Theoretical and Applied Genetics (2025) 138:132
https://doi.org/10.1007/s00122-025-04910-2

REVIEW

Improving plant breeding through AI-supported data integration

Worasit Sangjan[1] · **Daniel R. Kick**[1] · **Jacob D. Washburn**[1]

Received: 3 October 2024 / Accepted: 21 April 2025

Abstract
Integrating, learning from, and predicting using vast datasets from various scales, platforms, and species is crucial for advancing crop improvement through breeding. Artificial intelligence (AI) is a broad category of methods, many of which have been used in breeding for decades. Recent years have seen an explosion of new AI tools (or old ones at new scales), with exciting applications, both demonstrated and potential, to improve or maybe even revolutionize plant breeding! Example use cases and data types included data mining, phenotyping, monitoring, genetics, multi-omics, environment, management practices, cross-species inference, sustainability, economics, and many others. Improvements in these areas could increase predictive accuracy for plant traits, thereby expediting breeding cycles and optimizing resource management. Aside from improving predictions, AI methods can potentially enhance biological inferences and enable more informed approaches to areas like gene discovery, gene editing, and transformation. At the same time, AI is not going to solve every breeding challenge, and studies so far have shown mixed results depending on the application, dataset, and other factors. AI continues to transform plant breeding, yet its full potential remains unclear, with many possibilities still to be realized. This review explores the transformative potential of AI in plant breeding with a particular focus on its ability to integrate the many diverse streams of data involved. Success in this would open opportunities to improve crop resilience, yield, and sustainability, thus supporting global food security and inspiring the next generation of plant breeding technologies.

Introduction

Integrating vast amounts of data from various scales, platforms, and species is crucial for advancing crop improvement through model-driven plant breeding. This involves collecting and analyzing diverse data types, including genomic sequences, phenotypic observations, environmental conditions, agricultural management practices, and various omics data. The heterogeneity of these data poses significant challenges, complicating integration and impeding valuable insights (Cobb et al. 2019; Crossa et al. 2021; Volk et al. 2021). Additionally, enormous amounts of standard image, spectral, LiDAR (light detection and ranging), X-ray (X-radiation), and other types of data are being generated for use in plant breeding and related applications. Many tools, resources, methods, and technologies have been developed over decades to try and extract meaningful information from all of these sources and use that information for better, smarter, and faster plant breeding. The progress that has been made is phenomenal, with plant breeding methods and technologies advancing so quickly that applied breeding programs can struggle to keep up. A significant amount of this advancement is due to statistical modeling efforts ranging from simple linear models to very complex and intricate prediction, decision-making, integration, extraction, and interpretation tools.

Defining and categorizing artificial intelligence methods

Interest in using "artificial intelligence" (AI) to tackle complex challenges has surged as its usefulness in myriad applications (e.g., protein folding prediction, object recognition, natural language modeling), including many related to plant breeding (e.g., phenotyping, genomic prediction,

Communicated by Diego Jarquin.

✉ Jacob D. Washburn
jacob.washburn@usda.gov

Worasit Sangjan
worasitsangjan.ws@gmail.com

Daniel R. Kick
hello@danielkick.com

[1] Plant Genetics Research Unit, United States Department of Agriculture, Agricultural Research Service, Columbia, MO 65211, USA

Published online: 02 June 2025

etc.), has been demonstrated. However, AI terminology has become increasingly ambiguous due to its varied use in science and the popular press. The authors do not wish to adjudicate which methods "count" as AI or any other category. We use AI expansively to refer to statistical learning models and non-statistical learning models (e.g., expert systems, see Fig. 1), although we will discuss non-statistical learning models only in passing. Similarly, we use statistical learning (following James et al. 2023) to include models ranging from linear regression to tree-based models to deep learning methods. We divide this category into "statistical models," which refers to linear regression and its extensions, and "machine learning models" (ML). The latter category we divide into "classical or traditional machine learning" (K-nearest neighbors (KNN), support vector machines (SVM), etc.) and "deep learning" (DL). It is worth noting that the boundaries between these methods are not always clear. For example, some might categorize penalized regression models and kernel regression models (e.g., reproducing kernel Hilbert space best linear unbiased predictors) as bordering statistical modeling and ML. In contrast, others might classify such models into one or the other groups. A useful framing presented in Breiman (2001) is that of "data modeling" and "algorithmic modeling," with the former approach seeking a stochastic data model for a process while the latter seeks a model that maps input variables to a target response. The former approach leans more toward models with fewer, more readily interpretable parameters, while the latter tends toward less interpretable and more complex but often highly accurate models.

Another challenge in discussing the use of AI in plant breeding is that these technologies have become so ubiquitous that many everyday tasks use AI, even without the user recognizing it. When one uses spell check or autocorrects on email, text messages, or documents, they are using AI. Internet searches, local computer searches, social media feeds, and reading recommendations are all connected to, if not directly using, natural language processing (NLP), large language models (LLM), computer vision, and other forms of AI. For the purposes of this review, the authors will primarily ignore these "mundane" AI applications and focus on applications that are unique to breeding and/or research.

Introduction to AI methods for data integration in breeding

The ever-expanding tool kit of AI methods, many of which have only become practical to use in recent years, offers potentially superior capabilities for many aspects of managing and analyzing large-scale high-dimensional datasets compared to the more limited number of tools available in the recent past. Some studies have demonstrated that AI can enhance the accuracy of plant phenotype and environmental response predictions and, in some cases (but certainly not all), provide insights that may deepen our understanding of plant biology, ultimately contributing to improved breeding outcomes (Kuriakose et al. 2020;

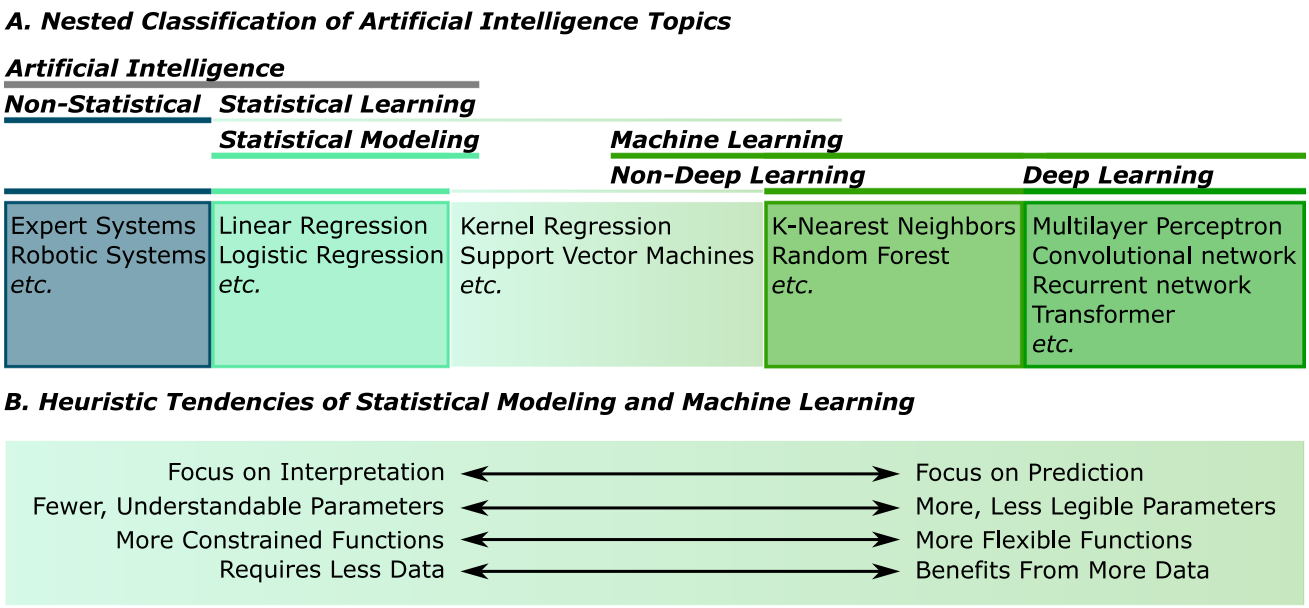

Fig. 1 **A** The categorization scheme used here is influenced by James et al. 2023 and Negus et al. 2024. For the purpose of this review, we focus on statistical learning approaches to artificial intelligence. This category contains statistical and machine learning models, with deep learning models being a subset of the latter. **B** While the distinction between statistical and machine learning models is at times vague, the latter tend toward greater flexibility, often at the cost of interpretability

Yoosefzadeh-Najafabadi et al. 2023). Diverse applications of AI technologies in plant breeding are being explored and hypothesized, offering complementary insights into how neural networks, ensemble learning, and other techniques address challenges in crop improvement and data integration (Negus et al. 2024).

However, it is critical to recognize with some humility that the number of AI methods and applications that have been explored in breeding applications is only a small fraction of what is currently in use in computer science and other domains. Additionally, new and promising methods are being developed every day! Even with extensive efforts and significant progress made by plant scientists and breeders, the AI methods sufficiently tested to demonstrate clear utility are a small drop in the proverbial bucket. Additionally, for many methods, perhaps those categorized as DL in particular, the way these methods have been tested in breeding applications may not be sufficient to bring out their most valuable and well-demonstrated strengths (more on this topic in Sect. "AI for genomic and phenomic data integration in plant breeding").

Convolution neural network (CNN), gradient boosting machine (GBM), and random forest (RF), among other models, have been applied to genomic selection in various crops, including wheat, maize, and potato (Alemu et al. 2024), and have sometimes resulted in enhanced genomic prediction accuracy, especially with sufficient genetic diversity. Moreover, fusing phenotypic data, including hyperspectral imaging, with genomic information using CNN techniques has been shown to improve trait prediction, such as wheat fusarium head blight (Thapa et al. 2024). AI approaches have also shown great promise for merging multi-scale data essential for plant breeding, ranging from molecular-level gene expression profiles to field-level data, like soil conditions and climate variables (Washburn et al. 2021; Kick et al. 2023; Togninalli et al. 2023; Ren et al. 2024). Some of these approaches leverage multi-omics technologies to provide a holistic view of plant development and/or performance.

In combination with technological improvements resulting in the wide availability of drones, cameras, and other phenotyping instruments at relatively low costs, AI methods have proven themselves fundamental to data extraction, mining, and analysis. The importance of these tools in plant breeding is clearly demonstrated and is only likely to increase in future. While phenotyping data that is analyzed weeks or months after its collection are still useful for breeding research, faster turnaround times are needed for many within-season decisions. Real-time or near real-time data integration and analysis, such as early disease detection and trait monitoring in plants, remains challenging (Thompson et al. 2020; Trippa et al. 2023; Wang et al. 2024a). Overcoming these obstacles is essential for improving the accuracy of trait predictions, expediting breeding cycles, optimizing resource management, and gaining comprehensive insights into plant growth and development (Tyagi et al. 2024).

Furthermore, AI can facilitate the integration of data across different platforms and species. Cross-species data integration is particularly promising, though notoriously difficult, in plant breeding as it allows researchers to transfer knowledge and insights from model species to less-studied crops. For example, deep transfer learning was applied to predict disease severity across species, specifically cassava, strawberry, and grape. The method improved classification accuracy for plant disease severity when transferring knowledge between these distantly related species (Yan et al. 2021). Extreme gradient boosting (XGBoost) and RF-based models, initially developed using maize and Arabidopsis data, were later applied to predict traits in rice using rice-specific data (Cheng et al. 2021). Moreover, transfer learning has been employed in conjunction with other methods to predict specialized metabolism genes in tomatoes using data from Arabidopsis (Moore et al. 2020). This cross-species applicability could enhance the efficiency of breeding programs and accelerate the development of improved crop varieties, but these methods do not solve (at least not yet) the common challenges of linkage drag, knock-on effects, and genetic background impacts.

