntegrator                                                                                                                  | A Nextflow pipeline for integrated omics analyses                                                                    | General | <a href="https://github.com/Bianca-Pasat/MOI">https://github.com/Bianca-Pasat/MOI</a> (Pasat et al. 2024)                               |
| MOVE                                                                                                                                  | Multi-omics variational autoencoder frame for multi-omics integration and identification of cross modal associations | General | <a href="https://github.com/RasmussenLab/MOVE">https://github.com/RasmussenLab/MOVE</a> (Allesøe et al. 2023)                           |

resistance with yield. Furthermore, emerging techniques including single-cell and spatial omics techniques could improve prediction accuracy in legumes, as it provides a clearer picture of changes during legume-pathogen interactions. Similarly, deep learning technologies such as large language models have the potential to generate high impact questions influencing legume-pathogen interactions and complement researchers' investigations (Agathokleous et al. 2024). By addressing these challenges and using advanced computational approaches we can harness the full potential of multi-omics data, to provide more accurate prediction, which will ultimately aid in developing disease-resistant legumes using sustainable practices to meet global food security.

## Conclusions

Multi-omics approaches provide insights into the different biological layers including genes, transcripts, proteins, and metabolites associated with complex traits such as disease resistance. These approaches are important for legumes, in which a narrow genetic diversity restricts the sources of variation available for resistance breeding. As pathogen evolution breaks down plant disease resistance, breeders require greater genetic diversity in their breeding pools to develop disease-resistant varieties. Legumes play an important role in meeting global food security, and integrating multi-omics with machine learning has the potential to decipher complex molecular mechanisms associated with disease resistance, improving prediction accuracy, supporting faster breeding cycles, and enabling more targeted gene editing.

**Acknowledgements** This work was carried out while Shameela Mohamedikbal was in receipt of an Australian Government International Research Training Program Stipend and a University of Western Australia International Fee Scholarship.

**Author contributions** SM, HAA and DE jointly conceived the review; SM reviewed the literature and wrote the manuscript with editing from HAA, MSB, JB and DE. All authors read and approved the final manuscript.

**Funding** Open Access funding enabled and organized by CAUL and its Member Institutions. Australian Research Council, LP230100351, David Edwards.

## Declarations

**Conflict of interest** The authors declare that they have no conflicts of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are

included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Agathokleous E, Rillig MC, Peñuelas J, Yu Z (2024) One hundred important questions facing plant science derived using a large language model. *Trends Plant Sci* 29(2):210–218. <https://doi.org/10.1016/j.tplants.2023.06.008>
- Alemu A, Batista L, Singh PK, Ceplitis A, Chawade A (2023) Haplotype-tagged SNPs improve genomic prediction accuracy for Fusarium head blight resistance and yield-related traits in wheat. *Theor Appl Genet* 136:92. <https://doi.org/10.1007/s00122-023-04352-8>
- Alemu A, Åstrand J, Montesinos-López OA, Sanchez JI, Fernandez-Gonzalez J, Tadesse W, Vetukuri RR, Carlsson AS, Ceplitis A, Crossa J, Ortiz R (2024) Genomic selection in plant breeding: Key factors shaping two decades of progress. *Mol Plant* 17(4):552–578. <https://doi.org/10.1016/j.molp.2024.03.007>
- Alexandratos N, Bruinsma J, Alexandratos N, Bruinsma J (2012) World agriculture towards 2030/2050: the 2012 revision. *PSI Struct Genom Knowled*. <https://doi.org/10.22004/ag.econ.288998>
- Allesøe RL, Lundgaard AT, Hernández Medina R et al (2023) Discovery of drug-omics associations in type 2 diabetes with generative deep-learning models. *Nat Biotechnol* 41:399–408. <https://doi.org/10.1038/s41587-022-01520-x>
- Almeida-Silva F, Venancio TM (2021) Integration of genome-wide association studies and gene coexpression networks unveils promising soybean resistance genes against five common fungal pathogens. *Sci Rep* 11(1):24453. <https://doi.org/10.1038/s41598-021-03864-x>
- Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8(1):53. <https://doi.org/10.1186/s40537-021-00444-8>
- Ambrose C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA* 99:6562–6566. <https://doi.org/10.1073/pnas.102102699>
- Annacondia ML, Markovic D, Reig-Valiente JL, Scaltsoyiannes V, Pieterse CMJ, Ninkovic V, Slotkin RK, Martinez G (2021) Aphid feeding induces the relaxation of epigenetic control and the associated regulation of the defense response in Arabidopsis. *New Phytol* 230(3):1185–1200. <https://doi.org/10.1111/nph.17226>
- Antonelli J, Claggett BL, Henglin M, Kim A, Ovsak G, Kim N, Deng K, Rao K, Tyagi O, Watrous JD, Lagerborg KA, Hushcha PV, Demler OV, Mora S, Niiranen TJ, Pereira AC, Jain M, Chen S (2019) Statistical workflow for feature selection in human metabolomics data. *Metabolites* 9(7):143. <https://doi.org/10.3390/metabo9070143>
- Aono AH, Costa EA, Rody HVS et al (2020) Machine learning approaches reveal genomic regions associated with sugarcane brown rust resistance. *Sci Rep* 10:20057. <https://doi.org/10.1038/s41598-020-77063-5>
- Athieniti E, Spyrou GM (2023) A guide to multi-omics data collection and integration for translational medicine. *Comput Struct Biotechnol J* 21:134–149. <https://doi.org/10.1016/j.csbj.2022.11.050>
- Ayyappan V, Kalavacharla V, Thimmapuram J, Bhide KP, Sripathi VR, Smolinski TG, Manoharan M, Thurston Y, Todd A, Kingham B (2015) Genome-wide profiling of histone modifications

