International Journal of Environment, Agriculture and Biotechnology
Vol-11, Issue-2; Mar-Apr, 2026
Peer-Reviewed International Journal
Journal Home Page Available: https://ijeab.com/
Journal DOI: 10.22161/ijeab

AI-Driven Integration of Multi-Omics Data for Gene Function Discovery and Prediction of Complex Crop Phenotypes

Binsy Karattuchali*, Mohammed Fazil Chalattilkalladithodi

*Corresponding author

Received: 16 Mar 2026; Received in revised form: 19 Apr 2026; Accepted: 23 Apr 2026; Available online: 30 Apr 2026

Abstract— *Recent advances in high-throughput genotyping, phenotyping, and multi-omics technologies have generated large, heterogeneous datasets spanning genomics, transcriptomics, proteomics, metabolomics, epigenomics, and phenomics across plant species. Conventional statistical and machine-learning approaches often struggle to integrate these modalities, limiting mechanistic insight and predictive accuracy for complex genotype-environment-phenotype relationships. Artificial intelligence (AI), particularly deep learning and graph-based architectures, provides powerful tools for multi-omics data fusion, regulatory network reconstruction, and prediction of polygenic, environmentally modulated traits. This review synthesizes current applications of AI-driven multi-omics integration in plant science, with a focus on gene function annotation, cellular network inference, and complex crop phenotype prediction. We examine key methodological frameworks including autoencoders, variational generative models, multimodal transformers, graph neural networks, and self-supervised foundation models, highlighting representative case studies in Arabidopsis and major crops such as rice, maize, wheat, and tomato. We critically assess challenges related to data quality, batch effects, domain shift, metadata standardization, model interpretability, and reproducibility, and outline future directions encompassing plant-specific foundation models, pan-omics integration, digital-twin cropping systems, and community-driven benchmarking.*

Keywords— ***Artificial intelligence, Multi-omics integration, Gene function prediction, Crop phenotypes, Graph neural networks, Reproducibility.***

I. INTRODUCTION

Over the past decade, plant science has undergone rapid expansion in biological and agronomic data generation (Marks et al., 2023; Shiu & Lehti-Shiu, 2024). High-quality reference genomes and pangenomes for major crops are now complemented by large-scale resequencing, single-cell and spatial transcriptomics, chromatin accessibility profiling, proteomics, metabolomics, and high-throughput phenotyping using drones, robotics, and imaging platforms (Gill et al., 2022). Despite this data richness, integrating heterogeneous multi-omics datasets to achieve mechanistic insight and predictive accuracy remains a central challenge (Kajrolkar, 2025). Conventional analytical approaches fail to capture nonlinear dependencies, conditional interactions, and cross-modal relationships that underpin complex traits (Jamil et al., 2020), with limitations amplified by technical noise, missing values, batch effects, and strong tissue-, time-, and environment-specific variation typical of plant datasets (Jain et al., 2024).

AI provides a complementary framework for addressing these challenges. Deep learning architectures including autoencoders, GNNs, CNNs, recurrent networks, and transformers learn latent representations capturing nonlinear and hierarchical structure across multiple biological layers (Cembrowska-Lech et al., 2023). By integrating genomics through phenomics, these models enable inference of gene function, reconstruction of regulatory and metabolic networks, and prediction of complex phenotypes under diverse environmental conditions (Montesinos-Lopez et al., 2024; Xu et al., 2025).

This review provides an integrated overview of AI methodologies for multi-omics data fusion in plant science, summarizing key model architectures, representative applications, critical limitations, and future directions.

II. MULTI-OMICS DATA IN PLANT SYSTEMS: OPPORTUNITIES AND INTEGRATION CHALLENGES

The multi-omics paradigm reflects the hierarchical flow of biological information from DNA sequence variation to organism-level phenotypes (Mahmood et al., 2022). No single omics layer can fully explain complex traits such as yield, stress tolerance, or nutritional quality (Balotf et al., 2025). Genomics defines the static genetic blueprint; transcriptomics characterizes dynamic regulatory activity; proteomics reveals functional protein machinery; metabolomics captures biochemical endpoints; epigenomics identifies heritable regulatory modifications; and phenomics quantifies observable outcomes shaped by gene-environment interactions (Amin et al., 2025).

2.1 Omics Layers

Genomics constitutes the foundational layer of multi-omics integration. Chromosome-scale reference genomes are now available for major crops enabled by PacBio HiFi and Oxford Nanopore sequencing (Wang et al., 2023; Hufford et al., 2021). Pangenomes capture intraspecific diversity including structural variation and presence-absence variation absent from linear reference genomes; the maize pangenome comprises over 100,000 non-redundant genes and highlights extensive copy-number variation (Hufford et al., 2021; Jayakodi et al., 2024). However, genomic variation alone rarely explains phenotypic outcomes without functional context, underscoring the necessity of multi-omics integration.

Transcriptomics provides a dynamic view of gene expression. In rice, time-resolved transcriptome analyses during dehydration/rehydration and functional studies of OsCCA1 controlling ABA signaling revealed key transcriptional hubs governing osmotic adjustment (Park & Jeong, 2023; Wei et al., 2022). In maize, Stereo-seq spatial transcriptomics of the developing ear revealed spatially resolved regulatory networks (Y. Wang et al., 2024). Proteomics captures post-transcriptional and post-translational regulation; proximity-labeling techniques such as TurboID enable in vivo detection of transient interactions (Arora et al., 2020). Metabolomics in wheat linked flavonoid accumulation to specific loci via metabolite-based GWAS (J. Chen et al., 2020; D. Ma et al., 2022), although comprehensive metabolite annotation remains challenging (Mahieu & Patti, 2017; Nguyen et al., 2024). Epigenomic mechanisms including DNA methylation, histone modifications, and chromatin accessibility drive phenotypic plasticity; stress-responsive TE activation in maize and heat-induced 3D genome reorganization in Arabidopsis illustrate epigenomic variation as a key regulatory layer (Liang et al., 2021; L. Sun et al., 2020). High-throughput phenomics translates molecular variation into quantitative traits using RGB, hyperspectral, and 3D scanning platforms, with UAV-based imaging successfully predicting soybean yield (Herrero-Huerta et al., 2020).

2.2 Integration Challenges

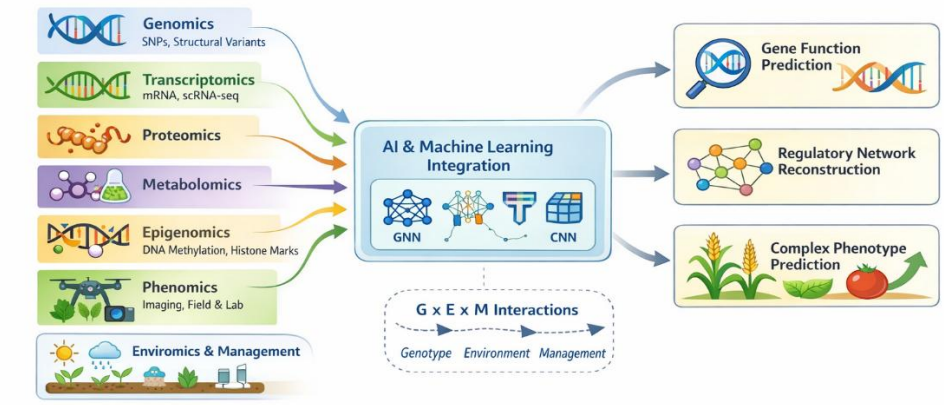

Fig.1. Conceptual overview of AI-driven multi-omics integration in plants.