AI has the potential to contribute to optimizing resource management in plant breeding, which is particularly relevant in sustainable agriculture (Rai 2022). For example, CNN and long short-term memory (LSTM) neural networks have been applied to predict soybean maturity from drone imagery, significantly reducing the need for manual field assessments and optimizing the timing of drone flights. These models enhanced prediction accuracy and showed cost-saving potential by requiring fewer flights while maintaining precision in maturity predictions (Moeinizade et al. 2022). RF, least absolute shrinkage and selection operator (LASSO), and multi-trait ensemble genomic prediction models were applied to enhance wheat's complex trait prediction accuracy. These models optimized trait selection, such as grain yield, while managing trade-offs with protein content, improving long-term genetic gains, and preserving genetic diversity (Fradgley et al. 2023).

This review explores how AI technologies can incorporate complex datasets across various scales, platforms, and species to enhance plant breeding strategies. While more "traditional" ML methods have been used extensively to improve plant breeding, examples of the broad application of newer and more complex AI methods across many crops and use cases remain limited. For example, the clear utility of DL approaches for certain image labeling and detection tasks is well established, while its effectiveness in genomic prediction remains underexplored, with some studies showing improvements and others not. This is not surprising, given that the largest recent advancements in AI

technologies have been in image and language tasks, which are not always directly transferable to genomic prediction applications. We give examples and discuss both demonstrated and potential use cases where AI can improve data integration, with a specific focus on data management, genomic and phenomic selection, high-throughput phenotyping, multi-omics data integration, and cross-species integration—areas that represent some of the most impactful and actively explored applications. These advancements contribute to higher prediction performance, accelerated breeding cycles, and optimized resource management. Ultimately, this review highlights AI's future potential in plant breeding, along with a healthy amount of skepticism and a recognition that many AI methods remain in their infancy.

Data integration challenges and opportunities in plant breeding

The importance of data integration

Integrating data across different scales, platforms, and species in plant breeding using AI and other technologies is essential for many reasons, including informed and timely decision-making, efficient use of resources, scalability, sustainability, and effective collaboration across programs (See Fig. 2). Modern AI approaches such as ML and DL can enhance predictive accuracy by analyzing complex datasets to uncover patterns and correlations that may be missed by other methods, leading to more robust predictive models and a more reliable selection of desirable traits (Yang et al. 2022; Kick et al. 2023; Montesinos-López et al. 2024; and many others). Moreover, AI can potentially expedite the identification and development of desirable traits (via genomic selection), reducing the time and cost associated with conventional breeding methods (Rai 2022; Bose et al. 2024). In one example, Togninalli et al. (2023) demonstrated the effectiveness of multi-modal DL by combining genomic, phenotypic, and environmental data to improve the accuracy of wheat grain yield predictions, suggesting it could help breeders identify high-performing lines earlier and support the optimization of resource allocation. AI-driven data integration has the potential to streamline plant breeding by furthering genomics-assisted breeding, thereby improving resource use (Bhat et al. 2023). However, there are barriers to this use, and many breeding programs have yet to explore and utilize modern AI capabilities. Many of these capabilities also remain untested and inaccessible to smaller breeding programs. In fact, a large number of breeding programs still struggle to implement genomic selection methods that

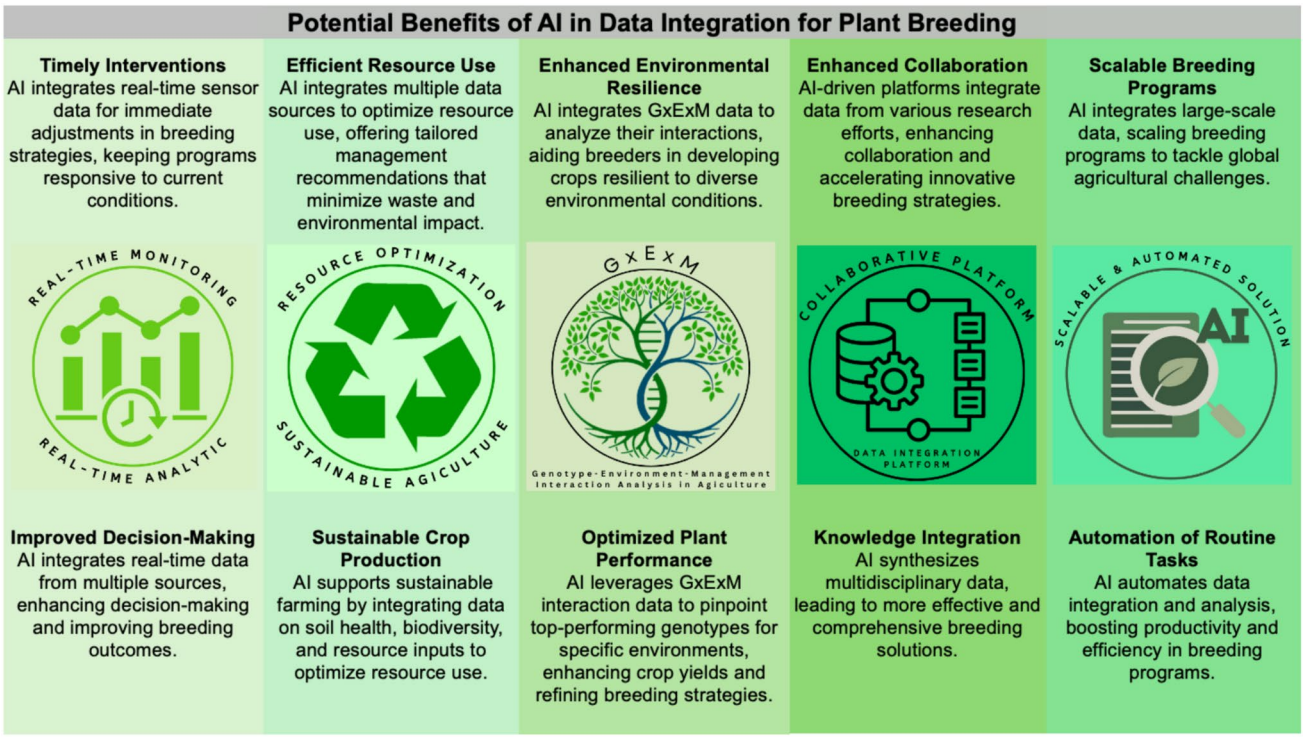

Fig. 2 AI-driven data integration for plant breeding. This figure highlights AI's potential in integrating diverse datasets to enhance breeding strategies. G, Genotype; E, Environment; M, Management; AI, Artificial Intelligence. Reference: (Zhang et al. 2022; Cembrowska-Lech et al. 2023; Yoosefzadeh-Najafabadi et al. 2023; Chen et al. 2024)

are broadly available and well demonstrated due to resource limitations and logistical challenges.

DL methods have been shown in some cases to enhance genomic selection by combining different types of data, providing more accurate estimations of breeding values, and facilitating cross-disciplinary research. Måløy et al. (2021) demonstrated how performer-based DL models improved genomic selection by combining genomic data (single nucleotide polymorphism (SNP) markers) with environmental factors to predict barley yields more accurately. The model effectively captured complex interactions between genotype and environment, outperforming traditional methods. Their attention mechanisms provided insights into the most influential genomic markers and environmental variables, facilitating more informed breeding decisions. Similarly, our group has shown that combining genomic, environmental, and management data using DL and/or ensembles of various models, including best linear unbiased prediction (BLUP), DL, and ML, can improve maize yield prediction, particularly when cross-environment prediction is desired (Washburn et al. 2021; Kick et al. 2023; Kick and Washburn 2023).

AI, in combination with other technologies and sensors, offers real-time monitoring, fast analysis, and adaptive management solutions, improving efficiency and reducing waste by providing timely recommendations based on sensor data. For instance, Yao et al. (2024) employed the seed germination rate-you only look once (SGR-YOLO) model to detect seed germination rates in wild rice. The model utilized an efficient channel attention (ECA) mechanism to extract important features from the images and a bidirectional feature pyramid network (BiFPN) to combine and refine these features. This approach, combined with fused image data from different germination environments over multiple days, enhances the accuracy and reliability of germination rate detection. Wang et al. (2024a) developed the ghost pyramid 2 (GhP2)-YOLO model for real-time rapeseed flower counting. This model incorporated the P2 detection head and ghost modules to improve sensitivity to small objects. By fusing video data from various stages of rapeseed flowering, this model significantly enhanced the tracking and detection accuracy of flowers and buds from the video. The model was also incorporated with the strong simple online and real-time tracking (StrongSORT) multitarget tracking algorithm, allowing real-time monitoring of rapeseed flower development in the field. In addition, Zhang and Li (2022) developed the EPSA-YOLO-V5s model, a YOLO-based architecture, combined with the efficient pyramid split attention (EPSA) mechanism for detecting the survival rate of rapeseed in plant factories. This model, applied across multiple key growth stages, significantly improved detection accuracy by merging environmental data and data augmentation techniques, enabling real-time monitoring and swift interventions to ensure crop survival. These advanced AI-driven methods present a significant opportunity to develop vital tools for real-time monitoring, evaluation, and decision-making in plant breeding programs.

Although AI offers exciting opportunities in plant breeding, much remains to be uncovered before its full potential can be realized. Its ability to fuse data from various sources presents new ways to analyze complex interactions, such as genotype-by-environment-by-management (G x E x M), which are crucial for developing resilient crops (Xu et al. 2022; Aziz and Masmoudi 2024). However, AI's application in this field is still evolving, and more research is needed, such as in combining datasets from wild relatives and landraces—valuable resources for accelerating pre-breeding efforts (Bohra et al. 2022). Continued research and refinement are essential to fully harness AI's potential in plant breeding and address this field's unique challenges.

Categories of data

The diversity and complexity of data types in plant breeding present both challenges and opportunities. These data encompass genetic sequences, omics data, phenotypic observations, environmental conditions, and agricultural practices, each offering insights for advancing crop improvement. Effectively fusing and leveraging these data sources is essential to maximizing their potential (Pommier et al. 2019; Tong and Nikoloski 2021). Table 1, while not an exhaustive list, attempts to categorize some of these foundational data types, highlighting their role (or potential role) in developing robust crop varieties. For example, genetic data reveal hereditary factors associated with traits, while phenotypic data link these traits to observable plant characteristics. Environmental and management data guide adjustments for climate and cultivation techniques to enhance performance.

Omics data provide insights into molecular mechanisms, enhancing the understanding of plant development from gene expression to phenotypic manifestation and enabling targeted trait modifications. Additionally, biotic interaction data, along with historical, socioeconomic, and agronomic information, offer context for developing strategies that address both internal plant responses and external influences like pests, diseases, market demands, and agriculture practices. Combining these diverse data types is crucial for creating comprehensive breeding strategies that improve crop performance, quality, and adaptability, ensuring that new varieties are well-suited to both current and future agriculture demands.

Challenges of data integration

Plant breeding faces many data integration challenges as diverse datasets from high-throughput sequencing, phenotyping platforms, and real-time monitoring become

Table 1 Categories of data used in plant breeding

Category	Definition	Type of data
Genetic & Genomic	Genetic information and variations within plant species	Genomic sequences, genetic markers, GWAS[1], pedigrees, pathogen genomic data
Phenotypic	Observations of physical and physiological traits of plants	Morphological traits, physiological traits, high-throughput plant phenotyping data
Environmental	Information on environmental conditions affecting plant growth	Climate data, soil characteristics, weather conditions, abiotic stress data
Management Practices	Data on techniques and inputs used in crop cultivation	Agricultural inputs (fertilizers, pesticides), cultural practices, field management strategies
Omics	Molecular data on gene expression, proteins, metabolites, and epigenetic modifications	Transcriptomics, proteomics, metabolomics, epigenomics
Species & Evolutionary	Classification, diversity, and evolutionary relationships of plant species	Taxonomic data, biodiversity records, phylogenetic data
Remote Sensing, Sensor, & Phenomic	Data from remote and proximal sensing technologies, including spectral data related to a plant's endophenotype	Satellite imagery, drone imagery, proximal sensing data, soil moisture sensors, weather station data
Biotic Interaction	Information on plant interactions with other organisms	Symbiotic relationships, competitive interactions, plant microbiome interactions
Historical & Socioeconomic	Historical and socioeconomic factors affecting plant breeding	Historical farming practices, breeding lineages, yield records, past climate data, market trends, policy impact, economic data
Agronomic & Experimental	Data on plant performance, resource use, and experimental trial outcomes	Nutrient uptake measurements, water use efficiency, experimental trial data, G^2 x E^3 interaction data

[1]GWAS, genome-wide association study; [2]G, genotype; [3]E, environment; Reference: Washburn et al. 2021; Lassoued et al. 2022; Sangjan et al. 2022a; Kick et al. 2023; Washburn et al. 2023;Sangjan et al. 2024

more prevalent (Gill et al. 2022; Nizamani et al. 2023). The findability of data is a core issue, as information is often stored in non-digital formats or isolated locations, making it difficult to locate and reuse efficiently. Data heterogeneity presents another challenge, with variations in data types, formats, and scales creating inconsistencies that complicate fusion. These differences must be harmonized to ensure compatibility across datasets. The volume and complexity of large datasets generated by breeding technologies necessitate efficient storage and computation capabilities. Without appropriate infrastructure, these datasets can be challenging to handle and analyze. Additionally, the dynamic nature of data, particularly from real-time sensors and monitoring systems, requires immediate processing and fusing to make timely analysis and actionable decisions.