- (H3K9me2 and H4K12ac) and gene expression in rust (*Uromyces appendiculatus*) inoculated common bean (*Phaseolus vulgaris* L.). *PLoS ONE* 10(7):e0132176. <https://doi.org/10.1371/journal.pone.0132176>
- Bailey RL, West KP, Black RE (2015) The epidemiology of global micronutrient deficiencies. *Ann Nutr Metab* 66(Suppl 2):22–33. <https://doi.org/10.1159/000371618>
- Bai X, Fang H, He Y, Zhang J, Tao M, Wu Q, Yang G, Wei Y, Tang Y, Tang L, Lou B, Deng S, Yang Y, Feng X (2023) Dynamic UAV phenotyping for rice disease resistance analysis based on multi-source data. *Plant Phenomics* 5:0019. <https://doi.org/10.34133/plantphenomics.0019>
- Bai W, Li C, Li W, Wang H, Han X, Wang P, Wang L (2024) Machine learning assists prediction of genes responsible for plant specialized metabolite biosynthesis by integrating multi-omics data. *BMC Genom* 25(1):418. <https://doi.org/10.1186/s12864-024-10258-6>
- Bangaru K, Mathew A, Bagudam R et al (2023) Next-generation crop breeding approaches for improving disease resistance in groundnut (*Arachis hypogaea* L.). In: Jha UC, Nayyar H, Sharma KD et al (eds) *Diseases in legume crops: next generation breeding approaches for resistant legume crops*. Springer Nature Singapore, Singapore, pp 195–232
- Bayer PE, Edwards D (2021) Machine learning in agriculture: from silos to marketplaces. *Plant Biotechnol J* 19(4):648–650. <https://doi.org/10.1111/pbi.13521>
- Bayer PE, Goliz AA, Tirnaz S, Chan C-KK, Edwards D, Batley J (2019) Variation in abundance of predicted resistance genes in the Brassica oleracea pangenome. *Plant Biotechnol J* 17(4):789–800. <https://doi.org/10.1111/pbi.13015>
- Bayer PE, Goliz AA, Scheben A, Batley J, Edwards D (2020) Plant pan-genomes are the new reference. *Nat Plants* 6(8):914–920. <https://doi.org/10.1038/s41477-020-0733-0>
- Black RE, Victora CG, Walker SP, Bhutta ZA, Christian P, de Onis M, Ezzati M, Grantham-McGregor S, Katz J, Martorell R, Uauy R, Maternal and Child Nutrition Study Group (2013) Maternal and child undernutrition and overweight in low-income and middle-income countries. *Lancet* 382(9890):427–451. [https://doi.org/10.1016/S0140-6736\(13\)60937-X](https://doi.org/10.1016/S0140-6736(13)60937-X)
- Blagus R, Lusa L (2013) SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform* 14:106. <https://doi.org/10.1186/1471-2105-14-106>
- Bock CH, Barbedo JGA, Mahlein A-K, Del Ponte EM (2022) A special issue on phytopathometry — visual assessment, remote sensing, and artificial intelligence in the twenty-first century. *Trop Plant Pathol* 47(1):1–4. <https://doi.org/10.1007/s40858-022-00498-w>
- Bommert A, Welchowski T, Schmid M, Rahnenführer J (2022) Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Brief Bioinform* 23(1):bbab354. <https://doi.org/10.1093/bib/bbab354>
- Boyd LA, Ridout C, O'Sullivan DM, Leach JE, Leung H (2013) Plant-pathogen interactions: disease resistance in modern agriculture. *Trends Genet* 29(4):233–240. <https://doi.org/10.1016/j.tig.2012.10.011>
- Braich S, Sudheesh S, Forster J, Kaur S (2017) Characterisation of Faba Bean (*Vicia faba* L.) transcriptome using RNA-Seq: sequencing, De Novo assembly, annotation, and expression analysis. *Agronomy* 7(3):53. <https://doi.org/10.3390/agronomy7030053>
- Breiman L (2001) Random forests. *Springer Science and Business Media LLC* 45:5–32. <https://doi.org/10.1023/a:1010933404324>
- Bulut M, Wendenburg R, Bitocchi E, Bellucci E, Kroc M, Gioia T, Susek K, Papa R, Fernie AR, Alseikh S (2023) A comprehensive metabolomics and lipidomics atlas for the legumes common bean, chickpea, lentil and lupin. *Plant J* 116(4):1152–1171. <https://doi.org/10.1111/tpj.16329>
- Cao P, Zhao Y, Wu F, Xin D, Liu C, Wu X, Lv J, Chen Q, Qi Z (2022) Multi-omics techniques for soybean molecular breeding. *Int J Mol Sci* 23(9):4994. <https://doi.org/10.3390/ijms23094994>
- Castro-Moretti FR, Gentzel IN, Mackey D, Alonso AP (2020) Metabolomics as an emerging tool for the study of plant-pathogen interactions. *Metabolites* 10(2):52. <https://doi.org/10.3390/metabo10020052>
- Cembrowska-Lech D, Krzemińska A, Miller T et al (2023) An integrated multi-omics and artificial intelligence framework for advance plant phenotyping in horticulture. *Biology (Basel)* 12. <https://doi.org/10.3390/biology12101298>
- Champigny MJ, Unda F, Skyba O, Soolanayakanahally RY, Mansfield SD, Campbell MM (2020) Learning from methylomes: epigenomic correlates of *Populus balsamifera* traits based on deep learning models of natural DNA methylation. *Plant Biotechnol J* 18:1361–1375. <https://doi.org/10.1111/pbi.13299>
- Chandrashekar G, Sahin F (2014) A survey on feature selection methods. *Comput Electr Eng* 40(1):16–28. <https://doi.org/10.1016/j.compeleceng.2013.11.024>
- Chen T, Lv Y, Zhao T, Li N, Yang Y, Yu W, He X, Liu T, Zhang B (2013) Comparative transcriptome profiling of a resistant vs. susceptible tomato (*Solanum lycopersicum*) cultivar in response to infection by tomato yellow leaf curl virus. *PLoS ONE* 8(11):80816. <https://doi.org/10.1371/journal.pone.0080816>
- Chen L, Zhao Q, Dong K, Yang Z, Dong Y (2019) Physiological mechanism of faba bean *Fusarium* wilt promoted by benzoic acid and cinnamic acid. *J Plant Protect* 46(2):298–304
- Chen L, Wu Q, He T, Lan J, Ding L, Liu T, Wu Q, Pan Y, Chen T (2020) Transcriptomic and metabolomic changes triggered by *Fusarium solani* in common bean (*Phaseolus vulgaris* L.). *Genes* 11(2):177. <https://doi.org/10.3390/genes11020177>
- Clarke R, Ransom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* 8:37–49. <https://doi.org/10.1038/nrc2294>
- Claude E, Leclercq M, Thébaud P, Droit A, Uricaru R (2024) Optimizing hybrid ensemble feature selection strategies for transcriptomic biomarker discovery in complex diseases. *NAR Genom Bioinform* 6(3):lqae79. <https://doi.org/10.1093/nargab/lqae079>
- Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3):273–297. <https://doi.org/10.1007/BF00994018>
- Danilevicz MF, Tay Fernandez CG, Marsh JI, Bayer PE, Edwards D (2020) Plant pangenomics: approaches, applications and advancements. *Curr Opin Plant Biol* 54:18–25. <https://doi.org/10.1016/j.pbi.2019.12.005>
- Danilevicz MF, Gill M, Anderson R, Batley J, Bennamoun M, Bayer PE, Edwards D (2022) Plant genotype to phenotype prediction using machine learning. *Front Genet* 13:822173. <https://doi.org/10.3389/fgene.2022.822173>
- Demirjian C, Vailleau F, Berthomé R, Roux F (2023) Genome-wide association studies in plant pathosystems: success or failure? *Trends Plant Sci* 28(4):471–485. <https://doi.org/10.1016/j.tplants.2022.11.006>
- Deng Y, Ning Y, Yang D-L, Zhai K, Wang G-L, He Z (2020) Molecular basis of disease resistance and perspectives on breeding strategies for resistance improvement in crops. *Mol Plant* 13(10):1402–1419. <https://doi.org/10.1016/j.molp.2020.09.018>
- Derbyshire MC, Batley J, Edwards D (2022) Use of multiple 'omics techniques to accelerate the breeding of abiotic stress tolerant crops. *Current Plant Biology* 32:100262. <https://doi.org/10.1016/j.cpb.2022.100262>
- Derbyshire MC, Newman TE, Thomas WJW, Batley J, Edwards D (2024) The complex relationship between disease resistance and yield in crops. *Plant Biotechnol J* 22(9):2612–2623. <https://doi.org/10.1111/pbi.14373>