Heterogeneous omics (genomics, transcriptomics, proteomics, metabolomics, epigenomics, phenomics) and environmental/management data are integrated using advanced AI frameworks (graph neural networks, CNNs, transformers, multimodal deep networks) to predict gene function, reconstruct regulatory networks, and forecast complex crop phenotypes. G×E×M interactions influence outputs, enabling climate-resilient crop design and precision breeding.

Integrating multi-omics datasets poses substantial technical challenges. Omics layers differ markedly in data structure, dimensionality, and preprocessing requirements, necessitating modality-specific normalization (Huang et al., 2017). Batch effects arising from differences in instrumentation frequently obscure biological signals (Leek et al., 2010). Missing data are pervasive because comprehensive profiling across all layers is rarely feasible (Flores et al., 2023). Nonlinear gene-environment interactions, epistasis, pleiotropy, and feedback regulation exceed the representational capacity of traditional linear models. Consequently, classical integration approaches such as PCA and canonical correlation analysis often fail to capture meaningful biological relationships, whereas emerging deep learning frameworks better model nonlinear, hierarchical interactions.

III. AI APPROACHES FOR MULTI-OMICS INTEGRATION

Effective integration of multi-omics data requires computational frameworks capable of jointly modeling high-dimensional, heterogeneous, and frequently incomplete datasets while capturing nonlinear, context-dependent interactions. Deep learning and AI methodologies have consistently outperformed traditional statistical and classical machine learning approaches in predictive accuracy, scalability, and biological relevance (Libbrecht & Noble, 2015; Reel et al., 2021).

3.1 Classical Machine Learning

Before deep learning, classical machine learning methods including random forests, SVMs, LASSO, elastic net, PLS regression, and Bayesian networks were widely applied to QTL mapping, genomic selection, and eQTL studies (Crossa et al., 2010; Rohart et al., 2017). These approaches offer useful interpretability but exhibit substantial limitations for large-scale, multi-omics integration: most assume linear or additive relationships, limiting their capacity to model epistatic interactions, nonlinear regulation, or G×E effects. They scale poorly to the extreme dimensionality of omics datasets (Libbrecht & Noble, 2015; Reel et al., 2021). Classical methods are now primarily used as benchmarking tools within hybrid pipelines.

3.2 Autoencoders and Variational Autoencoders

Autoencoders (AEs) compress diverse inputs into a low-dimensional latent bottleneck via an encoder-decoder architecture (Hinton & Salakhutdinov, 2006). In plants, denoising autoencoders applied to maize RNA-seq and LC-MS profiles under nitrogen limitation identified latent factors linked to nitrogen assimilation while mitigating batch effects (Kannan et al., 2025). Variational autoencoders (VAEs) model latent variables as probability distributions, optimizing the evidence lower bound (ELBO) to balance reconstruction accuracy against prior divergence (Kingma & Welling, 2013), enabling uncertainty quantification, generative sampling, and missing data imputation. Multimodal VAEs integrating SNPs, transcriptomes, and metabolomes in rice achieved approximately 25% higher accuracy than PCA or CCA baselines and imputed up to 30% missing metabolomic features (Zargar et al., 2022). Conditional VAEs integrating time-series scRNA-seq and metabolomics during Arabidopsis floral transition revealed coordinated changes in hormone signaling and secondary metabolism (Jeong et al., 2024).

3.3 Multimodal Deep Neural Networks

Multimodal deep neural networks (MDNNs) integrate heterogeneous data types via modality-specific subnetworks with flexible fusion strategies (Tabakhi et al., 2023). Early fusion concatenates features from all modalities; intermediate (hybrid) fusion combines modality-specific encoders through attention or tensor operations, preserving modality-specific structure while enabling nonlinear cross-modal interactions (Montesinos-Lopez et al., 2023); late fusion aggregates predictions from independently trained unimodal models via ensembling. MDNNs integrating genotypic information, remote-sensing imagery, and environmental variables have demonstrated improved prediction of agronomic traits compared with unimodal approaches (Shamsuddin et al., 2024), and advanced architectures incorporating cross-attention and gating units improve interpretability and dynamic modality weighting (Gong et al., 2023).

3.4 Transformers

Transformers use self-attention mechanisms to model long-range dependencies and contextual relationships, making them well-suited for sequence-based and heterogeneous omics data (Vaswani et al., 2017). Genomic transformers such as DNABERT predict regulatory elements and chromatin accessibility from k-mer tokenized sequences and have been applied to plant genomes for promoter and enhancer identification (Ji et al., 2021). Cross-attention mechanisms link genomic variants with transcriptomic or metabolomic profiles, and diverse data types including SNPs, gene expression tokens, and image patches can be unified into a single sequence for joint modeling (Rao et al., 2021). Vision transformers (ViT) outperform CNNs in feature extraction from high-resolution or hyperspectral plant imagery, and coupling ViT with transcriptomic data via cross-attention reveals spectral-transcriptomic correlations relevant to nutrient use efficiency and stress responses (Dosovitskiy et al., 2020). Practical limitations of

large transformer models include high computational and memory demands, mitigated by model distillation and parameter-efficient fine-tuning (Zaheer et al., 2020).

3.5 Graph Neural Networks

GNNs learn representations on graph-structured data, propagating information across nodes (genes, proteins, metabolites) and edges (PPI, co-expression, metabolic reactions) to integrate heterogeneous multi-omics measurements (Kipf & Welling, 2016; Velickovic et al., 2017). Graph convolutional networks (GCNs) learn node representations by aggregating local neighborhood information; graph attention networks (GATs) assign condition-specific weights to edges, prioritizing regulatory interactions under stress. Heterogeneous GNNs model multiple node and edge types simultaneously, enabling joint learning of TF-gene, protein-protein, and metabolic interactions. In plants, integration of TF binding data, pathway annotations, and single-cell expression profiles has improved GRN reconstruction accuracy compared with classical approaches (Huynh-Thu et al., 2010). Intrinsic attention mechanisms and post-hoc tools such as GNNExplainer reveal influential subgraphs, enabling tracing of signal propagation through regulatory hierarchies (Ying et al., 2019).

3.6 Self-Supervised Learning and Foundation Models

Self-supervised learning (SSL) leverages unlabeled data to learn generalizable representations via surrogate tasks including masked-omics modeling, contrastive learning, and cross-modal prediction, addressing the scarcity of labeled multi-omics datasets in plant biology (LeCun et al., 2015; Liu et al., 2021). Foundation models such as protein language models trained on hundreds of millions of sequences generate embeddings predicting enzyme function and variant effects (Rives et al., 2021; Elnaggar et al., 2021). Vision transformers pretrained on field images enable robust trait extraction and transfer learning for phenotyping (Dosovitskiy et al., 2020). SSL and foundation models provide scalable, flexible frameworks to accelerate gene function discovery and crop trait prediction, particularly when combined with domain adaptation and interpretable fine-tuning.

IV. AI FOR GENE FUNCTION PREDICTION

In plant genomics, approximately 40-60% of genes in major crop genomes remain annotated as hypothetical or uncharacterized proteins (Michael & VanBuren, 2020). Conventional annotation strategies based on sequence similarity searches, domain identification, and co-expression network analyses perform poorly for lineage-specific or rapidly evolving plant genes, exhibiting high false-negative rates and failing to capture conditional regulation, epistatic interactions, or post-transcriptional control (Bolser et al., 2016). AI-driven frameworks can reduce annotation errors and improve predictive accuracy by approximately 20-30% in early benchmarking studies by integrating multi-omics data to uncover nonlinear biological patterns (Jumper et al., 2021).