Ensuring data quality and consistency is crucial, as variability in data collection methods, accuracy, and completeness can compromise the reliability of merged datasets. Missing values or inconsistent data collection methods can hinder accurate analysis. Moreover, computational challenges arise from the need for sophisticated analytical approaches, combining plant biology and data science expertise to analyze complex data effectively. This necessitates collaboration and advanced training across disciplines. Data interoperability is a key for merging and sharing data across various platforms, often using different formats and protocols. This challenge requires adherence to standardized exchange protocols to ensure seamless data transfer.

Privacy and security are also significant concerns, as sensitive genetic and breeding data (that is not in the public domain) must be protected from unauthorized access to safeguard intellectual property. Cost and resource constraints can limit the ability to implement advanced data integration techniques, especially in resource-limited breeding programs. Efficient resource allocation and shared tool development are essential for addressing these constraints. Table 2 outlines these challenges, emphasizing their implications and potential solutions. It includes how AI and ML can potentially address these issues through strategies like standardization, real-time processing, and robust computational frameworks (Cembrowska-Lech et al. 2023).

These challenges are neither unique to plant breeding nor AI but manifest at a scale that often precludes or limits the efficacy of ad hoc solutions. A dataset containing tens to hundreds of observations with few variables may be manageable with manual curation and manipulation and certain technologies (e.g., relational databases, version-controlled scripting, etc.). However, conventional approaches can become insufficient as datasets grow in complexity and scale. A hybrid or adaptive strategy that integrates conventional data management methods with advanced AI-driven techniques could be beneficial across different scales. This integrated approach would allow breeding programs to

Table 2 Challenges and solutions in data integration for plant breeding

Challenge	Implication	Potential solution
Findability of Data	Difficulty accessing, reusing, and sharing data due to non-digital or obscure locations	Create centralized, searchable data repositories using AI[1]-driven data mining to automatically extract, categorize, and tag metadata, enhancing both human and machine accessibility (e.g., OmicsDI[2]: partially uses AI tools to improve metadata searchability in multi-omics data) (Perez-Riverol et al. 2017)
Heterogeneity of Data	Inconsistencies in data type, formats, and scales lead to incorporation difficulties and unbalanced datasets	Apply standardization and normalization techniques (Deng et al. 2023); Use within-model combination (e.g., CNN[3] layer concatenation) and ensemble methods (Ren et al. 2024)
Volume & Complexity	Large datasets require efficient storage infrastructure and advanced computational capabilities to combine data	Use of storage file formats with compression and formats with unambiguous data types; Employ robust computational infrastructure, AI algorithms, and dimensionality reduction techniques (e.g., PCA[4], feature engineering, etc.) (Wang et al. 2023a; Sheikh et al. 2024)
Dynamic Nature	Decision-making requires quick, even real-time data processing, combining, and analysis	Implement fast or real-time data processing frameworks and models with optimized computational and/or power efficiency (Sangjan et al. 2022b). Pre-trained AI models and frameworks offer extremely fast turnaround times for large and complicated datasets
Data Quality & Consistency	Variability in data collection and missing and/or errant values affect the reliability of the data integration process	Standardized data collection protocols and apply quality control (Seck et al. 2023); Use shared ontologies and data-gathering protocols; Estimate metadata for nuisance variables as random effects (e.g., differences in collection methods across teams); Use larger datasets to compensate for quality with AI methods (Sheikh et al. 2024)
Computational Challenges	High analytical demands necessitate cross-disciplinary expertise and complex data models	Promote interdisciplinary knowledge and advanced analytical model cross-training and support between plant biology and data science (Kusmec et al. 2021; Jeon et al. 2023)
Data Interoperability	Incompatibility across platforms hinders seamless data sharing and combining efforts	Use standardized data exchange protocols (e.g., BrAPI[5]) and data versioning through digital identifiers; Develop workflows for programmatic data handling; Use of AI tools for data mining and re-formatting (Togninalli et al. 2023)
Privacy & Security	Risk of authorized access to sensitive genetic and breeding information	Implement robust data security measures and protocols (Morales et al. 2022; Zhao et al. 2024) and carefully vet AI tools for safe data use
Cost & Resources	High costs restrict access to advanced integration tools and infrastructure	Optimize resource allocation and cost strategies (Khan et al. 2024); Identify diminishing returns in model performance relative to program goals and budget; Foster collaborations among breeders and scientists for shared tool development. (e.g., BreedBase[6], BrAPI, Breeding Insights[7], Soybase[8], etc.)

[1]AI, artificial intelligence; [2]OmicsDI, omics discovery Index; [3]CNN, convolution neural networks; [4]PCA, principal component analysis; [5]BrAPI, breeding application programming interface (Selby et al. 2019); [6]BreedBase (Morales et al. 2022); [7]Breeding Insights (https://breedinginsight.org/, accessed on August 8, 2024); [8]Soybase (Brown et al. 2021)

optimize efficiency, manageability, and analytical power according to their available resources. Ultimately, the optimal solution depends on clearly assessing the breeding program's specific goals, scale, resources, and complexity.

AI for genomic and phenomic data integration in plant breeding

Over the past 150 years, plant breeding has significantly advanced. In the parlance of Wallace et al. (2018), breeding has evolved from basic phenotypic selection (breeding 1.0) to the integration of statistical theory and experimental design (breeding 2.0) and now to the use of genetic and genomic data in more advanced statistical and computational models (breeding 3.0). These models aim to select genotypes with desirable traits, sometimes identifying specific associated genes or SNPs. Breeding 4.0 has been variously characterized as including the further integration of genetic data, improved computational methods, and plant transformation/gene editing more widely into breeding programs (Ramstein et al. 2019; Kuriakose et al. 2020). Some have suggested that Breeding 5.0 should be characterized by the use of big data and advanced AI methods (Kuriakose et al. 2020; Yoosefzadeh-Najafabadi et al. 2023). While linear modeling has been, and will likely continue to be, productive in breeding, ML and DL methods are emerging as promising options.

DL is notable for its ability to approximate any continuous function (Hornik et al. 1989; Zhou 2020). It has shown excellent practical efficacy in other domains, from medical diagnostics to NLP (Houssein et al. 2023; Alzubaidi et al. 2024). While these methods often perform well with data containing nonlinearities, in practice, their performance in genomic prediction is frequently similar to conventional methods (Montesinos-López et al. 2021a). The performance of these methods may improve with more observations, richer input data, new or different model architectures, or model combinations (Washburn et al. 2020; Montesinos-López et al. 2021a, 2021b).

As plant breeding advances, the modernization of practices through advanced technologies and digital innovations is reshaping the field. AI-driven techniques for merging diverse datasets have become essential tools in this transformation (Ramstein et al. 2019; Fu et al. 2022; Wang et al. 2023b). Using multiple data sources can potentially provide a deeper understanding of the complex interplay between various factors, enhancing the efficiency and effectiveness of breeding strategies to improve plant breeding outcomes relative to simpler models. Envirotyping and enviromics data from satellites and other sources can also enhance predictive models for use in breeding (Resende et al. 2024; Yunbi et al. 2022). This section explores current AI approaches for model-driven selection from genomic, phenomic, or multimodal datasets.

Genomic prediction and selection

Genomic prediction and selection have transformed plant breeding by enabling more accurate predictions of desirable traits through advanced computational models. AI techniques are increasingly being merged to enhance the predictive power of these models. This section explores the use of AI and ML in genomic selection models, recent work focusing on merging genetic and genomic information with other sources, and the benefits of utilizing multiple models in conjunction.

Operating on genotypes

While many approaches for AI and ML-based genomic prediction models have been tested, no one approach has appeared to supersede the rest. Some performance variability is likely due to differences in model tuning procedures, and a method's suitability largely depends on the architecture (purely additive effects, nonlinearities, interactions, etc.) of the trait in question and the specific use of the model. For instance, a study of New Mexican chile peppers tested a variety of AI methods, finding that multilayer perceptron (MLP) models performed best for some traits and linear models performed best for others (Lozada et al. 2023). An examination of models for rust resistance selection found that RF models performed best when predicting all individuals, but an SVM was more effective at selecting the top 15% of individuals (González-Camacho et al. 2018). The specific trait and purpose of the model must be considered when selecting the best model. Another important consideration is that data formatting, quality checking, and dimensionality reduction measures can impact different methods in different ways. While testing multiple methods using the same version of a dataset is simple and convenient, the results may not be representative of the true potential of diverse models and datasets.

Although these factors preclude a universal prescription, published studies do suggest model classes that may be effective. Within genomic prediction DL models, many use MLPs or CNNs, although LSTM has been effectively used in maize and eucalyptus (Maldonado et al. 2020). CNNs seem more effective generally, with specifics varying across studies (Montesinos-López et al. 2021a). Ma et al. (2018b) observed this, finding that the CNN (DeepGS, DL method to predict phenotypes from genotypes using CNN) outperformed MLP and regression models for wheat grain length and other traits. CNNs continue to be applied and refined across species. Promising results in soybeans were reported with a CNN model featuring two parallel processing streams

(Liu et al. 2019), available through an open-access web server, G2PDeep: provides a DL framework for quantitative phenotype prediction and discovery of genomics markers (Zeng et al. 2021). Recently, Wang et al. (2023a) corroborated the efficacy of CNNs, finding that a multilayer CNN with batch normalization layer (DNNGP, deep neural network for genomic prediction) performs similarly to or better than benchmarking models, including other CNN models.

While CNNs and MLPs have proven useful, further exploring data representation strategies and network architecture may improve their performance. Although neural networks can discover relationships from input data, potentially reducing the need for feature engineering, these models may still benefit from specific representations of the data. For example, genotypes can be represented in myriad ways, including as nucleotides (4 discrete values), as homozygous, heterozygous, or missing (3 discrete values) (Liu et al. 2019), as principal components (continuous values) (Washburn et al. 2021), or even as slices of kinship matrices (continuous values) (Nazzicari and Biscarini 2022). Different model types and architectures may be better suited to certain scenarios and require different representations of the data for optimal performance. Continued investigation into matching trait architecture, data representation, and model structure will likely yield increased performance.

In cases where network size and overfitting influence model performance, reducing the number of connections in an MLP by "pruning" the network or initially constraining connections may be profitable avenues for exploration. The latter approach is of particular interest as it can potentially increase model interpretability, e.g., by having the model structure mimic known biological relationships. In the past decade, several efforts have been in this vein (Gazestani and Lewis 2019). Promising results have been obtained in yeast where biologically, an informed model has been able to predict the effect of gene knockouts (Ma et al. 2018a) and to predict treatment resistance in prostate cancer (Elmarakeby et al. 2021). One study showed that this approach assigned higher weights to causal SNPs in a simulation. This method used biological data to identify genes associated with simple human traits (e.g., hair and eye color). However, it was less effective with complex traits (schizophrenia) (van Hilten et al. 2021). The extent to which similar approaches benefit prediction efforts or identification of associated SNPs remains to be seen. Additionally, MLP or CNN models do not provide information on additive and dominance effects in the way other genomic prediction models do.