- Campos de los G, Gianola D, Allison DB (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat Rev Genet* 11(12):880–886. <https://doi.org/10.1038/nrg2898>
- Ding B, del Bellizzi M, Ning Y, Meyers BC, Wang GL (2012) HDT701, a histone H4 deacetylase, negatively regulates plant innate immunity by modulating histone H4 acetylation of defense-related genes in rice. *Plant Cell* 24(9):3783–3794. <https://doi.org/10.1105/tpc.112.101972>
- Dolatabadian A, Neik TX, Danilevich MF, Upadhyaya SR, Batley J, Edwards D (2025) Image-based crop disease detection using machine learning. *Plant Pathol* 74:18–38. <https://doi.org/10.1111/ppa.14006>
- Dou B, Zhu Z, Merkurjev E, Ke L, Chen L, Jiang J, Zhu Y, Liu J, Zhang B, Wei G-W (2023) Machine learning methods for small data challenges in molecular science. *Chem Rev* 123(13):8736–8780. <https://doi.org/10.1021/acs.chemrev.3c00189>
- Downen RH, Pelizzola M, Schmitz RJ, Lister R, Downen JM, Nery JR, Dixon JE, Ecker JR (2012) Widespread dynamic DNA methylation in response to biotic stress. *Proc Natl Acad Sci USA* 109(32):E2183–E2191. <https://doi.org/10.1073/pnas.1209329109>
- Effrosynidis D, Arampatzis A (2021) An evaluation of feature selection methods for environmental data. *Ecol Inform* 61:101224. <https://doi.org/10.1016/j.ecoinf.2021.101224>
- Elmore JM, Griffin BD, Walley JW (2021) Advances in functional proteomics to study plant-pathogen interactions. *Curr Opin Plant Biol* 63:102061. <https://doi.org/10.1016/j.pbi.2021.102061>
- Fang C, Luo J (2019) Metabolic GWAS-based dissection of genetic bases underlying the diversity of plant metabolism. *Plant J* 97(1):91–100. <https://doi.org/10.1111/tpj.14097>
- Feldner-Busztin D, Firbas Nisantzis P, Edmunds SJ et al (2023) Dealing with dimensionality: the application of machine learning to multi-omics data. *Bioinformatics* 39. <https://doi.org/10.1093/bioinformatics/btad021>
- Feng W, Gao P, Wang X (2023) AI breeder: genomic predictions for crop breeding. *New Crops*. <https://doi.org/10.1016/j.ncrops.2023.12.005>
- Fernandes IK, Vieira CC, Dias KOG, Fernandes SB (2024) Using machine learning to combine genetic and environmental data for maize grain yield predictions across multi-environment trials. *Theor Appl Genet* 137:189. <https://doi.org/10.1007/s00122-024-04687-w>
- Finotello F, Di Camillo B (2015) Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief Funct Genom* 14(2):130–142. <https://doi.org/10.1093/bfpg/elu035>
- Flor HH (1971) Current status of the gene-for-gene concept. *Annu Rev Phytopathol* 9(1):275–296. <https://doi.org/10.1146/annurev.py.09.090171.001423>
- Flores JE, Claborne DM, Weller ZD, Webb-Robertson B-J, Waters KM, Bramer LM (2023) Missing data in multi-omics integration: Recent advances through artificial intelligence. *Front Artif Intell* 6:1098308. <https://doi.org/10.3389/frai.2023.1098308>
- Foti C, Zambounis A, Bataka EP, Kalloniati C, Panagiotaki E, Nakas CT, Flemetakis E, Pavli OI (2024) Metabolic aspects of lentil-fusarium interactions. *Plants* 13(14):2005. <https://doi.org/10.3390/plants13142005>
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 29(5):1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Gangurde SS, Xavier A, Naik YD, Jha UC, Rangari SK, Kumar R, Reddy MSS, Channale S, Elango D, Mir RR, Zwart R, Laxuman C, Sudini HK, Pandey MK, Punhuri S, Mendu V, Reddy UK, Guo B, Gangarao NVPR, Sharma VK, Wang X, Zhao C, Thudi M (2022) Two decades of association mapping: Insights on disease resistance in major crops. *Front Plant Sci* 13:1064059. <https://doi.org/10.3389/fpls.2022.1064059>
- Gao J, Li P, Chen Z, Zhang J (2020) A survey on deep learning for multimodal data fusion. *Neural Comput* 32(5):829–864. [https://doi.org/10.1162/neco\\_a\\_01273](https://doi.org/10.1162/neco_a_01273)
- Garrett KA, Dendy SP, Frank EE, Rouse MN, Travers SE (2006) Climate change effects on plant disease: genomes to ecosystems. *Annu Rev Phytopathol* 44:489–509. <https://doi.org/10.1146/annurev.phyto.44.070505.143420>
- Geddes-McAlister J, Uhrig RG (2025) The plant proteome delivers from discovery to innovation. *Trends Plant Sci*. <https://doi.org/10.1016/j.tplants.2025.03.003>
- Gemmer MR, Richter C, Jiang Y, Schmutzer T, Raorane ML, Junker B, Pillen K, Maurer A (2020) Can metabolic prediction be an alternative to genomic prediction in barley? *PLoS ONE* 15:e0234052. <https://doi.org/10.1371/journal.pone.0234052>
- Gill T, Gill SK, Saini DK et al (2022) A comprehensive review of high throughput phenotyping and machine learning for plant stress phenotyping. *Phenomics* 2:156–183. <https://doi.org/10.1007/s43657-022-00048-z>
- Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CKK, Severn-Ellis A, McCombie WR, Parkin IAP, Paterson AH, Pires JC, Sharpe AG, Tang H, Teakle GR, Town CD, Batley J, Edwards D (2016) The pangenome of an agronomically important crop plant Brassica oleracea. *Nat Commun* 7:13390. <https://doi.org/10.1038/ncomms13390>
- Graham PH, Vance CP (2003) Legumes: importance and constraints to greater use. *Plant Physiol* 131(3):872–877. <https://doi.org/10.1104/pp.017004>
- Greener JG, Kandathil SM, Moffat L, Jones DT (2022) A guide to machine learning for biologists. *Nat Rev Mol Cell Biol* 23(1):40–55. <https://doi.org/10.1038/s41580-021-00407-0>
- Guo Z, Magwire MM, Basten CJ, Xu Z, Wang D (2016) Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theor Appl Genet* 129(12):2413–2427. <https://doi.org/10.1007/s00122-016-2780-5>
- Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *Journal of machine learning research*
- Hale B, Ratnayake S, Flory A, Wijeratne R, Schmidt C, Robertson AE, Wijeratne AJ (2023) Gene regulatory network inference in soybean upon infection by *Phytophthora sojae*. *PLoS ONE* 18(7):e0287590. <https://doi.org/10.1371/journal.pone.0287590>
- Hamazaki K, Iwata H (2020) RAINBOW: Haplotype-based genome-wide association study using a novel SNP-set method. *PLoS Comput Biol* 16:e1007663. <https://doi.org/10.1371/journal.pcbi.1007663>
- Hansen PB, Ruud AK, de Los Campos G et al (2022) Integration of DNA methylation and transcriptome data improves complex trait prediction in *hordeum vulgare*. *Plants* 11. <https://doi.org/10.3390/plants11172190>
- Hao Y, Zhang Z, Luo E, Yang J, Wang S (2025) Plant metabolomics: applications and challenges in the era of multi-omics big data. *aBIOTECH* 6:116–132. <https://doi.org/10.1007/s42994-024-00194-0>
- He H, Bai Y, Garcia EA, Li S (2008) ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE, pp 1322–1328
- Hong J, Yang L, Zhang D, Shi J (2016) Plant metabolomics: an indispensable system biology tool for plant science. *Int J Mol Sci* 17(6). <https://doi.org/10.3390/ijms17060767>
- Hu X, Xie W, Wu C, Xu S (2019) A directed learning strategy integrating multiple omic data improves genomic prediction. *Plant Biotechnol J* 17(10):2011–2020. <https://doi.org/10.1111/pbi.13117>