4.1 Predicting Gene Function Using Integrated Omics

Deep learning-based multimodal frameworks learn shared latent representations from heterogeneous omics datasets including VAEs, multimodal transformers, GNNs, and multimodal autoencoders integrating genomic variation, transcriptomic profiles, epigenomic information, and proteomic measurements. In Arabidopsis, machine-learning frameworks combining genomic, transcriptomic, and methylomic data outperform genome-only models in predicting complex traits and identifying regulatory genes, with models achieving predictive accuracies exceeding 80% for gene expression based on flanking sequence inputs (P. Wang et al., 2024; Peleke et al., 2024). Transfer learning extends applicability to data-limited crops; cross-species learning strategies have improved regulatory network inference in poplar and maize (Mummadi et al., 2025). Multimodal deep learning frameworks effectively capture nonlinear and context-dependent regulatory effects including epistasis and environment-specific gene activity, and can impute missing omics layers such as predicting chromatin states from genotype and expression data.

4.2 Transcription Factor-Target Prediction

TF-target prediction requires integration of sequence-based binding potential with chromatin accessibility, epigenetic modifications, and condition-specific expression dynamics. Deep learning reframes TF-target inference as a heterogeneous graph learning problem: GNNs and GATs represent TFs, target genes, and regulatory elements as nodes, with edges encoding evidence from TF binding assays (DAP-seq, ChIP-seq), ATAC-seq accessibility, cis-motif occurrence, and histone modifications (Song et al., 2016). In Arabidopsis, large-scale DAP-seq datasets have provided a foundation for integrative regulatory network reconstruction, with subsequent studies showing that epigenomic context substantially refines TF-target predictions under stress conditions (Maher et al., 2017). Transformer-based architectures further advance TF-target prediction by modeling long-range dependencies, and integrative learning-based frameworks consistently outperform motif-only tools such as FIMO when evaluated against experimentally validated TF-target interactions (Fornes et al., 2019).

4.3 Enzyme Function Prediction and Case Studies

Plant enzymes frequently exhibit lineage-specific expansion, neo-functionalization, and relaxed substrate specificity, limiting homology-based annotation (Schlapfer et al., 2017). Protein language models (PLMs) such as ESM-1b and ProtBERT, pretrained on hundreds of millions of sequences, capture evolutionary constraints and functional motifs without explicit sequence alignment (Rives et al., 2021; Elnaggar et al., 2021). PLM-derived embeddings integrated with transcriptomic and metabolomic data enabled reconstruction of large portions of tomato flavonoid and phenylpropanoid biosynthetic pathways (Zhu et al., 2018). GNNs applied to curated metabolic graphs infer EC numbers and pathway positions; in maize, attention-based GNNs resolved isoform-level functions within oxylipin and jasmonate biosynthesis (Zhang et al., 2021). PLM- and GNN-based approaches outperform classical sequence similarity methods by approximately 30-50% for orphan plant enzymes. Key case studies illustrate these advances: in rice, transformer-based architectures prioritized CIPKs involved in osmotic stress adaptation, with CRISPR-Cas9 validation confirming altered proline accumulation and enhanced salt tolerance (Guo et al., 2024); in maize, graph attention networks identified previously uncharacterized lipid transfer proteins associated with kernel filling, confirmed by near-isogenic lines and field trials (Zhang et al., 2021); and in tomato, graph convolutional networks successfully reconstructed anthocyanin and flavonoid biosynthetic branches with enzymatic and regulatory roles validated by qRT-PCR and enzyme activity assays (Zhu et al., 2018).

5. AI for Reconstructing Cellular and Regulatory Networks

Cellular and regulatory networks including GRNs, PPI networks, metabolic pathways, and signaling cascades constitute the molecular circuitry governing plant development, homeostasis, and stress responses. Reconstructing these networks from high-dimensional omics data remains challenging due to nonlinear interactions, feedback regulation, and pervasive crosstalk (Zitnik et al., 2018). Conventional approaches such as WGCNA, ARACNE, and Bayesian networks primarily infer static or pairwise associations. AI approaches including GNNs, RNNs, VAEs, and transformers substantially improve network reconstruction by learning higher-order dependencies and context-specific regulatory logic (Zitnik et al., 2018).

5.1 Network Inference and Dynamic Modeling

GNNs operate directly on graph-structured data with multi-omics features incorporated as node or edge attributes, enabling context-aware regulatory inference (Kipf & Welling, 2016). Heterogeneous GNNs enhance inference by explicitly modeling multiple node and edge types, enabling simultaneous learning of TF-gene, protein-protein, and metabolic interactions. Integration of TF binding data, pathway annotations, and single-cell expression profiles has improved GRN reconstruction accuracy compared with classical approaches such as GENIE3 and Inferelator (Huynh-Thu et al., 2010). GNN-based models support modular network decomposition, facilitating rational prioritization of candidate regulators for genome editing.

Regulatory networks are inherently dynamic, undergoing extensive temporal and condition-dependent rewiring. RNNs including LSTM and GRU architectures infer delayed regulatory effects and nonlinear feedback loops from time-series transcriptomic and metabolomic data (Hochreiter & Schmidhuber, 1997). Neural ODEs model regulatory dynamics as continuous-time systems governed by learnable differential equations, and in plant biology have reconstructed developmental continua from scRNA-seq data, capturing cell-fate transitions and oscillatory modules involved in meristem maintenance and floral transition (Trapnell et al., 2014; R.T.Q. Chen et al., 2018). Pathway crosstalk such as JA-SA antagonism in immunity is modeled by multimodal transformers with multi-head attention, where distinct biological layers are integrated via cross-attention mechanisms (Vaswani et al., 2017). In Arabidopsis, integrative network analyses consistently identify NPR1 and AP2/ERF transcription factors as central hubs mediating JA-SA antagonism and multi-hormone crosstalk (J. Yang et al., 2019).

5.2 Single-Cell Network Reconstruction

Bulk omics analyses average signals across heterogeneous cell populations, obscuring cell-type-specific regulatory programs. ScRNA-seq and scATAC-seq combined with multi-omics integration have enabled reconstruction of cell-type-resolved GRNs at unprecedented resolution (Ryu et al., 2019). VAEs and graph-based models denoise dropout-prone single-cell data, infer developmental trajectories, and reconstruct TF-target interactions at the cell-cluster level. In Arabidopsis roots, large-scale scRNA-seq atlases revealed 149 expression modules along developmental trajectories, including regulators of quiescent center maintenance, metabolic specialization, and auxin-mediated gravitropism (Han et al., 2024). Integration of chromatin accessibility and transcriptomic data further revealed thousands of cell-type-specific regulatory elements (Dorrity et al., 2021). Comparable studies in rice root tips established transferable frameworks for GRN reconstruction in crops (Feng et al., 2022). These AI-driven single-cell network models support precision genome engineering and integration with spatial transcriptomics for three-dimensional regulatory mapping.