Leveraging multiple, and non-traditional, data sources

AI methods are exceptionally flexible to the use of multiple data sources and interactions. For example, decision tree-based methods allow for interactions between data types (e.g., G x E, G x E x M, etc.) without being explicitly specified. This can be a valuable property, especially for traits where environmental effects explain a meaningful component of the trait's variance, as is the case for many traits of interest in plant breeding (Rogers et al. 2021). This is not to suggest that interactions between diverse and high-dimensional covariates cannot be effectively included in linear models (for example, see Jarquín et al. 2014) but that AI methods increase the variety of approaches to include these interactions. Using a decision tree-based GBM, Westhues et al. (2021) found that maize yield predictions (but not height predictions) were substantially improved by the use of genotypic and environmental data for predictions of a new genotype and year combination. A recent investigation by Fernandes et al. (2024) reported similar findings with the inclusion of environmental data improving GBM performance relative to a factor analytic model.

DL models provide an exciting approach to incorporating interactions between data types by dedicating regions of the model to process each type separately, directly within their architectures. Typically, these models handle multimodal data inputs through parallel processing streams, where each data type is processed separately. These streams allow each data modality to be transformed independently before being combined in feature fusion layers or as components of a weighted prediction (Wang et al. 2023a). Attention mechanisms can be incorporated to aid in learning the most relevant features from each dataset, resulting in a more useful unified representation for improved prediction accuracy (Ren et al. 2024). This approach has been tested for maize yield prediction using data from the Genomes to Fields Initiative (G2F; https://www.genomes2fields.org/, accessed on August 8, 2024) (McFarland et al. 2020; Lima et al. 2023a, b; Washburn et al. 2024) and found to be an effective strategy (Washburn et al. 2021), particularly when the portions of the network operating on each data type are optimized independently (Kick et al. 2023). Using a portion of the G2F data, Sharma et al. (2022) took a distinct but conceptually related approach by creating embeddings of soil, management, weather, and genomic data before applying a multimodal cross-attention module to capture G x E effects and passing the resulting embeddings into a shared portion of the model, reporting better performance for prediction in new environments and similar performance for new genetics.

A different way to leverage multiple datasets is by creating additional model targets rather than inputs. Multi-trait prediction has been effectively deployed with linear and more advanced AI models (e.g., Montesinos-López et al. 2019). The exciting possibility here is not necessarily the multi-trait prediction of secondary phenotypes but the use of unconventional secondary phenotypes. Sandhu et al. (2021) incorporated high-throughput plant phenotyping data into a model using vegetation indices

as secondary phenotypes for genomic prediction. They report an improvement in prediction accuracies for grain yield and grain protein content using multi-trait models rather than unitrait models. With increased availability and diversity of data, the capability of AI models to flexibly incorporate a wide range of data will only increase in usefulness. It is important to note that each model has strengths and weaknesses, and breeding programs also have practical and resource constraints. For example, MLP and CNN models require different computational resources than linear models. The time required for training CNN models is often far greater than that required for linear models.

Using multiple models

In selecting models for a plant breeding program, the decision may not be a matter of "which model" but "which models." Combining predictions from multiple models and types of models into an "ensemble" is often effective for prediction. This can take the form of training one model to predict a variable and others to predict the *error* of the first (model stacking) or aggregating predictions from multiple models (e.g., through a weighted average). Both strategies may be effective, with the latter approach having the additional benefit of being potentially computationally inexpensive, provided that the base models are already trained and allow for leveraging performant models such as interaction containing BLUPs and deep neural networks to further increase prediction accuracy (Kick and Washburn 2023).

Ma et al. (2018b) demonstrate that an ensemble of a CNN and ridge regression (RR)-BLUP improved the selection of top-performing individuals and found that their CNN model and ensemble outperform RR-BLUP when extreme values are removed owing to the model's differential sensitivity. Diverse model types presumably can better represent different relationships in the data and thus stand to benefit from their differential sensitivity in an ensemble. Furthermore, model averaging appears robust to implementation details with unoptimized combinations of multimodal models and weightings, often improving performance (Kick and Washburn 2023). The effectiveness of using multiple models is highlighted by a recent competition in which teams from industry, academia, and government sought to most accurately predict the yield of diverse maize genotypes across North America (Lima et al. 2023c). Interestingly, 52% of the teams, including the winning team, used an ensemble model (Washburn et al. 2024). Also of interest, there were highly ranked teams using linear models, DL, and other methods both alone and as ensembles.

High-throughput plant phenotyping and phenomic selection

High-throughput phenotyping (HTP) has significantly expanded the availability and scale of phenotyping data, enhancing its application in breeding programs and management decisions. HTP enables the assessment of traits that were previously too costly or labor-intensive to measure at scale (Sangjan et al. 2023). Spectral data, such as near-infrared (NIR), are valuable for plant breeding as it can serve as a proxy for traits of interest, supplementing genetic information in phenomic prediction or providing secondary traits for selection (Murray et al. 2008; Cook et al. 2012; Washburn et al. 2013; Sandhu et al. 2021). When these traits can be measured from the seed or early in the growing season, they allow for early selection before pollination. Data collection from unoccupied aerial vehicles/systems (UAV, UAS, Drone) has become commonplace in plant breeding, but the technologies and methodologies continue to evolve. This section will discuss integrating HTP into prediction models for plant breeding.

Data extraction for breeding from HTP

Perhaps the most advanced and well-demonstrated uses of non-traditional AI methods in plant breeding today are computer vision methods used to extract numerical information from images and other sensor data. The phenotyping of traits that are easy to measure and correlated with more important traits that are difficult to measure has always been important for breeding. Imaging methods, including microscopy, spectroscopy, and photography, have a long history of use in breeding and other research endeavors. However, both the data collection and extraction tools used have progressed substantially. The trend toward more complex and automated methods, if simply to increase the number of data points generated with a given amount of funds or time, is happening at a rapid pace. This topic has been reviewed extensively in other recent articles, so we mention only a few examples here (Murphy et al. 2024; Sheikh et al. 2024).

Phenotype extraction from image data has been broadly applied. For example, a team led by researchers at the Oak Ridge National Laboratory applied a few-shot learning (training few samples) and CNN approach, along with other image-processing tools, to identify and quantify leaf traits from thousands of field-collected tree leaf images (Lagergren et al. 2023). Others have successfully used deep neural networks (DNNs) and more traditional ML to identify and quantify diseases on corn leaves (DeChant et al. 2017; Wu et al. 2019), detect and quantify Fusarium head blight on wheat spikes (Cooper et.al. 2023), phenotype seeds and other reproductive organs (Miller et al. 2017), and create digital twins and virtual reality models of plants and fields

(Mitsanis et al. 2024). Significant progress has also been made in root phenotyping and data extraction, including various image analysis and extraction tools, as well as X-ray computed tomography (Bucksch et al. 2014; Seethepalli et al. 2021; Ju et al. 2024).

Integrating HTP data in predictive models

In the context of prediction and selection, reflectance measurements provide a non-destructive means of quantifying traits that may not be visible to the naked eye. Using NIR spectra in "phenomic selection" can enable similar performance relative to the use of molecular markers while reducing the per-sample cost (Rincent et al. 2018). Model predictive ability is influenced by the organ assayed (e.g., leaf, grain, stem, etc.), making tissue and timing essential implementation details.

The architecture of the trait or traits under selection appears to influence the efficacy of phenomic selection without genetic information. Analysis of wheat in multi-environmental trials (using linear models) supports this, finding competitive performance between genomic and phenomic prediction for grain yield but slight underperformance for heading date, a trait with predominantly additive variance (Robert et al. 2022). A similar study in maize finds that phenomic prediction is most accurate for predicting the performance of new genotypes in tested environments and functionally equivalent to genomic prediction for new genotypes in new environments (DeSalvio et al. 2024). Phenomic prediction using a CNN outperformed or matched genomic prediction traits related to Fusarium head blight in winter wheat (Thapa et al. 2024). Within phenomic prediction, different complex traits seem to be best predicted using different bands (Zhu et al. 2021). Taken holistically, phenomic prediction can compete with genomic selection, but performance varies across traits, and if multiple traits are considered, collecting data covering the entire spectrum is prudent. It is also important to understand that phenomic and genomic selection are not always directly comparable, as different use cases will favor one or the other.

Phenomic selection may be especially valuable for crops where the cost or availability of genetic tools remains a barrier. An analysis of a plant breeding program for one such crop, *Coffea canephora* (Robusta/Conilon coffee), observed that NIR outperformed genomics for several environments. Having NIR does not remove the usefulness of genomic data; the best performance was achieved using both (Adunola et al. 2024) and can be used to model G x E effects (Robert et al. 2022). The specific costs involved are important to the utility of these methods. Analysis of a large soybean panel suggests that phenomic prediction models can be competitive with genomic prediction and perform well at making predictions in new environments, even given data for a single environment (Zhu et al. 2021). They noted that while using NIR spectroscopy and genotyping together is optimal for model performance, the increase may not be worth the cost of genotyping. As with most technologies, determining the optimal use requires balancing trade-offs. In the first case presented here, NIR was collected on the coffee cherry, requiring more time for plant development than a tissue sample for sequencing would. In many cases, the specific costs, opportunity costs, and existing data available to a breeding program may dictate the most useful modeling strategy.

In terms of the modeling approach, while some of the studies above that have considered AI methods have found them to be effective (Sandhu et al. 2021; Thapa et al. 2024), others have preferred variations of linear models instead (Zhu et al. 2021; Winn et al. 2023; Adunola et al. 2024). Compared to genomic selection work outlined in the previous section, model architectures have been less explored. This suggests promising exploration opportunities, especially when combining phenomic data with genetic, environmental, and management data. These studies indicate that phenomic selection is a valuable approach and can be competitive with genomic selection, particularly when combined with additional data sources.

AI for multi-omics and cross-species data integration in plant breeding

Multi-omics data integration

AI techniques have been used and hold even more future potential for integrating multi-omics data, which includes genomics (deoxyribonucleic acid (DNA) sequence information), transcriptomics (ribonucleic acid (RNA) expression data), proteomics (protein abundance and interaction data), and metabolomics (metabolite profiles). Combining these diverse data types into a unified analytical framework could uncover a deeper understanding of the biological processes that govern plant traits. This integration is essential for developing predictive models and biological networks that reveal the interactions among various molecular entities (Ahmed et al. 2024; Raza et al. 2024). However, as AI's application in this area continues to evolve, some objectives may take time to achieve or may not be fully attainable, and applying these multi-omics frameworks broadly in breeding programs is resource-intensive. The following section examines the current research on AI's role in combining multi-omics data (see Table 3).