- Hu H, Campbell MT, Yeats TH, Zheng X, Runcie DE, Covarrubias-Pazarán G, Broeckling C, Yao L, Caffè-Tremi M, Gutiérrez LA, Smith KP, Tanaka J, Hoekenga OA, Sorrells ME, Gore MA, Jannink J-L (2021) Multi-omics prediction of oat agronomic and seed nutritional traits across environments and in distantly related populations. *Theor Appl Genet* 134(12):4043–4054. <https://doi.org/10.1007/s00122-021-03946-4>
- Hu B, Zheng Y, Lv J, Zhang J, Dong Y (2023) Proteomic analysis of the faba bean-wheat intercropping system in controlling the occurrence of faba bean fusarium wilt due to stress caused by *Fusarium oxysporum* f. sp. *fabae* and benzoic acid. *BMC Plant Biol* 23(1):472. <https://doi.org/10.1186/s12870-023-04481-8>
- Hu H, Zhao J, Thomas WJW, Batley J, Edwards D (2025) The role of pangenomics in orphan crop improvement. *Nat Commun* 16:118. <https://doi.org/10.1038/s41467-024-55260-4>
- Hyten DL, Song Q, Zhu Y, Choi I-Y, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB (2006) Impacts of genetic bottlenecks on soybean genome diversity. *Proc Natl Acad Sci USA* 103(45):16666–16671. <https://doi.org/10.1073/pnas.0604379103>
- Iizumi T, Kotoku M, Kim W et al (2018) Uncertainties of potentials and recent changes in global yields of major crops resulting from census- and satellite-based yield datasets at multiple resolutions. *PLoS ONE* 13(9):e0203809. <https://doi.org/10.1371/journal.pone.0203809>
- Ikbāl S (2025) Fig 1. Created in BioRender. <https://BioRender.com/k42q351>
- Jeger M, Beresford R, Berlin A, Bock C, Fox A, Gold KM, Newton AC, Vicent A, Xu X (2024) Impact of novel methods and research approaches in plant pathology: Are individual advances sufficient to meet the wider challenges of disease management? *Plant Pathol* 73(7):1629–1655. <https://doi.org/10.1111/ppa.13927>
- Jeni LA, Cohn JF, De La Torre F (2013) Facing imbalanced data recommendations for the use of performance metrics. *Int Conf Affect Comput Intell Interact Workshops* 2013:245–251. <https://doi.org/10.1109/AACII.2013.47>
- Jiang S, Cheng Q, Yan J, Fu R, Wang X (2020) Genome optimization for improvement of maize breeding. *Theor Appl Genet* 133(5):1491–1502. <https://doi.org/10.1007/s00122-019-03493-z>
- Jones JDG, Dangl JL (2006) The plant immune system. *Nature* 444(7117):323–329. <https://doi.org/10.1038/nature05286>
- Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. *Science* 349(6245):255–260. <https://doi.org/10.1126/science.aaa8415>
- Kaler AS, Purcell LC, Beissinger T, Gillman JD (2022) Genomic prediction models for traits differing in heritability for soybean, rice, and maize. *BMC Plant Biol* 22:87. <https://doi.org/10.1186/s12870-022-03479-y>
- Kamber T, Buchmann JP, Pothier JF, Smits THM, Wicker T, Duffy B (2016) Fire blight disease reactome: RNA-seq transcriptional profile of apple host plant defense responses to *Erwinia amylovora* pathogen infection. *Sci Rep* 6:21600. <https://doi.org/10.1038/srep21600>
- Kandel SL, Hulse-Kemp AM, Stoffel K, Koike ST, Shi A, Mou B, Van Deynze A, Klosterman SJ (2020) Transcriptional analyses of differential cultivars during resistant and susceptible interactions with *Peronospora effusa*, the causal agent of spinach downy mildew. *Sci Rep* 10(1):6719. <https://doi.org/10.1038/s41598-020-63668-3>
- Kang M, Ko E, Mersha TB (2022) A roadmap for multi-omics data integration using deep learning. *Brief Bioinform*. <https://doi.org/10.1093/bib/bbab454>
- Keen NT (1990) Gene-for-gene complementarity in plant-pathogen interactions. *Annu Rev Genet* 24:447–463. <https://doi.org/10.1146/annurev.ge.24.120190.002311>
- Khan AW, Garg V, Sun S, Gupta S, Dudchenko O, Roorkiwal M, Chitkineni A, Bayer PE, Shi C, Upadhyaya HD, Bohra A, Bharadwaj C, Mir RR, Baruch K, Yang B, Coyne CJ, Bansal KC, Nguyen HT, Ronen G, Aiden EL, Veneklaas E, Siddique KHM, Liu X, Edwards D, Varshney RK (2024) Cicer super-pangenome provides insights into species evolution and agronomic trait loci for crop improvement in chickpea. *Nat Genet* 56(6):1225–1234. <https://doi.org/10.1038/s41588-024-01760-4>
- Kiani AK, Dhuli K, Donato K, Aquilanti B, Velluti V, Matera G, Iaconelli A, Connelly ST, Bellinato F, Gisondi P, Bertelli M (2022) Main nutritional deficiencies. *J Prev Med Hyg* 63:E93–E101. <https://doi.org/10.15167/2421-4248/jpmh2022.63.2S3.2752>
- Kliebenstein DJ (2012) Plant defense compounds: systems approaches to metabolic analysis. *Annu Rev Phytopathol* 50:155–173. <https://doi.org/10.1146/annurev-phyto-081211-172950>
- Krassowski M, Das V, Sahu SK, Misra BB (2020) State of the field in multi-omics research: from computational needs to data mining and sharing. *Front Genet* 11:610798. <https://doi.org/10.3389/fgene.2020.610798>
- Kubus M (2019) The problem of redundant variables in random forests. *Acta Universitatis Lodzensis Folia Oeconomica* 6(339):7–16. <https://doi.org/10.18778/0208-6018.339.01>
- Kumar R, Bohra A, Pandey AK, Pandey MK, Kumar A (2017) Metabolomics for plant improvement: status and prospects. *Front Plant Sci* 8:1302. <https://doi.org/10.3389/fpls.2017.01302>
- Lan P, Li W, Schmidt W (2012) Complementary proteome and transcriptome profiling in phosphate-deficient *Arabidopsis* roots reveals multiple levels of gene regulation. *Mol Cell Proteom* 11(11):1156–1166. <https://doi.org/10.1074/mcp.M112.020461>
- Leach JE, Vera Cruz CM, Bai J, Leung H (2001) Pathogen fitness penalty as a predictor of durability of disease resistance genes. *Annu Rev Phytopathol* 39:187–224. <https://doi.org/10.1146/annurev.phyto.39.1.187>
- Lee G, DiBiase CN, Liu B, Li T, McCoy AG, Chilvers MI, Sun L, Wang D, Lin F, Zhao M (2024) Transcriptomic and epigenetic responses shed light on soybean resistance to *Phytophthora sansomeana*. *Plant Genome* 17:e20487. <https://doi.org/10.1002/tpg2.20487>
- Li Y, Wu F-X, Ngom A (2018) A review on machine learning principles for multi-view biological data integration. *Brief Bioinform* 19(2):325–340. <https://doi.org/10.1093/bib/bbw113>
- Li Y, Mansmann U, Du S, Hornung R (2022) Benchmark study of feature selection strategies for multi-omics data. *BMC Bioinform* 23(1):412. <https://doi.org/10.1186/s12859-022-04962-x>
- Li J, Gmitter FG, Zhang B, Wang Y (2023) Uncovering interactions between plant metabolism and plant-associated bacteria in Huanglongbing-affected citrus cultivars using multiomics analysis and machine learning. *J Agric Food Chem* 71(43):16391–16401. <https://doi.org/10.1021/acs.jafc.3c04460>
- Li J, Zhang D, Yang F, Zhang Q, Pan S, Zhao X, Zhang Q, Han Y, Yang J, Wang K, Zhao C (2024) TrG2P: A transfer-learning-based tool integrating multi-trait data for accurate prediction of crop yield. *Plant Commun* 5(7):100975. <https://doi.org/10.1016/j.xplc.2024.100975>
- Lin Y-C, Mayer M, Valle Torres D et al (2024) Genomic prediction within and across maize landrace derived populations using haplotypes. *Front Plant Sci* 15:1351466. <https://doi.org/10.3389/fpls.2024.1351466>
- Liu X, Locasale JW (2017) Metabolomics: a primer. *Trends Biochem Sci* 42:274–284. <https://doi.org/10.1016/j.tibs.2017.01.004>
- Liu X, Wang H, Wang H, Guo Z, Xu X, Liu J, Wang S, Li W-X, Zou C, Prasanna BM, Olsen MS, Huang C, Xu Y (2018) Factors affecting genomic selection revealed by empirical evidence in maize. *Crop J* 6(4):341–352. <https://doi.org/10.1016/j.cj.2018.03.005>
- Liu Q, Fang L, Yu G, Wang D, Xiao C-L, Wang K (2019) Detection of DNA base modifications by deep recurrent neural network on