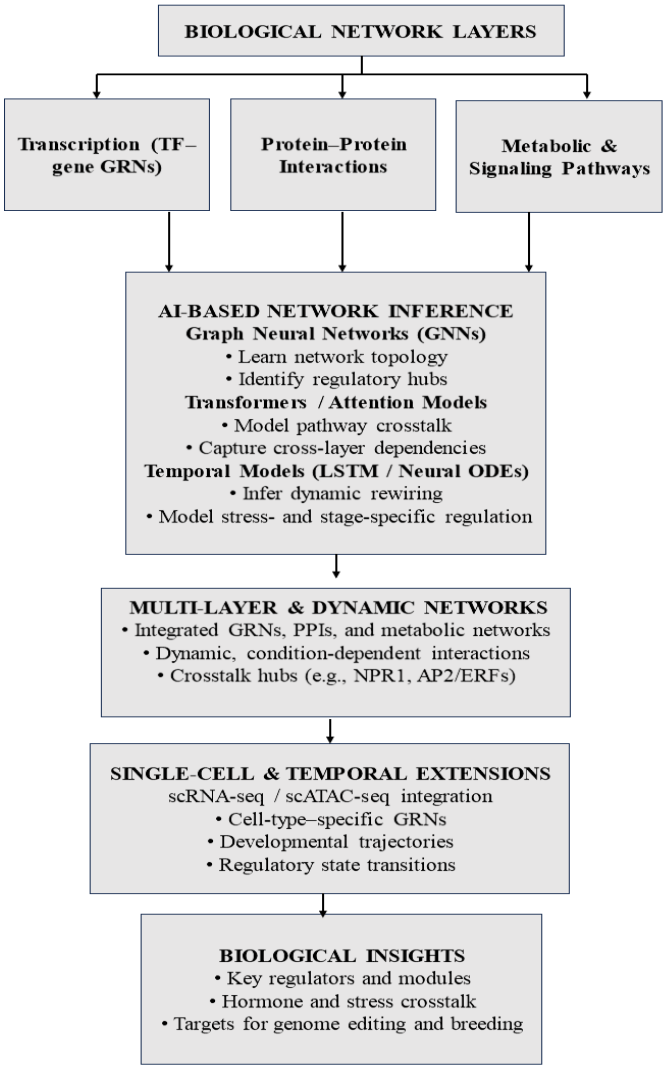

Fig.2. AI frameworks for regulatory network reconstruction and crosstalk modelling

The figure illustrates AI-based integration of transcriptional, protein–protein, and metabolic networks using graph neural networks, transformer models, and temporal learning approaches. These frameworks enable dynamic, cell-type–resolved reconstruction of regulatory networks and the identification of key crosstalk regulators relevant to plant development, stress responses, and crop improvement.

V. AI FOR PREDICTING COMPLEX CROP PHENOTYPES

Complex agronomic phenotypes including yield, phenological development, stress tolerance, nutritional quality, and resource-use efficiency are governed by highly polygenic architectures shaped by nonlinear G×E×M interactions. Conventional GS methods such as RR-BLUP and GBLUP predominantly assume additive genetic effects and linear responses, limiting their capacity to capture epistasis, pleiotropy, rare alleles, and environment-dependent phenotypic plasticity (Crossa et al., 2017). AI approaches including deep neural networks, kernel-based learning, and transformer architectures offer a powerful alternative by learning hierarchical and nonlinear representations from high-dimensional data.

6.1 Genotype-Phenotype and Multi-Omics Prediction

Deep learning enhances genomic prediction by automatically extracting multi-scale features capturing linkage disequilibrium, epistasis, and long-range genetic interactions. CNNs applied to SNP matrices enable detection of local haplotype patterns and non-additive effects beyond the capabilities of linear models (W. Ma et

al., 2018; O.A. Montesinos-Lopez et al., 2021). Pretrained DNA language models such as DNABERT embed sequence variation into contextual representations, and attention-based models have consistently outperformed GBLUP across environments by modeling epistatic interactions, rare variants, and G×E effects in wheat and maize multi-environment trials (Ji et al., 2021; Crossa et al., 2019). MDNNs address the limitations of single-omics genomic selection by integrating multiple omics modalities through intermediate fusion architectures (Subramanian et al., 2020). In sorghum and maize, MDNNs integrating SNPs with transcriptomic and metabolomic profiles significantly improved predictions of drought tolerance, nitrogen-use efficiency, and biomass accumulation (Gu et al., 2025). In tomato, late-fusion ensemble models combining transcriptomic data with hyperspectral phenotyping achieved high classification accuracy for fruit quality traits including carotenoid content and firmness (Xiang et al., 2022). VAEs and generative models improve robustness by imputing missing omics layers, enabling stable phenotype prediction under partial data availability (Way & Greene, 2017).

6.2 Modeling G×E×M Interactions

AI-based models encode enviromics including weather, soil, and management variables as structured, time-resolved inputs. RNNs and LSTM architectures model seasonal and intra-seasonal environmental dynamics (Khaki & Wang, 2019), while transformer architectures integrate genomic markers with sequential environmental data and high-throughput phenotyping, capturing long-range temporal dependencies and context-specific genetic effects such as heat- or drought-induced shifts in QTL expression (O.A. Montesinos-Lopez et al., 2021). These frameworks support climate-resilient breeding by enabling in silico evaluation of genotype performance under future climate scenarios. Key examples include: multi-modal AI frameworks combining dense SNP markers, UAV-derived multispectral indices, and meteorological data substantially improving wheat yield predictions across multi-environment trials (Merrick et al., 2022); integrated transcriptomic-metabolomic models predicting tomato fruit quality traits including carotenoid, sugar, and organic acid content (Zhu et al., 2018; Xiang et al., 2022); and time-resolved multi-omics integration in rice distinguishing tolerant and susceptible genotypes under drought while identifying key dehydration-responsive transcription factor networks (V.H. Nguyen et al., 2023). AI-enabled G×E×M modeling shortens breeding cycles, improves early-generation selection accuracy, and facilitates climate-smart cultivar development.

VI. DATA QUALITY, REPRODUCIBILITY, AND STANDARDIZATION CHALLENGES

The successful translation of AI-driven multi-omics approaches into production environments is critically dependent on data quality, rigorous experimental design, and robust computational reproducibility. Plant omics datasets are frequently compromised by technical noise, batch effects, incomplete metadata, and platform heterogeneity, collectively inflating spurious correlations and restricting generalizability across laboratories, species, and environments.

7.1 Metadata, FAIR Principles, and Benchmarking

Reproducibility in omics research fundamentally depends on comprehensive and standardized metadata. A substantial fraction of publicly available plant omics datasets lack critical contextual information including growth conditions, stress intensity, sampling time points, and soil properties (Rajesh et al., 2021). The FAIR (Findable, Accessible, Interoperable, Reusable) data principles promote structured metadata annotated with community-recognized ontologies. In plant phenomics, the MIAPPE standard specifies more than 120 metadata fields; despite widespread endorsement, compliance remains below 30% in major repositories such as GEO (Rajesh et al., 2021). Frameworks such as ISA-Tab and the Plant Ontology support harmonized annotation across studies, though inconsistent terminology continues to confound meta-analyses (Dumschott et al., 2023). Emerging FAIR assessment tools including FAIRshake automate the evaluation of FAIR metrics and identify datasets unsuitable for AI model training (Cuellar et al., 2022). In contrast to computer vision and NLP, plant AI research lacks widely adopted benchmarks for multi-omics applications (Amin et al., 2025). Community-driven initiatives such as PlantGenie and OpenPlantML have proposed benchmark tasks encompassing GRN inference, yield prediction, and stress classification. A major limitation is the scarcity of held-out test datasets with experimentally validated ground truth, increasing the risk of overfitting. Federated benchmarking approaches, in which institutions share models or evaluation metrics without exchanging raw data, are emerging as a promising solution (Burankova et al., 2024).