Table 3 AI techniques to integrate different scales and platforms of data in plant breeding and other non-plant research applications

Application example	Technique	How it integrates data
Phenotypic prediction and gene mining from multi-omics data in Arabidopsis and maize (Ren et al. 2024)	Deep Learning	Extract relevant features from each omics layer using multi-head self-attention. Then, features are combined using fully connected layers to make predictions
Genomic prediction of plant traits using multi-omics data such as genomics, transcriptomics, and proteomics in wheat, maize, and tomato (Wang et al. 2023a)	Deep CNN[1]	Uses a multi-layered CNN architecture to dynamically learn features from multi-omics data, integrating them for phenotypic predictions
Predicting disease-resistance genes in oilseed rape from multi-omics data (e.g., genomics, transcriptomics) (Wang et al. 2024b)	Random Forest	Process multi-omics features to prioritize genes within resistance-associated loci. The integration occurs during model training, combining omics features for gene prediction
Grain yield prediction in wheat breeding using genomics and phenomics data (Togninalli et al. 2023)	Deep Learning	Merge multiple data sources using attention mechanisms in a multiple-instance learning framework. Features are extracted and fused into a unified representation to predict traits
Trait prediction such as Fusarium-damaged kernels and deoxynivalenol in winter wheat using multiple datasets (genomics, phenomics, and hyperspectral imaging data) (Thapa et al. 2024)	CNN	CNN extracts features from multiple data types and combines them in its architecture during model training to predict traits. Integration occurs as the model learns from multiple sources simultaneously
Cancer recurrence prediction in human and biomarker discovery using multi-omics data (e.g., SNV[2], mRNA[3], and miRNA[4] expression) (Lan et al. 2024)	Deep Learning	Establish connections based on biological pathways, with self-attention mechanisms to learn correlations between different data. Features from these omics datasets are further merged through fully connected layers to make predictions

[1]CNN, convolution neural networks; [2]SNV, single nucleotide variant; [3]mRNA, messenger ribonucleic acid; [4]miRNA, micro-ribonucleic acid

AI techniques and predictive accuracy for multi-omics data integration

Integrating multi-omics data through advanced AI and ML techniques shows great potential for enhancing predictive modeling in plant breeding. For example, Wang et al. (2024b) exemplified this by utilizing the iMAP (integrative multi-omics analysis and ML for target gene prediction) algorithm to combine genome-wide association study (GWAS) and transcriptomics data, effectively identifying calcium signaling genes linked to disease resistance in oilseed rape. This study illustrates RF's potential to manage complex multi-omics data to enable faster and more accurate identification of resistance genes. Wang et al. (2023a) developed DNNGP, a DNN-based model, to fuse multi-omics data, including SNPs and transcriptomics, for predicting complex traits such as grain yield. DNNGP captures nonlinear interactions across biological layers by leveraging a multi-layered neural network structure, allowing the model to uncover complex relationships between genes, proteins, and other omics data. This enhances prediction accuracy for traits such as plant height and kernel number in maize. These techniques can be applied to merge omics data, presenting valuable opportunities for identifying genetic markers that enhance plant resilience to diseases and stress responses.

Ren et al. (2024) used dual-extraction modeling (DEM), a DL-based multi-modal architecture designed to fuse diverse omics layers for accurate phenotypic prediction and functional gene mining. The DEM platform was developed as a software tool to facilitate its application within the research community, allowing researchers to seamlessly incorporate multi-modal omics datasets and phenotypic data for classification and regression tasks. This underscores the growing accessibility of AI-driven multi-omics integration in plant research.

Leveraging AI techniques on combined datasets also offers the potential to improve the accuracy and reliability of predictions related to various plant traits, such as yield, disease resistance, and stress tolerance. For instance, Wu et al. (2024) applied ML models on fused multi-omics data in maize, combining SNPs, phenomics traits, and metabolic profiles. Their models showed improved maize yield prediction accuracy, with the RF model performing best due to its nonlinear feature selection capabilities, and it was able to identify associated genomic regions. Cheng et al. (2023) demonstrated that RF, SVM, and artificial neural networks (ANN) could offer insights into plant resilience. Their study combined metagenomic data and functional gene profiling with soil and plant traits to analyze root-associated microbial communities in rice cultivars under cadmium stress. The models identified microbial biomarkers involved in rice's cadmium accumulation and stress tolerance. Yang et al. (2024) demonstrated that CNNs can predict epigenetic

signals and uncover complex genomic interactions from integrated, large-scale multi-omics datasets, including genomes, transcriptomes, and epigenetic data from thousands of accessions.

As shown in Table 4, various AI techniques can enhance predictive accuracy when analyzing combined data across omics layers and from multiple scales and platforms.

Network construction through multi-omics data integration

AI-driven methods offer promising possibilities not only for predictive modeling but also for constructing and analyzing intricate biological networks from multi-omics data and other diverse sources. These approaches may provide a deeper understanding of the molecular pathways that influence plant traits, potentially advancing crop improvement and management for more precise, efficient, and sustainable outcomes (Ko and Brandizzi 2020; Chen et al. 2023; Manickam et al. 2023). Some of the networks that can be developed and/or analyzed through such integration include gene regulatory networks (GRNs), protein–protein interaction networks (PPINs), and metabolic networks.

GRNs are developed by incorporating gene expression data with genomic information to map gene regulatory relationships. These networks offer insights into gene expression and trait regulation, helping researchers identify regulatory genes and interactions. Understanding GRNs is useful for deciphering plant stress responses (Kulkarni and Vandepoele 2020). AI and ML techniques can enhance GRN inference by effectively handling and combining complex, high-dimensional data, leading to the opportunity for improving accuracy in predicting regulatory relationships and gene interactions. Cassan et al. (2024) demonstrated that a weighted RF could enhance GRN inference by merging transcription factor binding motifs (TFBMs) and optimizing data integration to minimize prediction errors in Arabidopsis. Adjusting TFBMs integration improved model performance, enabling more precise regulatory interaction predictions. Similarly, Lin and Ou-Yang (2023) showed that a DL model, DeepMCL: deep metric learning for cell-type labeling, enhances GRN inference by leveraging multiple data sources, highlighting DL's potential in advancing GRN analysis.

PPINs illustrate protein interactions by combining proteomic and genomic data, which is useful for understanding cellular functions and pathways. Mapping these networks helps identify essential proteins involved in important processes, aiding in improving traits such as disease resistance and stress tolerance in plants (Mishra et al. 2022; Shi et al. 2023). AI and ML can potentially strengthen the PPINs analysis; for example, Zhang et al. (2019) introduced DeepFunc, a DL framework that fuses protein sequences and PPINs using the DeepWalk algorithm to improve protein function prediction. Similarly, Pan et al. (2022) developed DWPPI (DeepWalk-based protein–protein interaction), a model designed to predict PPIs in plants by combining multi-source data and large-scale biological networks—the model utilized graph embedding algorithms for data integration and analyzed protein sequences to predict interactions. This approach improved accuracy in mapping PPIs and identifying regulatory proteins across species such as Arabidopsis, maize, and rice.

Metabolic networks integrate metabolomic, transcriptomic, and proteomic data to map plant biochemical pathways, revealing how metabolites are produced and used. AI and ML-driven approaches can potentially advance understanding of these interactions (Zulfiqar et al. 2024). For instance, Li et al. (2020) applied k-mean clustering and principal component analysis (PCA) to analyze high-resolution spatiotemporal metabolome and transcriptome data, metabolic network, from MicroTom tomatoes, identifying regulatory networks that control metabolic processes throughout the growth cycle. Babadi et al. (2023) developed ShAdow pRice-based meTabolite pRotein intEraction (SARTRE), a computational framework merging RF classifier with constraint-based modeling to predict metabolite–protein interactions (MPIs) within metabolic networks. This method achieves high accuracy in model organisms. Although these studies focus on model organisms and systems, identifying regulatory networks and MPIs holds significant potential for plant breeding research. Understanding these interactions is one of the keys to improving plant stress resilience, growth, and nutritional quality (Niu et al. 2020; Manickam et al. 2023).

One notable innovation in multi-omics data integration is AlphaFold, an open-source AI tool developed by Google DeepMind for accurate three-dimensional (3D) protein structure modeling (Jumper et al. 2021). Even though primarily focused on protein folding, AlphaFold has significant potential for application in omics studies by providing crucial insights into predicting PPIs, exploring gene functions, and mapping metabolic pathways. This AI-driven tool enhances the understanding of biological networks, including GNNs, PPINs, and metabolic networks, empowering researchers to incorporate structural protein data into multi-omics analyses, thereby advancing plant biology and crop improvement (Yin et al. 2023; Tavis and Hettich 2024).

Cross-species data integration and transfer learning

Transferring knowledge represented in a model from one task to a related task can be an effective strategy to increase model performance or decrease the computational or data resources needed for a model to perform well. However, outside of image processing (e.g., Jiang and Li 2020), the application of this approach has been limited in agriculture.

Table 4 AI techniques for analyzing integrated data to enhance predictive accuracy

Application example	Technique	How it uses integrated data
Yield prediction in maize from the integration of genomics, phenomics, and metabolomics data (Wu et al. 2024)	Random Forest	Select key features and capture nonlinear relationships in the data by combining many rules into decision trees and aggregating these decision trees
	PLS[1]	Maximize covariance between predictors and response variables by projecting them into a lower-dimensional space
	Gaussian Process	Apply nonlinear regression to capture complex relationships in the data, and the RBF[2] kernel allows flexible modeling of patterns in the data
	BART[3]	Combine ensemble regression trees with a Bayesian framework to handle nonlinear interactions in the data while using a regularizing prior to limit the influence of individual trees
Grain yield, plant height, and heading date prediction from combined genomics, phenomics, and environmental data in soft white winter wheat (Montesinos-López et al. 2024)	Random Forest	Aggregate decision trees to capture non-linear relationships and use Boruta for feature selection alongside its own impurity-based variable importance
	BRR[4]	Estimates the data by fitting a Bayesian regression model, capturing interactions between genotype, phenomics, and environment
Predict genes involved in the biosynthesis of plant-specialized metabolites across Arabidopsis, maize, and tomato (Bai et al. 2024)	XGBoost[5]	Use boosting techniques to optimize predictions from the data, improving feature selection
	LightGBM[6] Boosted Trees	Use gradient boosting to handle high-dimensional data, efficiently capturing key features
	Ensemble Model[7]	Leverage the strengths of each individual model and compensate for their weaknesses to capture complex interactions across integrated data
Predicts interactions between multi-omics data from rice and *Magnaporthe oryzae* (a fungal pathogen causing rice blast disease) (Zhao et al. 2023)	GAE[8]	Encode the data into a graph structure and learn lower-dimensional latent representations to reveal hidden relationships between biological entities
	VGAE[9]	Extend GAE with probabilistic modeling to capture complex dependencies and similarities in the latent space representations
Predict gene expression in maize under biotic stress using transcriptomic data from multiple sources (Nazari et al. 2023)	BiLSTM[10]	Process transcriptomic sequence in treating expression data as a sequence to more accurately capture complex relationships
Predict epigenetic signals from the integrated multi-omics data of the soybean genome (Yang et al. 2024)	CNN[11]	Extract features using convolutional layers to capture patterns in the fused data

[1]PLS, partial least squares; [2]RBF, radial basis function; [3]BART, Bayesian additive regression trees; [4]BRR, Bayesian ridge regression;[5] XGBoost, extreme gradient boosting; [6]LightGBM, light gradient boosting machine; [7]Ensemble Model in the study comprise of Random Forests, LightGBM Boosted Trees, CatBoost Boosted Trees, Extremely Randomized Trees (ExtraTrees), XGBoost, and neural networks; [8]GAE graph autoencoder; [9]VGAE, variational graph autoencoder; [10]BiLSTM, bidirectional long short-term memory; [11]CNN: convolution neural networks

One approach to transfer learning is to include a pre-training phase using data similar to the target dataset, e.g., data could be from a related system (sorghum vs. maize) or at a different level of resolution (county-level yield vs. plot-level yield). Assuming similar relationships between the pre-training and target datasets, the model weights will encode a pattern more similar to the target task than they would if the weights were randomly initialized. Washburn et al. (2021) used this approach to successfully increase model accuracy by using historical data on maize yield in the USA prior to the plot-level experimental dataset. The authors noted that pre-training changes the relative importance of the input data (soil, rather than genetics being most influential with pre-training), highlighting that fundamental changes in how the model behaves may result.

Ubbens et al. (2023) proposed an interesting strategy for genomic prediction similar to pre-training but sufficiently distinct to warrant separate consideration. They proposed using a network fit with simulated data to predict observed data. The simulated data mimics the target population's population structure. After training on this synthetic data, the "Genomic Prior-Data Fitted Network," which consists of a transformer and operates on principal component loadings, can perform inference. This approach overperformed genomic best linear unbiased predictor (GBLUP) for a majority of traits in wheat and lentil datasets (unstructured population) and performed well across most locations used for the soybean nested association mapping panel (SoyNAM) but underperformed in some.