- Oxford Nanopore sequencing data. *Nat Commun* 10(1):2449. <https://doi.org/10.1038/s41467-019-10168-2>
- Liu Y, Zhang Y, Liu X, Shen Y, Tian D, Yang X, Liu S, Ni L, Zhang Z, Song S, Tian Z (2023) SoyOmics: A deeply integrated database on soybean multi-omics. *Mol Plant* 16(5):794–797. <https://doi.org/10.1016/j.molp.2023.03.011>
- Liu Q, Zuo S, Peng S et al (2024) Development of machine learning methods for accurate prediction of plant disease resistance. *Engineering*. <https://doi.org/10.1016/j.eng.2024.03.014>
- Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T (2017) Transcriptomics technologies. *PLoS Comput Biol* 13:e1005457. <https://doi.org/10.1371/journal.pcbi.1005457>
- Luo Y, Zhao C, Chen F (2024) Multi-omics research: principles and challenges in integrated analysis. *BioDesign Res*. <https://doi.org/10.34133/bdr.0059>
- Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sunderland, MA: Sinauer
- Mallick H, Porwal A, Saha S, Basak P, Svetnik V, Paul E (2024) An integrated Bayesian framework for multi-omics prediction and classification. *Stat Med* 43:983–1002. <https://doi.org/10.1002/sim.9953>
- Mannur DM, Babbar A, Thudi M, Sabbavarapu MM, Roorkiwal M, Yeri SB, Bansal VP, Jayalakshmi SK, Singh Yadav S, Rathore A, Chamarthi SK, Mallikarjuna BP, Gaur PM, Varshney RK (2019) Super Annigeri 1 and improved JG 74: two Fusarium wilt-resistant introgression lines developed using marker-assisted backcrossing approach in chickpea (*Cicer arietinum* L.). *Mole Breed: New Strat Plant Improve* 39(1):2. <https://doi.org/10.1007/s11032-018-0908-9>
- Marsh JJ, Hu H, Gill M, Batley J, Edwards D (2021) Crop breeding for a changing climate: integrating phenomics and genomics with bioinformatics. *Theor Appl Genet* 134(6):1677–1690. <https://doi.org/10.1007/s00122-021-03820-3>
- Marsh JJ, Petereit J, Johnston BA, Bayer PE, Tay Fernandez CG, Al-Mamun HA, Batley J, Edwards D (2023) crosshap: R package for local haplotype visualization for trait association analysis. *Bioinformatics* 39(8):btad518. <https://doi.org/10.1093/bioinformatics/btad518>
- McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 5(4):115–133. <https://doi.org/10.1007/BF02478259>
- Meng C, Zelezniak OA, Thallinger GG et al (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinformatics* 17:628–641. <https://doi.org/10.1093/bib/bbv108>
- Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4):1819–1829. <https://doi.org/10.1093/genetics/157.4.1819>
- Meyer JG (2021) Deep learning neural network tools for proteomics. *Cell Rep Methods* 1:100003. <https://doi.org/10.1016/j.crmeth.2021.100003>
- Michael TP, Jackson S (2013) The First 50 Plant Genomes. *Plant Genome* 6(2):1–7. <https://doi.org/10.3835/plantgenome2013.03.0001in>
- Michel S, Wagner C, Nosenko T et al (2021) Merging genomics and transcriptomics for predicting fusarium head blight resistance in wheat. *Genes* 12. <https://doi.org/10.3390/genes12010114>
- Montesinos-López OA, Chavira-Flores M, Kiasiantini C-H, Saint Pierre C, Li H, Fritsche-Neto R, Al-Nowibet K, Montesinos-López A, Crossa J (2024) A review of multimodal deep learning methods for genomic-enabled prediction in plant breeding. *Genetics*. <https://doi.org/10.1093/genetics/iyae161>
- Mukherjee A, Abraham S, Singh A, Balaji S, Mukunthan KS (2024) From data to cure: a comprehensive exploration of multi-omics data analysis for targeted therapies. *Mol Biotechnol*. <https://doi.org/10.1007/s12033-024-01133-6>
- Mustafa G, Komatsu S (2021) Plant proteomic research for improvement of food crops under stresses: a review. *Mol Omics* 17(6):860–880. <https://doi.org/10.1039/d1mo00151e>
- Nazari L, Aslan MF, Sabanci K, Ropelewska E (2023) Integrated transcriptomic meta-analysis and comparative artificial intelligence models in maize under biotic stress. *Sci Rep* 13:15899. <https://doi.org/10.1038/s41598-023-42984-4>
- Neik TX, Amas J, Barbetti M, Edwards D, Batley J (2020) Understanding host-pathogen interactions in brassica napus in the Omics Era. *Plants* 9. <https://doi.org/10.3390/plants9101336>
- Nelson R, Wiesner-Hanks T, Wissner R, Balint-Kurti P (2018) Navigating complexity to breed disease-resistant crops. *Nat Rev Genet* 19(1):21–33. <https://doi.org/10.1038/nrg.2017.82>
- Niks RE, Qi X, Marcel TC (2015) Quantitative resistance to biotrophic filamentous plant pathogens: concepts, misconceptions, and mechanisms. *Annu Rev Phytopathol* 53:445–470. <https://doi.org/10.1146/annurev-phyto-080614-115928>
- Norman A, Taylor J, Edwards J, Kuchel H (2018) Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. *G3 (Bethesda)* 8:2889–2899. <https://doi.org/10.1534/g3.118.200311>
- Oerke EC (2006) Crop losses to pests. *J Agric Sci* 144(01):31. <https://doi.org/10.1017/S0021859605005708>
- Padder BA, Kamfwa K, Awale HE, Kelly JD (2016) Transcriptome Profiling of the *Phaseolus vulgaris* - *Colletotrichum lindemuthianum* Pathosystem. *PLoS ONE* 11(11):e0165823. <https://doi.org/10.1371/journal.pone.0165823>
- Pan SJ, Yang Q (2010) A survey on transfer learning. *IEEE Trans Knowl Data Eng* 22(10):1345–1359. <https://doi.org/10.1109/TKDE.2009.191>
- Panahi B, Hamid R, Mohammad Zadeh Jalaly H (2025) Deciphering plant transcriptomes: leveraging machine learning for deeper insights. *Current Plant Biol* 41:100432. <https://doi.org/10.1016/j.cpb.2024.100432>
- Pasat BA, Pilalis E, Mnich K et al (2024) MultiOmicsIntegrator: a nextflow pipeline for integrated omics analyses. *Bioinform Adv* 4:vbae175. <https://doi.org/10.1093/bioadv/vbae175>
- Pavan Kumar BK, Kanakala S, Malathi VG, Gopal P, Usha R (2017) Transcriptomic and proteomic analysis of yellow mosaic diseased soybean. *J Plant Biochem Biotechnol* 26(2):224–234. <https://doi.org/10.1007/s13562-016-0385-3>
- Pérez-Rodríguez P, Gianola D, González-Camacho JM, Crossa J, Manès Y, Dreisigacker S (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3 (Bethesda)* 2(12):1595–1605. <https://doi.org/10.1534/g3.112.003665>
- Peyraud R, Dubiella U, Barbacci A, Genin S, Raffaele S, Roby D (2017) Advances on plant-pathogen interactions from molecular toward systems biology perspectives. *Plant J* 90(4):720–737. <https://doi.org/10.1111/tjp.13429>
- Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A (2021) Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J* 19:3735–3746. <https://doi.org/10.1016/j.csbj.2021.06.030>
- Plotkin JB (2010) Transcriptional regulation is only half the story. *Mol Syst Biol* 6:406. <https://doi.org/10.1038/msb.2010.63>
- Poland JA, Balint-Kurti PJ, Wissner RJ, Pratt RC, Nelson RJ (2009) Shades of gray: the world of quantitative disease resistance. *Trends Plant Sci* 14(1):21–29. <https://doi.org/10.1016/j.tplants.2008.10.006>
- Ponnala L, Wang Y, Sun Q, van Wijk KJ (2014) Correlation of mRNA and protein abundance in the developing maize leaf. *Plant J* 78(3):424–440. <https://doi.org/10.1111/tjp.12482>
- Portilla AE, Mayor-Duran VM, Buendia HF, Blair MW, Cichy K, Raatz B (2022) Climbing bean breeding for disease resistance