7.2 Domain Shift and Explainability

AI models trained under controlled experimental conditions frequently exhibit degraded performance in heterogeneous field environments due to domain shift. Batch effects can exceed true biological signal (Leek et al., 2010), and correction methods such as ComBat, Harmony, and surrogate variable analysis are effective when batch variables are accurately annotated (Johnson et al., 2006; Korsunsky et al., 2019), but their effectiveness diminishes

when metadata are incomplete or technical covariates are confounded with biological factors. Domain adaptation strategies including Domain Adversarial Neural Networks and VAEs have demonstrated improved transferability by aligning feature distributions across domains, with unsupervised domain adaptation enhancing plant disease recognition when transferring models from laboratory imagery to field conditions (Wu et al., 2023).

The increasing adoption of complex AI architectures has intensified concerns regarding model transparency, interpretability, and user trust (Tjoa & Guan, 2020). Attention-based architectures offer intrinsic interpretability by assigning weights to input features or graph nodes (Vaswani et al., 2017). Post-hoc explanation techniques such as SHAP and LIME decompose individual predictions into feature-level contributions (Lundberg & Lee, 2017); SHAP analyses of deep learning-based drought tolerance models in rice identified OsNAC transcription factor expression, proline accumulation, and ABA-related features as dominant contributors, consistent with well-characterized stress-response pathways (L. Xu et al., 2024). Pathway-centric explanations from GNNs enable higher-order interpretability by mapping predictions onto known biological pathways, and hybrid modeling strategies combining deep neural feature extractors with transparent downstream models further balance performance and interpretability (Molnar, 2022).

7.3 Ethical and Data-Sharing Issues

Large-scale proprietary breeding datasets are frequently siloed within private corporations, raising concerns regarding knowledge monopolization and unequal access to AI-driven innovation (Rotz et al., 2019). Open science infrastructures including CyVerse, PlantArrayNet, and ELIXIR Plants promote data sharing aligned with FAIR data principles (Goff et al., 2011). Federated learning enables collaborative model training across distributed datasets without direct exchange of raw data, preserving confidentiality while leveraging collective statistical power (Q. Li et al., 2021; T. Li et al., 2020). International ethical AI frameworks from the FAO and OECD emphasize transparency, accountability, inclusiveness, and sustainability in agricultural AI systems, calling for systematic bias audits to address the overrepresentation of temperate, high-input cropping systems in training datasets. Community-governed initiatives such as DivSeek International and IPPN play a pivotal role in establishing shared standards for metadata harmonization, interoperability, and benefit-sharing (Minervini et al., 2015).

Table 1. Quantitative Performance of AI-Based Multi-Omics Integration Methods in Representative Plant Studies

AI Approach	Data Modalities	Plant System / Trait	Quantitative Outcome	Benchmark Comparison	Key Advantage	Representative Studies
Classical ML (RF, LASSO, PLS)	SNPs; transcriptomics; metabolomics	Maize yield, drought tolerance	R2: 0.35-0.55	PCA/linear (R2 < 0.40)	Interpretability; feature selection	Crossa et al., 2010; Rohart et al., 2017
Autoencoders (AE)	Transcriptomics + metabolomics	Maize nitrogen stress	Variance: ~60-70%	PCA: ~40-50%	Noise reduction; batch correction	Du et al., 2021; Yu, 2022
VAE / cVAE	SNPs + RNA-seq + metabolomics	Rice drought resilience	Accuracy +20-25%; imputation ~30%	PCA/CCA	Generative modeling; imputation	Zargar et al., 2022
Multimodal DNNs	SNPs + UAV imagery + weather	Wheat yield prediction	RMSE reduction: 10-25%	Genomic selection	G×E modeling; cross-modal	Montesinos-Lopez et al., 2021
Attention-based MDNNs	Hyperspectral + RNA-seq	Disease classification (tomato)	Accuracy: 85-92%	Unimodal CNNs: 70-80%	Dynamic modality weighting	Li et al., 2023; Gong et al., 2023
Transformers (sequence)	DNA sequence	Regulatory element prediction	AUROC: 0.85-0.92	CNN: ~0.80	Long-range dependencies	Ji et al., 2021; Avsec et al., 2021

Multimodal transformers	Genomics + transcriptomics + images	Phenotype prediction	Accuracy +15-30%	CNN or RNN baselines	Cross-attention; unified tokenization	Rao et al., 2021
GNNs (gene function)	Co-expression + PPI + GWAS	Gene function (maize, rice)	Top-k prioritization +20-35%	Network propagation	Topology-aware learning	Zitnik et al., 2018; Pan et al., 2024
GNNs (phenotype)	Multi-omics graphs	Kernel weight; yield traits	R2: 0.55-0.70	Linear mixed models	Mechanistic interpretability	Zhang et al., 2021
Self-supervised (SSL)	Unlabeled omics data	Multiple crops	Downstream accuracy +10-40%	Training from scratch	Label efficiency; transferability	Rives et al., 2019; Liu et al., 2021
Foundation models	Protein sequences; images; omics	Enzyme function; phenotyping	Zero/few-shot comparable to supervised	Task-specific models	Cross-species generalization	Elnaggar et al., 2020

Abbreviations: AUROC, area under the receiver operating characteristic curve; RMSE, root mean square error; R2, coefficient of determination.

VII. FUTURE PERSPECTIVES

The convergence of AI and multi-omics is reshaping plant biology from a predominantly descriptive discipline into a predictive and hypothesis-generating science (Tardieu et al., 2017; Cooper & Messina, 2021). A major future direction lies in the development of large-scale, self-supervised foundation models trained on vast collections of unlabeled plant genomic, transcriptomic, and proteomic data. Analogous to protein and genomic language models demonstrating strong transferability, plant-focused foundation models are poised to enable cross-species generalization for gene function prediction and variant effect estimation even in data-scarce crops (Rives et al., 2021; Ji et al., 2021). Extending these models to multimodal embeddings jointly encoding genomics, epigenomics, transcriptomics, and phenomics will further support few-shot and zero-shot learning (Washburn et al., 2019).

Future AI systems will need to integrate increasingly diverse biological and environmental modalities including microbiome profiles, volatilomics, hyperspectral and thermal imaging, three-dimensional canopy architecture, root phenomics, and continuous soil and climate sensor data, requiring hierarchical multimodal architectures capable of modeling interactions across spatial, temporal, and organizational scales (Mataigne et al., 2022; Amin et al., 2025). Causal representation learning and hybrid mechanistic-data-driven models represent another critical frontier; incorporating causal inference frameworks and structured biological priors will improve interpretability, robustness, and decision relevance for breeding and agronomy (Y. Wang & Jordan, 2021). Generative AI models could substantially reduce the cost and duration of empirical field trials by simulating plausible molecular and phenotypic responses under future climate conditions (Hu et al., 2024). AI-enabled digital twins that simulate growth, development, and physiological responses across the life cycle by integrating neural ODEs with process-based crop models offer a promising platform for forecasting trait performance and guiding ideotype design (Hammer et al., 2019).

Progress in federated and privacy-preserving learning will be essential for scaling AI across institutional and national boundaries (T. Li et al., 2020; Warnat-Herresthal et al., 2021). The long-term success of AI in plant science will depend on sustained investment in FAIR-compliant data infrastructures, open benchmarks, and interdisciplinary training, with close collaboration among plant biologists, breeders, data scientists, and policymakers to ensure AI systems are interpretable, equitable, and aligned with global food security goals.