Beyond leveraging data at different scales, data from different regions or species can be leveraged. Wang et al. (2018) used a model of soybean yield based on satellite imagery from Argentina to predict soybean yield in Brazil via transfer learning. This approach is beneficial because it reduces the dataset size required for a performant model and can result in substantially faster training (fourfold speedup in this case). A different strategy is to combine datasets using multitarget learning. Unlike the multitarget prediction models discussed earlier, where multiple traits were used, Khaki et al. (2021) demonstrated that maize and soybean yield can be predicted with a single model by using a cleverly designed neural architecture. Conceptually, this allows for shared relationships (e.g., weather conditions conducive for growth in both species) to be learned from more observations while still permitting trait-specific (or species-specific) relationships to be learned. Phylogenetic relatedness between conserved genes across two species has also been used as a means for predicting RNA expression differences and similarities (Washburn et al. 2019).

Note, however, that these examples are not genomic prediction models. Using transfer learning for genomic prediction is an exciting possibility but presents certain challenges. Even if data from a single species are considered, the number of markers and their position in the genome across studies may vary meaningfully. Using transformed data does not necessarily circumvent this issue. For instance, PCA-transformed datasets are neither guaranteed to have an equal number of principal components nor similar loadings. These challenges are compounded for data from multiple species as factors such as ploidy, number of chromosomes, genes present, and relationships between genes and gene networks may differ substantially. Other potential solutions include the use of SNP chips, filtering methods, or imputation, but this becomes more challenging as the distance between species increases.

Modern phylogenomic methods and the use of gene synteny can provide hundreds or even thousands of common genes across species (Lyons and Freeling 2008; Washburn et al. 2017; Grass Phylogeny Working Group III et al. 2024; many others). While these approaches have been very successful for evolutionary studies, applying them directly to breeding has been more challenging. Non-genic SNPs, for example, can be very difficult to recover and compare across species, and functional conservation is not guaranteed even for genes. A significant effort was made recently, and is ongoing, to develop resources and tools (many within the umbrella of AI) for using species from across the grass (*Poaceae)* tribe Andropogoneae to improve crop species like corn and sorghum (Li et al. 2024; Zhai et al. 2024). This effort has resulted in many useful AI tools for cross-species analyses as well as sequenced genomes of crop wild relatives. A separate but related project was able to demonstrate successful cross-species modeling between maize and sorghum in relation to drought stress (Pardo et al. 2023). Ultimately, developing cross-species environmental embedding models may be desirable to enable model development to focus on learning crop-specific relationships. If network size and overfitting influence model performance, reducing the number of connections in an MLP by "pruning" the network or initially constraining connections may be profitable avenues for exploration. The latter approach is of particular interest as it can potentially increase model interpretability by having the model structure mimic known biological relationships.

Synthesis and future directions of AI in plant breeding

AI methods have been a vital part of breeding for years, and they have revolutionized the way breeding is done in many respects (e.g., genomic prediction and HTP methods in the past few decades). More recent AI advances, such as DL and generative AI, are also significantly impacting the field, and they appear poised to transform additional aspects of breeding. Which areas will be most impacted by the application

of new methods is unknown, but it seems likely that AI will increasingly touch every aspect of modern plant breeding just as it is doing for human society in general. This widespread integration of AI highlights the importance of building modern breeding teams comprising specialists in plant breeding, genomics, bioinformatics, data science, machine learning, and data management.

The amount and diversity of data collected by modern breeding programs are enormous and only likely to increase. Integrating, mining, analyzing, and making decisions from these data is challenging and requires ever-evolving methods. Recently developed AI methods have demonstrated the potential to manage and analyze large, complex datasets from various sources. These approaches show promise in addressing traditional challenges in data integration, such as heterogeneity, high dimensionality, and scalability limitations. AI methods can also combine data across different scales, platforms, and species, potentially leading to a more comprehensive view of plant development and performance. Data integration using AI offers the potential to identify the genetic basis of desirable traits. This could inform precise breeding strategies and improve program efficiency (Thapa et al. 2024).

Across the breeding landscape, AI applications exist on a spectrum of adoption rather than a simple current or future dichotomy. Technologies such as phenotype prediction, environmental response modeling, and multi-data source integration are operational realities in some advanced breeding programs while remaining aspirational for others. The number of AI tools and applications that have been tested and explored in breeding remains small in comparison with those available now and those being newly developed on a regular basis. New and different ways of feeding data into modern tools like DL and generative AI need to be tested and broadly demonstrated to enable their wider uptake into applied breeding programs. Moving new methods from the realm of promising results in a limited number of studies to fine-tuned, reliable, and accessible breeding tools will require time, resources, creativity, and very significant efforts.

While traditional statistical models, such as BLUPs, remain highly effective, the addition of DL and ensemble models provides potential new avenues for identifying promising genotypes, potentially reducing breeding cycles and saving both time and resources (Ma et al. 2018b; Togninalli et al. 2023; Azrai et al. 2024). Early phenotyping and fast turnover for decision-making using AI's capacity to process and merge data from diverse sources quickly can enable breeders to identify critical information earlier and at larger scales than traditional methods (de Castro et al. 2019). These enhanced methods would support proactive breeding decisions and accelerate the development of improved crop varieties (Leukel et al. 2023; Salehi et al. 2024). AI may also enable simulations and integration of various environmental and genetic scenarios, offering the possibility of reliable predictions that lessen the need for large-scale physical trials. Although the extent of technologies' capacity to minimize the need for extensive field trials and datasets is not yet known, the potential for additional and significant resource savings is certainly there. (Rai 2022; Fradgley et al. 2023).

AI models are being developed to assess genotype performance across various environments. These models could help breeders identify traits that contribute to resilience under environmental stress, though this remains an area of active research (Sinha et al. 2023; Mushtaq et al. 2024). AI can integrate molecular-level data with environmental and phenotypic data, potentially providing a holistic view of plant growth and development. This integration may lead to more effective breeding strategies by allowing breeders to see the bigger picture across scales (Kick and Washburn 2023; Fernandes et al. 2024).

AI methods show significant promise for cross-species data integration for breeding, including the use of crop wild relatives. They can be used to leverage data from well-studied model organisms to hopefully inform breeding strategies for less-studied crops, potentially speeding up the discovery and integration of beneficial alleles and reducing the time required to develop new varieties (Yan et al. 2021; Xu et al. 2022). Another area where this could have particularly important benefits is for breeders of poorly funded crop systems and those with long generation times and other impediments to traditional research methods. Utilizing existing data from well-studied species could reduce the need for extensive data collection in less-studied crops, lowering research costs (Yan et al. 2021).

The likely payoffs of these efforts include faster, better, more sustainable breeding and agricultural systems as well as improvements in resource use efficiency, but none of these things will come easily. AI is not a silver bullet that will magically fix over a hundred years of carefully identified plant breeding challenges! Significant, sustained, and collaborative efforts will be required to realize any of the potential gains described above. Interdisciplinary training and interdisciplinary teams spanning computer science, plant biology, breeding, data science, and other areas are vital to addressing agricultural challenges effectively (Interdisciplinary Plant Science Consortium 2023). Additionally, ongoing investment in infrastructure, data governance, and ethical AI use is crucial for ensuring data quality, security, and sustainability (Dara et al. 2022).

Acknowledgements This research was supported in part by an appointment to the Agricultural Research Service (ARS) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the U.S. Department of

Agriculture (USDA). ORISE is managed by ORAU under DOE contract number DE-SC0014664.

Author contributions WS contributed to methodology, validation, formal analysis, investigation, data curation, visualization, writing—original draft, writing—review and editing, supervision, and project administration. DK contributed to formal analysis, data curation, writing—original draft, and writing—review and editing. JW contributed to conceptualization, methodology, writing—review and editing, and funding acquisition.

Funding This work was also supported by the Agriculture and Food Research Initiative grant no. 2023–67012-39485 from the USDA National Institute of Food and Agriculture as well as the USDA-ARS. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of USDA, DOE, or ORAU/ORISE.

Declarations

Conflict of interests The authors declare no conflict of interest.

References

Adunola P, Tavares Flores E, Riva-Souza EM, Ferrão MAG, Senra JFB, Comério M, Espindula MC et al (2024) A comparison of genomic and phenomic selection methods for yield prediction in *Coffea canephora*. Plant Phenom J 7:1. https://doi.org/10.1002/ppj2.20109

Ahmed Z, Wan S, Zhang F, Zhong W (2024) Artificial intelligence for omics data analysis. BMC Methods 1:4. https://doi.org/10.1186/s44330-024-00004-5

Alemu A, Åstrand J, Montesinos-López OA, Y 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 17:1453–1467. https://doi.org/10.1016/j.molp.2024.03.007

Alzubaidi L, Khamael AD, Salhi A, Alammar Z, Fadhel MA, Albahri AS, Gu Y (2024) Comprehensive review of deep learning in orthopaedics: applications, challenges, trustworthiness, and fusion. Artif Intell Med 155:102935. https://doi.org/10.1016/j.artmed.2024.102935

Aziz MA, Masmoudi K (2024) Molecular breakthroughs in modern plant breeding techniques. Horticult Plant J 10:123–134. https://doi.org/10.1016/j.hpj.2024.01.004

Azrai M, Aqil M, Andayani NN, Efendi R, Jihad M, Zainuddin B, Muliadi A, Yasin M, Hannan MF, Syam A (2024) Optimizing ensembles machine learning, genetic algorithms, and multivariate modeling for enhanced prediction of maize yield and stress tolerance index. Front Sustain Food Syst 8:1334421. https://doi.org/10.3389/fsufs.2024.1334421

Babadi FS, Razaghi-Moghadam Z, Zare-Mirakabad F, Nikoloski Z (2023) Prediction of metabolite–protein interactions based on integration of machine learning and constraint-based modeling. Bioinform Adv. https://doi.org/10.1093/bioadv/vbad098

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 Genomics 25:418. https://doi.org/10.1186/s12864-024-10258-6

Bhat JA, Feng X, Mir ZA, Raina A, Siddique KM (2023) Recent advances in artificial intelligence, mechanistic models, and speed breeding offer exciting opportunities for precise and accelerated genomics-assisted breeding. Physiol Plant. https://doi.org/10.1111/ppl.13969

Bohra A, Kilian B, Sivasankar S, Caccamo M, Mba C, McCouch SR, Varshney RK (2022) Reap the crop wild relatives for breeding future crops. Trends Biotechnol 40:412–431. https://doi.org/10.1016/j.tibtech.2021.08.009

Bose S, Banerjee S, Kumar S, Saha A, Nandy D, Hazra S (2024) Review of applications of artificial intelligence (AI) methods in crop research. J Appl Genet 65:225–240. https://doi.org/10.1007/s13353-023-00826-z

Breiman L (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16:199–231. https://doi.org/10.1214/ss/1009213726

Brown AV, Conners SI, Huang W, Wilkey AP, Grant D, Weeks NT, Cannon SB, Graham MA, Nelson RT (2021) A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. https://doi.org/10.1093/nar/gkaa1107

Bucksch A, Burridge J, York LM, Das A, Nord E, Weitz JS, Lynch JP (2014) Image-based high-throughput field phenotyping of crop roots. Plant Physiol 166:470–486. https://doi.org/10.1104/pp.114.243519

Cassan O, Lecellier C, Martin A, Bréhélin L, Lèbre S (2024) Optimizing data integration improves gene regulatory network inference in *Arabidopsis thaliana*. Bioinformatics. https://doi.org/10.1093/bioinformatics/btae415

Cembrowska-Lech D, Krzemińska A, Miller T, Nowakowska A, Adamski C, Radaczyńska M, Mikiciuk G, Mikiciuk M (2023) An integrated multi-omics and artificial intelligence framework for advanced plant phenotyping in horticulture. Biol 12:10. https://doi.org/10.3390/biology12101298

Chen Y, Guo Y, Guan P, Wang Y, Wang X, Wang Z, Qin Z, Ma S, Xin M, Hu Z, Yao Y (2023) A wheat integrative regulatory network from large-scale complementary functional datasets enables trait-associated gene discovery for crop improvement. Mol Plant 16:393–414. https://doi.org/10.1016/j.molp.2022.12.019