- and grain quality traits. *Legume Sci* 4(2):e122. <https://doi.org/10.1002/leg3.122>
- Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM (2022) A review of feature selection methods for machine learning-based disease risk prediction. *Front Bioinform* 2:927312. <https://doi.org/10.3389/fbinf.2022.927312>
- Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575. <https://doi.org/10.1086/519795>
- Qureshi IA, Mehler MF (2011) Advances in epigenetics and epigenomics for neurodegenerative diseases. *Curr Neurol Neurosci Rep* 11:464–473. <https://doi.org/10.1007/s11910-011-0210-2>
- Ramakrishnan P, Manivannan N, Mothilal A, Mahaingam L, Prabhu R, Gopikrishnan P (2020) Marker assisted introgression of QTL region to improve late leaf spot and rust resistance in elite and popular variety of groundnut (*Arachis hypogaea* L.) cv TMV 2. *Austral Plant Pathol* 49(5):505–513. <https://doi.org/10.1007/s13313-020-00721-9>
- Ramirez-Prado JS, Piquerez SJM, Bendahmane A, Hirt H, Raynaud C, Benhamed M (2018) Modify the histone to win the battle: chromatin dynamics in plant-pathogen interactions. *Front Plant Sci* 9:355. <https://doi.org/10.3389/fpls.2018.00355>
- Ray DK, Mueller ND, West PC, Foley JA (2013) Yield trends are insufficient to double global crop production by 2050. *PLoS ONE* 8(6):e66428. <https://doi.org/10.1371/journal.pone.0066428>
- Raymond B, Bouwman AC, Schrooten C, Houwing-Duistermaat J, Veerkamp RF (2018) Utility of whole-genome sequence data for across-breed genomic prediction. *Genet Sel Evol* 50:27. <https://doi.org/10.1186/s12711-018-0396-8>
- Reel PS, Reel S, Pearson E, Trucco E, Jefferson E (2021) Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol Adv* 49:107739. <https://doi.org/10.1016/j.biotechadv.2021.107739>
- Rehman HM, Cooper JW, Lam H-M, Yang SH (2019) Legume biofortification is an underexploited strategy for combatting hidden hunger. *Plant Cell Environ* 42(1):52–70. <https://doi.org/10.1111/pce.13368>
- Ren Y, Wu C, Zhou H, Hu X (2024) Dual-extraction modeling: a multi-modal deep-learning architecture for phenotypic prediction and functional gene mining of complex traits. *Plant Commun* 5:101002. <https://doi.org/10.1016/j.xplc.2024.101002>
- Rimbaud L, Fabre F, Papaix J, Moury B, Lannou C, Barrett LG, Thrall PH (2021) Models of plant resistance deployment. *Annu Rev Phytopathol* 59:125–152. <https://doi.org/10.1146/annurev-phyto-020620-122134>
- Ristaino JB, Anderson PK, Bebbler DP, Brauman KA, Cunniffe NJ, Fedoroff NV, Finegold C, Garrett KA, Gilligan CA, Jones CM, Martin MD, MacDonald GK, Neenan P, Records A, Schmale DG, Tateosian L, Wei Q (2021) The persistent threat of emerging plant disease pandemics to global food security. *Proc Natl Acad Sci USA* 118(23):e2022239118. <https://doi.org/10.1073/pnas.2022239118>
- Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D (2015) Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* 16(2):85–97. <https://doi.org/10.1038/nrg3868>
- Roychowdhury R, Das SP, Gupta A, Parihar P, Chandrasekhar K, Sarker U, Kumar A, Ramrao DP, Sudhakar C (2023) Multi-omics pipeline and omics-integration approach to decipher plant's abiotic stress tolerance responses. *Genes* 14(6):1281. <https://doi.org/10.3390/genes14061281>
- Rutkoski J, Benson J, Jia Y, Brown-Guedira G, Jannink J-L, Sorrells M (2012) Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. *Plant Genome* 5:51–61. <https://doi.org/10.3835/plantgenome2012.02.0001>
- Sahu A, Psaroudakis D, Rolletschek H, et al (2025) panomiX: Investigating Mechanisms Of Trait Emergence Through Multi-Omics Data Integration. *BioRxiv*. <https://doi.org/10.1101/2025.04.11.648356>
- Samuel AL (1959) Some studies in machine learning using the game of checkers. *IBM J Res & Dev* 3(3):210–229. <https://doi.org/10.1147/rd.33.0210>
- Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A (2019) The global burden of pathogens and pests on major food crops. *Nat Ecol Evol* 3(3):430–439. <https://doi.org/10.1038/s41559-018-0793-y>
- Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE (2018) Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. *Genetics* 208(4):1373–1385. <https://doi.org/10.1534/genetics.117.300374>
- Scossa F, Alseikh S, Fernie AR (2021) Integrating multi-omics data for crop improvement. *J Plant Physiol* 257:153352. <https://doi.org/10.1016/j.jplph.2020.153352>
- Sereshki S, Lee N, Omirou M, Fasoula D, Lonardi S (2023) On the prediction of non-CG DNA methylation using machine learning. *NAR Genom Bioinform* 5:lqad045. <https://doi.org/10.1093/nargab/lqad045>
- Shi X, Chen Q, Liu S, Wang J, Peng D, Kong L (2021) Combining targeted metabolite analyses and transcriptomics to reveal the specific chemical composition and associated genes in the incompatible soybean variety PI437654 infected with soybean cyst nematode HG1.2.3.5.7. *BMC Plant Biol* 21(1):217. <https://doi.org/10.1186/s12870-021-02998-4>
- Shoaib M, Shah B, Sayed N, Ali F, Ullah R, Hussain I (2023) Deep learning for plant bioinformatics: an explainable gradient-based approach for disease detection. *Front Plant Sci* 14:1283235. <https://doi.org/10.3389/fpls.2023.1283235>
- Shomorony I, Cirulli ET, Huang L, Napier LA, Heister RR, Hicks M, Cohen IV, Yu H-C, Swisher CL, Schenker-Ahmed NM, Li W, Nelson KE, Brar P, Kahn AM, Spector TD, Caskey CT, Venter JC, Karow DS, Kirkness EF, Shah N (2020) An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. *Genome Med* 12(1):7. <https://doi.org/10.1186/s13073-019-0705-z>
- Sia J, Zhang W, Cheng M, Bogdan P, Cook DE (2025) Machine learning-based identification of general transcriptional predictors for plant disease. *New Phytol* 245:785–806. <https://doi.org/10.1111/nph.20264>
- Singh BK, Delgado-Baquerizo M, Egidi E, Guirado E, Leach JE, Liu H, Trivedi P (2023a) Climate change impacts on plant pathogens, food security and paths forward. *Nat Rev Microbiol* 21(10):640–656. <https://doi.org/10.1038/s41579-023-00900-7>
- Singh RN, Krishnan P, Bharadwaj C, Das B (2023b) Improving prediction of chickpea wilt severity using machine learning coupled with model combination techniques under field conditions. *Ecol Inform* 73:101933. <https://doi.org/10.1016/j.ecoinf.2022.101933>
- Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE (2008) Genomic selection using different marker types and densities. *J Anim Sci* 86:2447–2454. <https://doi.org/10.2527/jas.2007-0010>
- Song M, Greenbaum J, Luttrell J, Zhou W, Wu C, Shen H, Gong P, Zhang C, Deng H-W (2020) A review of integrative imputation for multi-omics datasets. *Front Genet* 11:570255. <https://doi.org/10.3389/fgene.2020.570255>
- Sperschneider J (2020) Machine learning in plant-pathogen interactions: empowering biological predictions from field scale to genome scale. *New Phytol* 228(1):35–41. <https://doi.org/10.1111/nph.15771>