VIII. CONCLUSION

AI-driven multi-omics integration is transforming plant science by enabling predictive modeling and mechanistic insights that were previously unattainable. By capturing nonlinear interactions across genomics, transcriptomics, proteomics, metabolomics, epigenomics, and phenomics, AI approaches can infer gene function, reconstruct regulatory networks, and predict complex phenotypes emerging from G×E×M interactions. In crops such as rice, maize, wheat, and tomato, these approaches have accelerated gene discovery including annotation of

thousands of orphan genes with high accuracy (>85%), revealed dynamic regulatory modules and crosstalk hubs, and enhanced genomic prediction for yield, stress tolerance, and quality traits (R2 gains of 15-30% over classical methods). Advanced architectures including GNNs, transformers, and foundation models enable scalable integration of heterogeneous multi-omics data while addressing G×E×M interactions critical for climate-resilient breeding.

However, challenges remain: data quality, batch effects, domain shift, metadata standardization, model interpretability, and ethical data sharing limit the reproducibility, transferability, and equitable application of AI models. Black-box architectures risk generating spurious causal inferences, while proprietary datasets and uneven global resources exacerbate inequities in crop improvement. Overcoming these limitations requires causal validation experiments, interpretable hybrid models, FAIR-compliant data infrastructures, and community-driven benchmarking. With these safeguards, AI-driven multi-omics integration can guide rational crop design, anticipate environmental responses, and accelerate breeding for resilient, high-yielding varieties, ultimately unlocking the molecular basis of complex traits to ensure food security in a rapidly changing environment while fostering equitable and reproducible plant science.

REFERENCES

[1] Allen, D. K., & Young, J. D. (2019). Tracing metabolic flux through time and space with isotope labeling experiments. Current Opinion in Biotechnology, 64, 92-100.

[2] Amin, A., Zaman, W., & Park, S. (2025). Harnessing Multi-Omics and Predictive modeling for Climate-Resilient crop breeding: from genomes to fields. Genes, 16(7), 809.

[3] Arora, D., Abel, N. B., Liu, C., et al. (2020). Establishment of Proximity-Dependent biotinylation approaches in different plant model systems. The Plant Cell, 32(11), 3388-3407.

[4] Avsec, Z., Agarwal, V., Visentin, D., et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods, 18(10), 1196-1203.

[5] Balotf, S., Wilson, R., Hemmati, R., et al. (2025). Plant-pathogen interactions: making the case for multi-omics analysis of complex pathosystems. Stress Biology, 5(1), 66.

[6] Bolser, D. M., Staines, D. M., Perry, E., & Kersey, P. J. (2016). ENSEMBL Plants: integrating tools for visualizing, mining, and analyzing plant genomic data. Methods in Molecular Biology, 1533, 1-31.

[7] Burankova, Y., Klemm, J., Lohmann, J. J. G., et al. (2024). FedRBE: a decentralized privacy-preserving federated batch effect correction tool for omics data based on limma. arXiv. https://doi.org/10.48550/arxiv.2412.05894

[8] Cembrowska-Lech, D., Krzeminska, A., Miller, T., et al. (2023). An Integrated Multi-Omics and Artificial Intelligence framework for advance plant phenotyping in horticulture. Biology, 12(10), 1298.

[9] Chen, J., Hu, X., Shi, T., et al. (2020). Metabolite-based genome-wide association study enables dissection of the flavonoid decoration pathway of wheat kernels. Plant Biotechnology Journal, 18(8), 1722-1735.

[10] Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. (2018). Neural ordinary differential equations. Advances in Neural Information Processing Systems, 31, 6572-6583.

[11] Choudhury, S. D., Samal, A., & Awada, T. (2019). Leveraging image analysis for high-throughput plant phenotyping. Frontiers in Plant Science, 10, 508.

[12] Cooper, M., & Messina, C. D. (2021). Can we harness Enviromics to accelerate crop improvement by integrating breeding and agronomy? Frontiers in Plant Science, 12, 735143.

[13] Crossa, J., De Los Campos, G., Perez, P., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics, 186(2), 713-724.

[14] Crossa, J., Martini, J. W., Gianola, D., et al. (2019). Deep kernel and deep learning for Genome-Based prediction of single traits in multienvironment breeding trials. Frontiers in Genetics, 10, 1168.

[15] Crossa, J., Perez-Rodriguez, P., Cuevas, J., et al. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. Trends in Plant Science, 22(11), 961-975.

[16] Cuellar, L. K., Friedrich, A., Gabernet, G., et al. (2022). A data management infrastructure for the integration of imaging and omics data in life sciences. BMC Bioinformatics, 23(1), 61.

[17] Dorrity, M. W., Alexandre, C. M., Hamm, M. O., et al. (2021). The regulatory landscape of Arabidopsis thaliana roots at single-cell resolution. Nature Communications, 12(1), 3334.

[18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv. https://doi.org/10.48550/arxiv.2010.11929

[19] Dumschott, K., Dorpholz, H., Laporte, M., et al. (2023). Ontologies for increasing the FAIRness of plant research data. arXiv. https://doi.org/10.48550/arxiv.2309.07129

[20] Elnaggar, A., Heinzinger, M., Dallago, C., et al. (2020). ProtTrans: towards cracking the language of life through self-supervised deep learning. arXiv. https://arxiv.org/abs/2007.06225

[21] Elnaggar, A., Heinzinger, M., Dallago, C., et al. (2021). ProtTrans: Toward understanding the language of life through self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 7112-7127.

[22] Feng, D., Liang, Z., Wang, Y., et al. (2022). Chromatin accessibility illuminates single-cell regulatory dynamics of rice root tips. BMC Biology, 20(1), 274.

[23] Flores, J. E., Claborne, D. M., Weller, Z. D., et al. (2023). Missing data in multi-omics integration: Recent advances

through artificial intelligence. Frontiers in Artificial Intelligence, 6, 1098308.

[24] Fornes, O., Castro-Mondragon, J. A., Khan, A., et al. (2019). JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Research, 48(D1), D87-D92.

[25] Gill, T., Gill, S. K., Saini, D. K., et al. (2022). A comprehensive review of high throughput phenotyping and machine learning for plant stress phenotyping. Phenomics, 2(3), 156-183.

[26] Goff, S. A., Vaughn, M., McKay, S., et al. (2011). The iPlant Collaborative: Cyberinfrastructure for plant biology. Frontiers in Plant Science, 2, 34.

[27] Gong, P., Cheng, L., Zhang, Z., et al. (2023). Multi-omics integration method based on attention deep learning network for biomedical data classification. Computer Methods and Programs in Biomedicine, 231, 107377.

[28] Gu, W., Feng, Y., Liu, C., et al. (2025). Transcriptomic and metabolomic unraveling of nitrogen use efficiency in sorghum. Plant Growth Regulation, 105(3), 687-706.

[29] Guo, Z., Wang, S., Zhang, F., et al. (2024). Common and specific genetic basis of metabolite-mediated drought responses in rice. Stress Biology, 4(1), 6.

[30] Hammer, G., Messina, C., Wu, A., & Cooper, M. (2019). Biological reality and parsimony in crop models. In Silico Plants, 1(1).

[31] Han, E., Geng, Z., Qin, Y., Wang, Y., & Ma, S. (2024). Single-cell network analysis reveals gene expression programs for Arabidopsis root development and metabolism. Plant Communications, 5(8), 100978.

[32] Hassan, A. H., Mokhtar, M. M., & Allali, A. E. (2024). Transposable elements: multifunctional players in the plant genome. Frontiers in Plant Science, 14, 1330127.

[33] Herrero-Huerta, M., Rodriguez-Gonzalvez, P., & Rainey, K. M. (2020). Yield prediction by machine learning from UAS-based multi-sensor data fusion in soybean. Plant Methods, 16(1), 78.