Chen L, Liu G, Zhang T (2024) Integrating machine learning and genome editing for crop improvement. aBIOTECH 5:262–277. https://doi.org/10.1007/s42994-023-00133-5

Cheng C, Li Y, Varala K, Bubert J, Huang J, Kim GJ, Halim J, Arp J, Shih H, Levinson G, Park SH, Cho HY, Moose SP, Coruzzi GM (2021) Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships. Nat Commun 12:1–15. https://doi.org/10.1038/s41467-021-25893-w

Cheng Z, Zheng Q, Shi J, He Y, Yang X, Huang X, Wu L, Xu J (2023) Metagenomic and machine learning-aided identification of biomarkers driving distinctive Cd accumulation features in the root-associated microbiome of two rice cultivars. ISME Commun 3:1–13. https://doi.org/10.1038/s43705-023-00213-z

Cobb JN, Juma RU, Biswas PS, Arbelaez JD, Rutkoski J, Atlin G, Hagen T, Quinn M, Ng EH (2019) Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. Theor Appl Genet 132:627–645. https://doi.org/10.1007/s00122-019-03317-0

Cook JP, McMullen MD, Holland JB, Tian F, Bradbury P, Buckler ES, Smith A (2012) Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant Physiol 158:824–834. https://doi.org/10.1104/pp.111.185033

Cooper J, Du C, Beaver Z, Zheng M, Page R, Wodarek JR, Matny O, Szinyei T, Quiñones A, Anderson JA, Smith KP, Yang C, Steffenson BJ, Hirsch CD (2023) An RGB-based deep neural network for high-fidelity Fusarium head blight phenotyping in wheat. bioRxiv 2023:20230920.558703. https://doi.org/10.1101/2023.09.20.558703

Crossa J, Fritsche-Neto R, Montesinos-López OA, Costa-Neto G, Dreisigacker S, Montesinos-López A, Bentley AR (2021) The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. Front Plant Sci 12:651480. https://doi.org/10.3389/fpls.2021.651480

Dara R, Hazrati Fard SM, Kaur J (2022) Recommendations for ethical and responsible use of artificial intelligence in digital agriculture. Front Artif Intell 5:884192. https://doi.org/10.3389/frai.2022.884192

de Castro AI, Rallo P, Suárez MP, Casanova LM, Jiménez MR (2019) High-throughput system for the early quantification of major architectural traits in olive breeding trials using UAV images and OBIA techniques. Front Plant Sci 10:01472. https://doi.org/10.3389/fpls.2019.01472

DeChant C, Wiesner-Hanks T, Chen S, Stewart EL, Yosinski J, Gore MA, Nelson RJ, Lipson H (2017) Automated identification of northern leaf blight-infected maize plants from field imagery using deep learning. Phytopathology 107:1426–1432. https://doi.org/10.1094/PHYTO-11-16-0417-R

Deng CH, Naithani S, Kumari S, Cobo-Simón I, Quezada-Rodríguez EH, Skrabisova M, Gladman N, Correll MJ et al (2023) Genotype and phenotype data standardization, utilization and integration in the big data era for agricultural sciences. Database. https://doi.org/10.1093/database/baad088

DeSalvio AJ, Adak A, Murray SC, Jarquín D, Winans ND, Crozier D, Rooney WL (2024) Near-infrared reflectance spectroscopy phenomic prediction can perform similarly to genomic prediction of maize agronomic traits across environments. Plant Genome 17:20454. https://doi.org/10.1002/tpg2.20454

Elmarakeby HA, Hwang J, Arafeh R, Crowdis J, Gang S, Liu D, AlDubayan SH et al (2021) Biologically informed deep neural network for prostate cancer discovery. Nature 598:348–352. https://doi.org/10.1038/s41586-021-03922-4

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

Fradgley N, Gardner KA, Bentley AR, Howell P, Mackay IJ, Scott MF, Mott R, Cockram J (2023) Multi-trait ensemble genomic prediction and simulations of recurrent selection highlight importance of complex trait genetic architecture for long-term genetic gains in wheat. In Silico Plants. https://doi.org/10.1093/insilicoplants/diad002

Fu J, Hao Y, Li H, Reif JC, Chen S, Huang C, Wang G, Li X, Xu Y, Li L (2022) Integration of genomic selection with doubled-haploid evaluation in hybrid breeding: from GS 1.0 to GS 4.0 and beyond. Mol Plant 15:577–580. https://doi.org/10.1016/j.molp.2022.02.005

Gazestani VH, Lewis NE (2019) From genotype to phenotype: augmenting deep learning with networks and systems biology. Curr Opin Syst Biol 15:68–73. https://doi.org/10.1016/j.coisb.2019.04.001

Gebresenbet G, Bosona T, Patterson D, Persson H, Fischer B, Mandaluniz N, Chirici G, Zacepins A, Komasilovs V, Pitulac T, Nasirahmadi A (2023) A concept for application of integrated digital technologies to enhance future smart agricultural systems. Smart Agric Technol 5:100255. https://doi.org/10.1016/j.atech.2023.100255

Gill T, Gill SK, Saini DK, Chopra Y, de Koff JP, Sandhu KS (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

González-Camacho JM, Ornella L, Pérez-Rodríguez P, Gianola D, Dreisigacker S, Crossa J (2018) Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. Plant Genome 11:170104. https://doi.org/10.3835/plantgenome2017.11.0104

Grass Phylogeny Working Group III, Arthan W, Baker WJ, Barrett MD, Barrett RL et al (2024) Nuclear phylogenomics of grasses (Poaceae) supports current classification and reveals repeated reticulation. bioRxiv:2024.2005.2028.596153. https://doi.org/10.1101/2024.05.28.596153

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2(5):359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Houssein EH, Mohamed RE, Ali AA (2023) Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques. Sci Rep 13(1):1–19. https://doi.org/10.1038/s41598-023-34294-6

Interdisciplinary Plant Science Consortium (2023) Inclusive collaboration across plant physiology and genomics: now is the time! Plant Direct 7:5. https://doi.org/10.1002/pld3.493

James G, Witten D, Hastie T, Tibshirani R, Taylor J (2023) An introduction to statistical learning: with applications in Python. Springer Nature

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, Burgueño J, de los Campos G (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127:595–607. https://doi.org/10.1007/s00122-013-2243-1

Jeon D, Kang Y, Lee S, Choi S, Sung Y, Lee T, Kim C (2023) Digitalizing breeding in plants: a new trend of next-generation breeding based on genomic prediction. Front Plant Sci 14:1092584. https://doi.org/10.3389/fpls.2023.1092584

Jiang Y, Li C (2020) Convolutional neural networks for image-based high-throughput plant phenotyping: a review. Plant Phenomics. https://doi.org/10.34133/2020/4152816

Ju Y, Liu AE, Oestreich K, Wang T, Topp CN, Ju T (2024) TopoRoot+: computing whorl and soil line traits of field-excavated maize roots from CT imaging. Plant Methods 20:132. https://doi.org/10.1186/s13007-024-01240-0

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SA, Ballard AJ, Cowie A et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2

Khaki S, Pham H, Wang L (2021) Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. Sci Rep 11:11132. https://doi.org/10.1038/s41598-021-89779-z

Khan AA, Iqbal B, Jalal A, Khan KA, Al-Andal A, Khan I, Suboktagin S, Qayum A, Elboughdiri N (2024) Advanced molecular approaches for improving crop yield and quality: a review. J Plant Growth Regul 43:2091–2103. https://doi.org/10.1007/s00344-024-11253-7

Kick DR, Wallace JG, Schnable JC, Kolkman JM, Alaca B, Beissinger TM, Edwards J, et al (2023) Yield prediction through integration of genetic, environment, and management data through deep learning. G3 Genes|Genomes|Genetics https://doi.org/10.1093/g3journal/jkad006

Kick DR, Washburn JD (2023) Ensemble of best linear unbiased predictor, machine learning and deep learning models predict maize yield better than each model alone. In Silico Plants. https://doi.org/10.1093/insilicoplants/diad015

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

Kulkarni SR, Vandepoele K (2020) Inference of plant gene regulatory networks using data-driven methods: a practical overview. Biochim Biophys Acta Gene Regul Mech 1863:194447. https://doi.org/10.1016/j.bbagrm.2019.194447

Kuriakose SV, Pushker R, Hyde EM (2020) Data-driven decisions for accelerated plant breeding. In: Gosal S, Wani S (eds) Accelerated plant breeding, volume 1. Springer, Cham, pp 418–433. https://doi.org/10.1007/978-3-030-41866-3_4

Kusmec A, Zheng Z, Archontoulis S, Ganapathysubramanian B, Hu G, Wang L, Yu J, Schnable PS (2021) Interdisciplinary strategies to enable data-driven plant breeding in a changing climate. One Earth 4:372–383. https://doi.org/10.1016/j.oneear.2021.02.005

Lagergren J, Pavicic M, Chhetri HB, York LM, Hyatt D, Kainer D, Rutter EM, Flores K, Bailey-Bale J, Klein M et al (2023) Few-shot learning enables population-scale analysis of leaf traits in *Populus trichocarpa*. Plant Phenomics 5:0072. https://doi.org/10.34133/plantphenomics.0072

Lan W, Liao H, Chen Q, Zhu L, Pan Y, Chen Y (2024) DeepKEGG: a multi-omics data integration framework with biological insights for cancer recurrence prediction and biomarker discovery. Brief Bioinform. https://doi.org/10.1093/bib/bbae185

Lassoued R, Macall DM, Smyth SJ, Phillips PWB, Hesseln H (2021) Data challenges for future plant gene editing: expert opinion. Transgenic Res 30:765–780. https://doi.org/10.1007/s11248-021-00264-9

Leukel J, Zimpel T, Stumpe C (2023) Machine learning technology for early prediction of grain yield at the field scale: a systematic review. Comput Electron Agric 207:107721. https://doi.org/10.1016/j.compag.2023.107721

Li Y, Chen Y, Zhou L, You S, Deng H, Chen Y, Alseekh S, Yuan Y, Fu R, Zhang Z, Su D (2020) MicroTom metabolic network: Rewiring tomato metabolic regulatory network throughout the growth cycle. Mol Plant 13:1203–1218. https://doi.org/10.1016/j.molp.2020.06.005

Li T, Xu H, Teng S, Suo M, Bahitwa R, Xu M, Qian Y, Ramstein GP, Song B, Buckler ES, Wang H (2024) Modeling 0.6 million genes for the rational design of functional cis-regulatory variants and de novo design of cis-regulatory sequences. Proc Natl Acad of Sci USA. https://doi.org/10.1073/pnas.2319811121

Lima DC, Aviles AC, Alpers RT, McFarland BA, Kaeppler S, Ertl D, Romay MC et al (2023a) 2018–2019 field seasons of the maize genomes to fields (G2F) G x E project. BMC Genomic Data 24:29. https://doi.org/10.1186/s12863-023-01129-2

Lima DC, Aviles AC, Alpers RT, Perkins A, Schoemaker DL, Costa M, Michel KJ et al (2023b) 2020–2021 field seasons of maize G x E project within the genomes to fields initiative. BMC Res Notes 16(1):219. https://doi.org/10.1186/s13104-023-06430-y

Lima DC, Washburn JD, Varela JI, Chen Q, Gage JL, Romay MC, Holland J et al (2023c) Genomes to fields 2022 maize genotype by environment prediction competition. BMC Res Notes 16:148. https://doi.org/10.1186/s13104-023-06421-z

Lin Z, Ou-Yang L (2023) Inferring gene regulatory networks from single-cell gene expression data via deep multi-view contrastive learning. Brief Bioinform 24:1. https://doi.org/10.1093/bib/bbac586

Liu Y, Wang D, He F, Wang J, Joshi T, Xu D (2019) Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. Front Genet 10:1091. https://doi.org/10.3389/fgene.2019.01091