- Strange RN, Scott PR (2005) Plant disease: a threat to global food security. *Annu Rev Phytopathol* 43:83–116. <https://doi.org/10.1146/annurev.phyto.43.113004.133839>
- St Clair DA (2010) Quantitative disease resistance and quantitative resistance Loci in breeding. *Annu Rev Phytopathol* 48:247–268. <https://doi.org/10.1146/annurev-phyto-080508-081904>
- Stuthman DD, Leonard KJ, Miller-Garvin J (2007) Breeding Crops for Durable Resistance to Disease. Elsevier, pp 319–367. [https://doi.org/10.1016/S0065-2113\(07\)95004-X](https://doi.org/10.1016/S0065-2113(07)95004-X)
- Subramanian I, Verma S, Kumar S, Jere A, Anamika K (2020) Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights* 14:1177932219899051. <https://doi.org/10.1177/1177932219899051>
- Tang X, Zheng P, Li X, Wu H, Wei D-Q, Liu Y, Huang G (2022) Deep6mAPred: a CNN and Bi-LSTM-based deep learning method for predicting DNA N6-methyladenosine sites across plant species. *Methods* 204:142–150. <https://doi.org/10.1016/j.ymeth.2022.04.011>
- Taylor MB, Ehrenreich IM (2015) Higher-order genetic interactions and their contribution to complex traits. *Trends Genet* 31(1):34–40. <https://doi.org/10.1016/j.tig.2014.09.001>
- Tay Fernandez CG, Nestor BJ, Danilevicz MF, Gill M, Petereit J, Bayer PE, Finnegan PM, Batley J, Edwards D (2022) Pangenomes as a resource to accelerate breeding of under-utilised crop species. *Int J Mol Sci* 23(5):2671. <https://doi.org/10.3390/ijms23052671>
- Tilman D, Cassman KG, Matson PA, Naylor R, Polasky S (2002) Agricultural sustainability and intensive production practices. *Nature* 418(6898):671–677. <https://doi.org/10.1038/nature01014>
- Tirnaz S, Batley J (2019) DNA methylation: toward crop disease resistance improvement. *Trends Plant Sci* 24(12):1137–1150. <https://doi.org/10.1016/j.tplants.2019.08.007>
- Tiwari N, Barpete S, Kumar T et al (2023) Disease resistance an essential for better adaptability and production of faba bean in India (*Vicia faba* L.). In: Jha UC, Nayyar H, Sharma KD et al (eds) Diseases in legume crops: next generation breeding approaches for resistant legume crops. Springer Nature Singapore, Singapore, pp 175–193
- Tong J, Tarekegn ZT, Jambuthenne D, Alahmad S, Periyannan S, Hickey L, Dinglasan E, Hayes B (2024) Stacking beneficial haplotypes from the Vavilov wheat collection to accelerate breeding for multiple disease resistance. *Theor Appl Genet* 137(12):274. <https://doi.org/10.1007/s00122-024-04784-w>
- Tse OYO, Jiang P, Cheng SH, Peng W, Shang H, Wong J, Chan SL, Poon LCY, Leung TY, Chan KCA, Chiu RWK, Lo YMD (2021) Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proc Natl Acad Sci USA* 118(5):e2019768118. <https://doi.org/10.1073/pnas.2019768118>
- United Nations Department of Economic and Social Affairs, Population Division (2022) World Population Prospects 2022: United Nations. [https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022\\_summary\\_of\\_results.pdf](https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf)
- Upadhyaya SR, Danilevicz MF, Dolatabadian A, Neik TX, Zhang F, Al-Mamun HA, Bennamoun M, Batley J, Edwards D (2024) Genomics-based plant disease resistance prediction using machine learning. *Plant Pathol*. <https://doi.org/10.1111/ppa.13988>
- Upasani ML, Limaye BM, Gurjar GS, Kasibhatla SM, Joshi RR, Kadoo NY, Gupta VS (2017) Chickpea-*Fusarium oxysporum* interaction transcriptome reveals differential modulation of plant defense strategies. *Sci Rep* 7(1):7746. <https://doi.org/10.1038/s41598-017-07114-x>
- Vahabi N, Michailidis G (2022) Unsupervised multi-omics data integration methods: a comprehensive review. *Front Genet* 13:854752. <https://doi.org/10.3389/fgene.2022.854752>
- van Esse HP, Reuber TL, van der Does D (2020) Genetic modification to improve disease resistance in crops. *New Phytol* 225(1):70–86. <https://doi.org/10.1111/nph.15967>
- Valous NA, Popp F, Zörnig I, Jäger D, Charoentong P (2024) Graph machine learning for integrated multi-omics analysis. *Br J Cancer* 131:205–211. <https://doi.org/10.1038/s41416-024-02706-7>
- Varona L, Legarra A, Toro MA, Vitezica ZG (2018) Non-additive effects in genomic selection. *Front Genet* 9:78. <https://doi.org/10.3389/fgene.2018.00078>
- Vogel C, Marcotte EM (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* 13(4):227–232. <https://doi.org/10.1038/nrg3185>
- Wallace JG, Rodgers-Melnick E, Buckler ES (2018) On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Annu Rev Genet* 52:421–444. <https://doi.org/10.1146/annurev-genet-120116-024846>
- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10:57–63. <https://doi.org/10.1038/nrg2484>
- Wang S, Wei J, Li R, Qu H, Chater JM, Ma R, Li Y, Xie W, Jia Z (2019) Identification of optimal prediction models using multi-omic data for selecting hybrid rice. *Heredity* 123(3):395–406. <https://doi.org/10.1038/s41437-019-0210-6>
- Wang K, Abid MA, Rasheed A, Crossa J, Hearne S, Li H (2023) DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Mol Plant* 16:279–293. <https://doi.org/10.1016/j.molp.2022.11.004>
- Wang P, Lehti-Shiu MD, Lotreck S, Segura Abá K, Krysan PJ, Shiu S-H (2024a) Prediction of plant complex traits via integration of multi-omics data. *Nat Commun* 15(1):6856. <https://doi.org/10.1038/s41467-024-50701-6>
- Wang X-Y, Ren C-X, Fan Q-W, Xu Y-P, Wang L-W, Mao Z-L, Cai X-Z (2024b) Integrated assays of genome-wide association study, multi-omics co-localization, and machine learning associated calcium signaling genes with oilseed rape resistance to *Sclerotinia sclerotiorum*. *Int J Mol Sci* 25(13):6932. <https://doi.org/10.3390/ijms25136932>
- Wani ZA, Ashraf N (2018) Transcriptomic studies revealing enigma of plant-pathogen interaction. In: Singh A, Singh IK (eds) Molecular aspects of plant-pathogen interaction. Springer Singapore, Singapore, pp 219–238
- Weber SE, Frisch M, Snowdon RJ, Voss-Fels KP (2023) Haplotype blocks for genomic prediction: a comparative evaluation in multiple crop datasets. *Front Plant Sci* 14:1217589. <https://doi.org/10.3389/fpls.2023.1217589>
- Weinstock L, Schambach J, Fisher A, Kunstadt C, Lee E, Koning E, Morrell W, Mays W, Davis W, Krishnakumar R (2024) A hybrid machine learning model for predicting gene expression from epigenetics across fungal species. *BioRxiv*. <https://doi.org/10.1101/2024.12.12.628183>
- Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. *J Big Data* 3(1):9. <https://doi.org/10.1186/s40537-016-0043-6>
- Weissenbach J (2016) The rise of genomics. *C R Biol* 339:231–239. <https://doi.org/10.1016/j.crvi.2016.05.002>
- Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453(7199):1239–1243. <https://doi.org/10.1038/nature07002>
- Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3:160018. <https://doi.org/10.1038/sdata.2016.18>