[34] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

[35] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

[36] Hu, T., Zhang, X., Khanal, S., et al. (2024). Climate change impacts on crop yields: A review of empirical findings, statistical crop models, and machine learning methods. Environmental Modelling and Software, 179, 106119.

[37] Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. Frontiers in Genetics, 8, 84.

[38] Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., et al. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science, 373(6555), 655-662.

[39] Humplík, J. F., Lazar, D., Husickova, A., & Spichal, L. (2015). Automated phenotyping of plant shoots using imaging methods for analysis of plant stress responses. Plant Methods, 11(1), 29.

[40] Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS ONE, 5(9), e12776.

[41] Jain, A., Sarsaiya, S., Singh, R., et al. (2024). Omics approaches in understanding the benefits of plant-microbe interactions. Frontiers in Microbiology, 15, 1391059.

[42] Jamil, I. N., Remali, J., Azizan, K. A., et al. (2020). Systematic Multi-Omics Integration (MOI) approach in plant systems biology. Frontiers in Plant Science, 11, 944.

[43] Jayakodi, M., Lu, Q., Pidon, H., et al. (2024). Structural variation in the pangenome of wild and domesticated barley. Nature, 636(8043), 654-662.

[44] Jeong, Y., Ronen, J., Kopp, W., et al. (2024). scMaui: a widely applicable deep learning framework for single-cell multiomics integration in the presence of batch effects and missing data. BMC Bioinformatics, 25(1), 257.

[45] Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics, 37(15), 2112-2120.

[46] Johnson, W. E., Li, C., & Rabinovic, A. (2006). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1), 118-127.

[47] Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.

[48] Kajrolkar, A. (2025). Integrating multi-omics data for plant stress response: Current advances and future directions. Premier Journal of Plant Biology. https://doi.org/10.70389/pjpb.100012

[49] Kannan, J., Palani, T., Selvakumar, D., et al. (2025). Computational approaches in multi-omics for crop improvement. Current Bioinformatics, 20.

[50] Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. Frontiers in Plant Science, 10, 621.

[51] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv. https://doi.org/10.48550/arxiv.1312.6114

[52] Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv. https://arxiv.org/abs/1609.02907

[53] Korsunsky, I., Millard, N., Fan, J., et al. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nature Methods, 16(12), 1289-1296.

[54] Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9(1), 559.

[55] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[56] Leek, J. T., Scharpf, R. B., Bravo, H. C., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics, 11(10), 733-739.

[57] Leek, J. T., & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genetics, 3(9), e161.

[58] Li, Q., Wen, Z., Wu, Z., et al. (2021). A survey on federated learning systems. IEEE Transactions on Knowledge and Data Engineering, 35(4), 3347-3366.

[59] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.

[60] Li, Z., Tao, W., Liu, J., et al. (2023). Tomato leaf disease recognition via optimizing deep learning methods. Horticulturae, 9(9), 1034.

[61] Liang, Z., Anderson, S. N., Noshay, J. M., et al. (2021). Genetic and epigenetic variation in transposable element expression responses to abiotic stress in maize. Plant Physiology, 186(1), 420-433.

[62] Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321-332.

[63] Liu, X., Zhang, F., Hou, Z., et al. (2021). Self-supervised learning: generative or contrastive. IEEE Transactions on Knowledge and Data Engineering, 1.

[64] Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[65] Ma, C., Xin, M., Feldmann, K. A., & Wang, X. (2014). Machine learning-based differential network analysis in Arabidopsis. The Plant Cell, 26(2), 520-537.

[66] Ma, D., Xu, B., Feng, J., et al. (2022). Dynamic metabolomics and transcriptomics analyses for characterization of phenolic compounds in wheat grain. Frontiers in Nutrition, 9, 844337.

[67] Ma, W., Qiu, Z., Song, J., et al. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta, 248(5), 1307-1318.

[68] Maher, K. A., Bajic, M., Kajala, K., et al. (2017). Profiling of accessible chromatin regions across multiple plant species and cell types. The Plant Cell, 30(1), 15-36.

[69] Mahieu, N. G., & Patti, G. J. (2017). Systems-level annotation of a metabolomics data set reduces 25,000 features to fewer than 1,000 unique metabolites. Analytical Chemistry, 89(19), 10397-10406.

[70] Mahmood, U., Li, X., Fan, Y., et al. (2022). Multi-omics revolution to promote plant breeding efficiency. Frontiers in Plant Science, 13, 1062952.

[71] Marks, R. A., Amezquita, E. J., Percival, S., et al. (2023). A critical analysis of plant science literature reveals ongoing inequities. Proceedings of the National Academy of Sciences, 120(10), e2217564120.

[72] Mataigne, A., Vannier, N., Vandenkoornhuyse, P., & Hacquard, S. (2022). Multi-genome metabolic modeling predicts functional inter-dependencies in the Arabidopsis root microbiome. Microbiome, 10(1), 217.

[73] Merrick, L. F., Herr, A. W., Sandhu, K. S., Lozada, D. N., & Carter, A. H. (2022). Optimizing plant breeding programs for genomic selection. Agronomy, 12(3), 714.

[74] Michael, T. P., & VanBuren, R. (2020). Building near-complete plant genomes. Current Opinion in Plant Biology, 54, 26-33.

[75] Minervini, M., Scharr, H., & Tsaftaris, S. A. (2015). Image analysis: The new bottleneck in plant phenotyping. IEEE Signal Processing Magazine, 32(4), 126-131.

[76] Molnar, C. (2022). Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/

[77] Montesinos-Lopez, A., Rivera, C., Pinto, F., et al. (2023). Multimodal deep learning methods enhance genomic prediction of wheat breeding. G3 Genes Genomes Genetics, 13(5).

[78] Montesinos-Lopez, O. A., Chavira-Flores, M., Kiasmiantini, et al. (2024). A review of multimodal deep learning methods for genomic-enabled prediction in plant breeding. Genetics. https://doi.org/10.1093/genetics/iyae161

[79] Montesinos-Lopez, O. A., Montesinos-Lopez, A., Perez-Rodriguez, P., et al. (2021). A review of deep learning applications for genomic selection. BMC Genomics, 22(1), 19.

[80] Muller, M., & Munne-Bosch, S. (2015). Ethylene response factors: a key regulatory hub in hormone and stress signaling. Plant Physiology, 169(1), 32-41.

[81] Mummadi, S. T., Islam, M. K., Busov, V., & Wei, H. (2025). Gene regulatory network prediction using machine learning, deep learning, and hybrid approaches. Forestry Research, 5(1), 0.

[82] Nguyen, Q., Nguyen, H., Oh, E. C., & Nguyen, T. (2024). Current approaches and outstanding challenges of functional annotation of metabolites. Briefings in Bioinformatics, 25(6).

[83] Nguyen, V. H., Morantte, R. I. Z., Lopena, V., et al. (2023). Multi-environment genomic selection in rice elite breeding lines. Rice, 16(1).

[84] Pan, Y., Ji, X., You, J., et al. (2024). CSGDN: Contrastive signed graph diffusion network for predicting crop gene-phenotype associations. arXiv. https://doi.org/10.48550/arxiv.2410.07511

[85] Park, S. Y., & Jeong, D. (2023). Comprehensive analysis of rice seedling transcriptome during dehydration and rehydration. International Journal of Molecular Sciences, 24(9), 8439.

[86] Peleke, F. F., Zumkeller, S. M., Gultas, M., Schmitt, A., & Szymanski, J. (2024). Deep learning the cis-regulatory code for gene expression in selected model plants. Nature Communications, 15(1), 3488.