- Wu P-Y, Stich B, Weisweiler M, Shrestha A, Erban A, Westhoff P, Inghelant DV (2022) Improvement of prediction ability by integrating multi-omic datasets in barley. *BMC Genom* 23(1):200. <https://doi.org/10.1186/s12864-022-08337-7>
- Wu C, Luo J, Xiao Y (2024) Multi-omics assists genomic prediction of maize yield with machine learning approaches. *Mol Breed* 44(2):14. <https://doi.org/10.1007/s11032-024-01454-z>
- Xie H, Wang L, Qian Y, Ding Y, Guo F (2025) Methyl-GP: accurate generic DNA methylation prediction based on a language model and representation learning. *Nucleic Acids Res* 53. <https://doi.org/10.1093/nar/gkaf223>
- Xu Y, Li P, Zou C, Lu Y, Xie C, Zhang X, Prasanna BM, Olsen MS (2017) Enhancing genetic gain in the era of molecular breeding. *J Exp Bot* 68(11):2641–2666. <https://doi.org/10.1093/jxb/erx135>
- Yang W, Feng H, Zhang X, Zhang J, Doonan JH, Batchelor WD, Xiong L, Yan J (2020) Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol Plant* 13(2):187–214. <https://doi.org/10.1016/j.molp.2020.01.008>
- Yang Y, Saand MA, Huang L, Abdelaal WB, Zhang J, Wu Y, Li J, Sirahi MH, Wang F (2021) Applications of multi-omics technologies for crop improvement. *Front Plant Sci* 12:563953. <https://doi.org/10.3389/fpls.2021.563953>
- Yang Z, Luo C, Pei X, Wang S, Huang Y, Li J, Liu B, Kong F, Yang Q-Y, Fang C (2024) SoyMD: a platform combining multi-omics data with various tools for soybean research and breeding. *Nucleic Acids Res* 52:D1639–D1650. <https://doi.org/10.1093/nar/gkad786>
- Yin L, Zhang H, Tang Z et al (2021) rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genom Proteomics Bioinform* 19:619–628. <https://doi.org/10.1016/j.gpb.2020.10.007>
- Zagorščak M, Abdelhakim L, Rodriguez-Granados NY et al (2025) Integration of multi-omics data and deep phenotyping provides insights into responses to single and combined abiotic stress in potato. *Plant Physiol* 197. <https://doi.org/10.1093/plphys/kiaf126>
- Zeng W, Gautam A, Huson DH (2022) MuLan-Methyl-multiple transformer-based language models for accurate DNA methylation prediction. *Gigascience* 12. <https://doi.org/10.1093/gigascience/giad054>
- Zeng S, Adusumilli T, Awan SZ, Immadi MS, Xu D, Joshi T (2024) G2PDeep-v2: a web-based deep-learning framework for phenotype prediction and biomarker discovery using multi-omics data. *BioRxiv*. <https://doi.org/10.1101/2024.09.10.612292>
- Zhang X, Zhang J, He X, Wang Y, Ma X, Yin D (2017) Genome-wide association study of major agronomic traits related to domestication in peanut. *Front Plant Sci* 8:1611. <https://doi.org/10.3389/fpls.2017.01611>
- Zhang H, Yin L, Wang M, Yuan X, Liu X (2019) Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Front Genet* 10:189. <https://doi.org/10.3389/fgene.2019.00189>
- Zhang F, Yu S, Wu L et al (2020a) Phenotype Classification using proteome data in a data-independent acquisition tensor format. *J Am Soc Mass Spectrom* 31:2296–2304. <https://doi.org/10.1021/jasms.0c00254>
- Zhang X-Q, Bai L, Sun H-B, Yang C, Cai B-Y (2020b) Transcriptomic and Proteomic Analysis Revealed the Effect of *Funnelformis* mosseae in Soybean Roots Differential Expression Genes and Proteins. *J Proteome Res* 19(9):3631–3643. <https://doi.org/10.1021/acs.jproteome.0c00017>
- Zhang R, Zhang C, Yu C, Dong J, Hu J (2022) Integration of multi-omics technologies for crop improvement: Status and prospects. *Front Bioinform* 2:1027457. <https://doi.org/10.3389/fbinf.2022.1027457>
- Zhao C, Zhang Y, Du J, Guo X, Wen W, Gu S, Wang J, Fan J (2019) Crop phenomics: current status and perspectives. *Front Plant Sci* 10:714. <https://doi.org/10.3389/fpls.2019.00714>
- Zhao E, Dong L, Zhao H, Zhang H, Zhang T, Yuan S, Jiao J, Chen K, Sheng J, Yang H, Wang P, Li G, Qin Q (2023) A relationship prediction method for magnaporthe oryzae-rice multi-omics data based on WGCNA and graph autoencoder. *J Fungi (Basel)* 9(10):1007. <https://doi.org/10.3390/jof9101007>
- Zhao L, Tang P, Luo J, Liu J, Peng X, Shen M, Wang C, Zhao J, Zhou D, Fan Z, Chen Y, Wang R, Tang X, Xu Z, Liu Q (2025) Genomic prediction with NetGP based on gene network and multi-omics data in plants. *Plant Biotechnol J*. <https://doi.org/10.1111/pbi.14577>
- Zhi P, Chang C (2021) Exploiting epigenetic variations for crop disease resistance improvement. *Front Plant Sci* 12:692328. <https://doi.org/10.3389/fpls.2021.692328>
- Zhu H, Bilgin M, Snyder M (2003) Proteomics. *Annu Rev Biochem* 72:783–812. <https://doi.org/10.1146/annurev.biochem.72.121801.161511>
- Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H, Tu Z, Brem RB, Bumgarner RE, Schadt EE (2012) Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biol* 10(4):e1001301. <https://doi.org/10.1371/journal.pbio.1001301>
- Zhu Q-H, Shan W-X, Ayliffe MA, Wang M-B (2016) Epigenetic mechanisms: an emerging player in plant-microbe interactions. *Mol Plant Microbe Interact* 29(3):187–196. <https://doi.org/10.1094/MPMI-08-15-0194-FI>
- Zhu L, Zhou Y, Li X, Zhao J, Guo N, Xing H (2018) Metabolomics analysis of soybean hypocotyls in response to *Phytophthora sojae* infection. *Front Plant Sci* 9:1530. <https://doi.org/10.3389/fpls.2018.01530>
- Zia B, Shi A, Olaoye D et al (2022) Genome-wide association study and genomic prediction for bacterial wilt resistance in common bean (*Phaseolus vulgaris*) core collection. *Front Genet* 13:853114. <https://doi.org/10.3389/fgene.2022.853114>
- Zipfel C (2014) Plant pattern-recognition receptors. *Trends Immunol* 35(7):345–351. <https://doi.org/10.1016/j.it.2014.05.004>
- Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM (2019) Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Inf Fusion* 50:71–91. <https://doi.org/10.1016/j.inffus.2018.09.012>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.