[87] Rajesh, A., Chang, Y., Abedalthagafi, M. S., et al. (2021). Improving the completeness of public metadata accompanying omics studies. Genome Biology, 22(1), 106.

[88] Rao, R., Liu, J., Verkuil, R., et al. (2021). MSA Transformer. bioRxiv. https://doi.org/10.1101/2021.02.12.430858

[89] Reel, P. S., Reel, S., Pearson, E., Trucco, E., & Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. Biotechnology Advances, 49, 107739.

[90] Rives, A., Meier, J., Sercu, T., et al. (2019). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. bioRxiv. https://doi.org/10.1101/622803

[91] Rives, A., Meier, J., Sercu, T., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15).

[92] Rohart, F., Gautier, B., Singh, A., & Cao, K. L. (2017). mixOmics: An R package for omics feature selection and multiple data integration. PLoS Computational Biology, 13(11), e1005752.

[93] Rotz, S., Duncan, E., Small, M., et al. (2019). The politics of digital agricultural technologies: A preliminary review. Sociologia Ruralis, 59(2), 203-229.

[94] Ryu, K. H., Huang, L., Kang, H. M., & Schiefelbein, J. (2019). Single-cell RNA sequencing resolves molecular relationships among individual plant cells. Plant Physiology, 179(4), 1444-1456.

[95] Schlapfer, P., Zhang, P., Wang, C., et al. (2017). Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. Plant Physiology, 173(4), 2041-2059.

[96] Shamsuddin, D., Danilevicz, M. F., Al-Mamun, H. A., et al. (2024). Multimodal deep learning integration of image, weather, and phenotypic data under temporal effects for early prediction of maize yield. Remote Sensing, 16(21), 4043.

[97] Shiu, S., & Lehti-Shiu, M. D. (2024). Assessing the evolution of research topics in a biological field using plant science as an example. PLoS Biology, 22(5), e3002612.

[98] Song, L., Huang, S. C., Wise, A., et al. (2016). A transcription factor hierarchy defines an environmental stress response network. Science, 354(6312).

[99] Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. Bioinformatics and Biology Insights, 14, 117793221989905.

[100] Sun, L., Jing, Y., Liu, X., et al. (2020). Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in Arabidopsis. Nature Communications, 11(1), 1886.

[101] Sun, Y., Fan, X., Cao, D., et al. (2010). Integration of brassinosteroid signal transduction with the transcription network for plant growth regulation in Arabidopsis. Developmental Cell, 19(5), 765-777.

[102] Tabakhi, S., Suvon, M. N. I., Ahadian, P., & Lu, H. (2022). Multimodal learning for multi-omics: A survey. World Scientific Annual Review of Artificial Intelligence, 01.

[103] Tardieu, F., Cabrera-Bosquet, L., Pridmore, T., & Bennett, M. (2017). Plant phenomics, from sensors to knowledge. Current Biology, 27(15), R770-R783.

[104] Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793-4813.

[105] Trapnell, C., Cacchiarelli, D., Grimsby, J., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology, 32(4), 381-386.

[106] Van Bel, M., Diels, T., Vancaester, E., et al. (2017). PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Research, 46(D1), D1190-D1196.

[107] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[108] Velickovic, P., Cucurull, G., Casanova, A., et al. (2017). Graph attention networks. arXiv. https://arxiv.org/abs/1710.10903

[109] Wang, J., Yang, W., Zhang, S., et al. (2023). A pangenome analysis pipeline provides insights into functional gene identification in rice. Genome Biology, 24(1), 19.

[110] Wang, P., Lehti-Shiu, M. D., Lotreck, S., et al. (2024). Prediction of plant complex traits via integration of multi-omics data. Nature Communications, 15(1), 6856.

[111] Wang, Q., Wang, B., Liu, H., et al. (2022). Comparative proteomic analysis for revealing the advantage mechanisms of salt-tolerant tomato. PeerJ, 10, e12955.

[112] Wang, X., Singh, D., Marla, S., Morris, G., & Poland, J. (2018). Field-based high-throughput phenotyping of plant height in sorghum. Plant Methods, 14(1), 53.

[113] Wang, Y., & Jordan, M. I. (2021). Desiderata for representation learning: a causal perspective. arXiv. https://doi.org/10.48550/arxiv.2109.03795

[114] Wang, Y., Luo, Y., Guo, X., et al. (2024). A spatial transcriptome map of the developing maize ear. Nature Plants, 10(5), 815-827.

[115] Warnat-Herresthal, S., Schultze, H., Shastry, K. L., et al. (2021). Swarm learning for decentralized and confidential clinical machine learning. Nature, 594(7862), 265-270.

[116] Washburn, J. D., Mejia-Guerra, M. K., Ramstein, G., et al. (2019). Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. Proceedings of the National Academy of Sciences, 116(12), 5542-5549.

[117] Way, G. P., & Greene, C. S. (2017). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. Pacific Symposium on Biocomputing, 23, 80-91.

[118] Wei, H., Xu, H., Su, C., Wang, X., & Wang, L. (2022). Rice CIRCADIAN CLOCK ASSOCIATED 1 transcriptionally regulates ABA signaling to confer multiple abiotic stress tolerance. Plant Physiology, 190(2), 1057-1073.

[119] Wu, X., Fan, X., Luo, P., et al. (2023). From laboratory to field: Unsupervised domain adaptation for plant disease recognition in the wild. Plant Phenomics, 5, 0038.

[120] Xiang, Y., Chen, Q., Su, Z., et al. (2022). Deep learning and hyperspectral images based tomato soluble solids content and firmness estimation. Frontiers in Plant Science, 13, 860656.

[121] Xu, F., Wu, T., Cheng, Q., Wang, X., & Yan, J. (2025). Foundation models in plant molecular biology: advances, challenges, and future directions. Frontiers in Plant Science, 16, 1611992.

[122] Xu, L., Ning, S., Xu, X., et al. (2024). Comparative analysis of machine learning models and explainable AI for agriculture drought prediction. Agricultural Water Management, 306, 109176.

[123] Yang, J., Duan, G., Li, C., et al. (2019). The crosstalks between jasmonic acid and other plant hormone signaling. Frontiers in Plant Science, 10, 1349.

[124] Yang, Y., Saand, M. A., Huang, L., et al. (2021). Applications of multi-omics technologies for crop improvement. Frontiers in Plant Science, 12, 563953.

[125] Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. arXiv. https://doi.org/10.48550/arxiv.1903.03894

[126] Yu, T. (2022). AIME: Autoencoder-based integrative multi-omics data embedding that allows for confounder adjustments. PLoS Computational Biology, 18(1), e1009826.

[127] Zaheer, M., Guruganesh, G., Dubey, A., et al. (2020). Big Bird: Transformers for longer sequences. Advances in Neural Information Processing Systems, 33, 17283-17297.

[128] Zargar, S. M., Mir, R. A., Ebinezer, L. B., et al. (2022). Physiological and multi-omics approaches for explaining drought stress tolerance and supporting sustainable production of rice. Frontiers in Plant Science, 12, 803603.

[129] Zhang, F., Wu, J., Sade, N., et al. (2021). Genomic basis underlying the metabolome-mediated drought adaptation of maize. Genome Biology, 22(1), 260.

[130] Zhu, G., Wang, S., Huang, Z., et al. (2018). Rewiring of the fruit metabolome in tomato breeding. Cell, 172(1-2), 249-261.

[131] Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13), i457-i466.