Lozada DN, Sandhu KS, Bhatta M (2023) Ridge regression and deep learning models for genome-wide selection of complex traits in New Mexican Chile peppers. BMC Genom Data 24:80. https://doi.org/10.1186/s12863-023-01179-6

Lyons E, Freeling M (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J 53:661–673. https://doi.org/10.1111/j.1365-313X.2007.03326.x

Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, Sharan R, Ideker T (2018a) Using deep learning to model the hierarchical structure and function of a cell. Nat Methods 15:290–298. https://doi.org/10.1038/nmeth.4627

Ma W, Qiu Z, Song J, Li J, Cheng Q, Zhai J, Ma C (2018b) A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta 248:1307–1318. https://doi.org/10.1007/s00425-018-2976-9

Maldonado C, Mora-Poblete F, Contreras-Soto RI, Ahmar S, Chen J-T, Amaral Júnior AT, Scapim CA (2020) Genome-wide prediction of complex traits in two outcrossing plant species through deep learning and Bayesian regularized neural network. Front Plant Sci 11:593897. https://doi.org/10.3389/fpls.2020.593897

Måløy H, Windju S, Bergersen S, Alsheikh M, Downing KL (2021) Multimodal performers for genomic selection and crop yield prediction. Smart Agric Technol 1:100017. https://doi.org/10.1016/j.atech.2021.100017

Manickam S, Rajagopalan VR, Kambale R, Rajasekaran R, Kanagarajan S, Muthurajan R (2023) Plant metabolomics: Current initiatives and future prospects. Curr Issues Mol Biol 45:8894–8906. https://doi.org/10.3390/cimb45110558

McFarland BA, AlKhalifah N, Bohn M, Bubert J, Buckler ES, Ciampitti I, Edwards J et al (2020) Maize genomes to fields (G2F): 2014–2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets. BMC Res Notes 13:71. https://doi.org/10.1186/s13104-020-4922-8

Miller ND, Haase NJ, Lee J, Kaeppler SM, de Leon N, Spalding EP (2017) A robust, high-throughput method for computing maize ear, cob, and kernel attributes automatically from images. Plant J 89:169–178. https://doi.org/10.1111/tpj.13320

Mishra B, Kumar N, Mukhtar MS (2022) A rice protein interaction network reveals high centrality nodes and candidate pathogen effector targets. Comput Struct Biotechnol J 20:2001–2012. https://doi.org/10.1016/j.csbj.2022.04.027

Mitsanis C, Hurst W, Tekinerdogan B (2024) A 3D functional plant modelling framework for agricultural digital twins. Comput Electron Agric 218:108733. https://doi.org/10.1016/j.compag.2024.108733

Moeinizade S, Pham H, Han Y, Dobbels A, Hu G (2022) An applied deep learning approach for estimating soybean relative maturity from UAV imagery to aid plant breeding decisions. Mach Learn Appl 7:100233. https://doi.org/10.1016/j.mlwa.2021.100233

Montesinos-López OA, Montesinos-López A, Tuberosa R, Maccaferri M, Sciara G, Ammar K, Crossa J (2019) Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods. Front Plant Sci 10:1311. https://doi.org/10.3389/fpls.2019.01311

Montesinos-López OA, Montesinos-López A, Hernandez-Suarez CM, Barrón-López JA, Crossa J (2021a) Deep-learning power and perspectives for genomic selection. Plant Genome 14:e20122. https://doi.org/10.1002/tpg2.20122

Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Barrón-López JA, Martini JWR, Fajardo-Flores SB, Gaytan-Lugo LS, Santana-Mancilla PC, Crossa J (2021b) A review of deep learning applications for genomic selection. BMC Genomics 22:19. https://doi.org/10.1186/s12864-020-07319-x

Montesinos-López OA, Herr AW, Crossa J, Montesinos-López A, Carter AH (2024) Enhancing winter wheat prediction with genomics, phenomics and environmental data. BMC Genomics 25:544. https://doi.org/10.1186/s12864-024-10438-4

Moore BM, Wang P, Fan P, Lee A, Leong B, Lou YR, Schenck CA, Sugimoto K, Last R, Lehti-Shiu MD, Barry CS, Shiu SH (2020) Within- and cross-species predictions of plant specialized metabolism genes using transfer learning. In Silico Plants. https://doi.org/10.1093/insilicoplants/diaa005

Morales N, Ogbonna AC, Ellerbrock BJ, Bauchet GJ, Tantikanjana T, Tecle IY, Powell AF, Lyon D, Menda N, et al (2022) Breedbase: a digital ecosystem for modern plant breeding. G3 Genes|Genomes|Genetics https://doi.org/10.1093/g3journal/jkac078

Murphy KM, Ludwig E, Gutierrez J, Gehan MA (2024) Deep learning in image-based plant phenotyping. Annu Rev Plant Biol 75:771–795. https://doi.org/10.1146/annurev-arplant-070523-042828

Murray SC, Sharma A, Rooney WL, Klein PE, Mullet JE, Mitchell SE, Kresovich S (2008) Genetic improvement of sorghum as a biofuel feedstock: I. QTL for stem sugar and grain nonstructural carbohydrates. Crop Sci 48:2165–2179. https://doi.org/10.2135/cropsci2008.01.0016

Mushtaq MA, Ahmed HG, Zeng Y (2024) Applications of artificial intelligence in wheat breeding for sustainable food security. Sustainability 16:5688. https://doi.org/10.3390/su16135688

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 13:1–12. https://doi.org/10.1038/s41598-023-42984-4

Nazzicari N, Nelson R, Biscarini F (2022) Stacked kinship CNN vs. GBLUP for genomic predictions of additive and complex continuous phenotypes. Sci Rep 12:19889. https://doi.org/10.1038/s41598-022-24405-0

Negus KL, Li X, Welch SM, Yu J (2024) The role of artificial intelligence in crop improvement. Adv Agron 184:1–66. https://doi.org/10.1016/bs.agron.2023.11.001

Niu Y, Wu L, Li Y, Huang H, Qian M, Sun W, Zhu H, Xu Y, Fan Y, Mahmood U, Xu B (2020) Deciphering the transcriptional regulatory networks that control size, color, and oil content in *Brassica rapa* seeds. Biotechnol Biofuels 13:1–20. https://doi.org/10.1186/s13068-020-01728-6

Nizamani MM, Zhang Q, Muhae-Ud-Din G, Wang Y (2023) High-throughput sequencing in plant disease management: a comprehensive review of benefits, challenges, and future perspectives. Phytopathol Res 5:44. https://doi.org/10.1186/s42483-023-00199-5

Pan J, Zhu Y, Li L, Huang W, Guo J, Yu C, Wang L, Zhao Z (2022) DWPPI: a deep learning approach for predicting protein–protein interactions in plants based on multi-source information with a large-scale biological network. Front Bioeng Biotechnol 10:807522. https://doi.org/10.3389/fbioe.2022.807522

Pardo J, Wai CM, Harman M, Nguyen A, Kremling KA, Romay MC, Lepak N, Bauerle TL, Buckler ES, Thompson AM, VanBuren R (2023) Cross-species predictive modeling reveals conserved drought responses between maize and sorghum. Proc Natl Acad Sci USA. https://doi.org/10.1073/pnas.2216894120

Perez-Riverol Y, Bai M, da Veiga LF, Squizzato S, Park YM, Haug K, Carroll AJ, Spalding D, Paschall J, Wang M, Del-Toro N (2017) Discovering and linking public omics data sets using the Omics Discovery Index. Nat Biotechnol 35:406–409. https://doi.org/10.1038/nbt.3790

Pommier C, Michotey C, Cornut G, Roumet P, Duchêne E, Flores R, Lebreton A, Alaux M, Durand S, Kimmel E et al (2019) Applying FAIR principles to plant phenotypic data management in GnpIS. Plant Phenomics. https://doi.org/10.34133/2019/1671403

Qin Q, Feng J (2017) Imputation for transcription factor binding predictions based on deep learning. PLoS Comput Biol 13(2):e1005403. https://doi.org/10.1371/journal.pcbi.1005403

Rai KK (2022) Integrating speed breeding with artificial intelligence for developing climate-smart crops. Mol Biol Rep 49:11385–11402. https://doi.org/10.1007/s11033-022-07769-4

Ramstein GP, Jensen SE, Buckler ES (2019) Breaking the curse of dimensionality to identify causal variants in Breeding 4. Theor Appl Genet 132:559–567. https://doi.org/10.1007/s00122-018-3267-3

Raza A, Salehi H, Bashir S, Tabassum J, Jamla M, Charagh S, Barmukh R, Mir RA, Bhat BA, Javed MA, Guan DX, Mir RR, Siddique KHM, Varshney RK (2024) Transcriptomics, proteomics, and metabolomics interventions prompt crop improvement against metal(loid) toxicity. Plant Cell Rep 43:80. https://doi.org/10.1007/s00299-024-03153-7

Ren Y, Wu C, Zhou H, Hu X, Miao Z (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

Resende RT, Hickey L, Amaral CH, Peixoto LL, Marcatti GE, Yunbi X (2024) Satellite-enabled enviromics to enhance crop improvement. Mol Plant 17:848–866. https://doi.org/10.1016/j.molp.2024.04.005Rincent

Rincent R, Charpentier J-P, Faivre-Rampant P, Paux E, Le Gouis J, Bastien C, Segura V (2018) Phenomic selection is a low-cost and high-throughput method based on indirect predictions: proof of concept on wheat and poplar. G3 Genes|Genomes|Genetics 8:3961–3972. https://doi.org/10.1534/g3.118.200760

Robert P, Goudemand E, Auzanneau J, Oury FX, Rolland B, Heumez E, Bouchet S et al (2022) Phenomic selection in wheat breeding: prediction of the genotype-by-environment interaction in multi-environment breeding trials. Theor Appl Genet 135:3337–3356. https://doi.org/10.1007/s00122-022-04170-4

Rogers AR, Dunne JC, Romay C, Bohn M, Buckler ES, Ciampitti IA, Edwards J, et al (2021) The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. G3 Genes|Genomes|Genetics 11: jkaa050. https://doi.org/10.1093/g3journal/jkaa050

Salehi B, Mireei SA, Jafari M, Hemmat A, Majidi MM (2024) Integrating in-field Vis-NIR leaf spectroscopy and deep learning feature extraction for growth-stage dependent and independent genotyping of wheat plants. Biosyst Eng 238:188–199. https://doi.org/10.1016/j.biosystemseng.2024.01.016

Sandhu K, Patil SS, Pumphrey M, Carter A (2021) Multitrait machine- and deep-learning models for genomic selection using spectral information in a wheat breeding program. Plant Genome. https://doi.org/10.1002/tpg2.20119

Sangjan W, Carpenter-Boggs LA, Hudson TD, Sankaran S (2022a) Pasture productivity assessment under mob grazing and fertility management using satellite and UAS imagery. Drones 6:232. https://doi.org/10.3390/drones6090232

Sangjan W, Pukrongta N, Carter AH, Pumphrey MO, Sankaran S (2022b) Development of IoT-based camera system for automated in-field monitoring to support crop breeding programs. ESS Open Arch. https://doi.org/10.22541/au.166758437.70063358/v1

Sangjan W, McGee RJ, Sankaran S (2023) Evaluation of forage quality in a pea breeding program using a hyperspectral sensing system. Comput Electron Agric 212:108052. https://doi.org/10.1016/j.compag.2023.108052

Sangjan W, Carter AH, Pumphrey MO, Hagemeyer K, Jitkov V, Sankaran S (2024) Effect of high-resolution satellite and UAV imagery plot pixel resolution in wheat crop yield prediction. Int J Remote Sens 45:1678–1698. https://doi.org/10.1080/01431161.2024.2313997

Seck F, Covarrubias-Pazaran G, Gueye T, Bartholomé J (2023) Realized genetic gain in rice: achievements from breeding programs. Rice 16:61. https://doi.org/10.1186/s12284-023-00677-6

Seethepalli A, Dhakal K, Griffiths M, Guo H, Freschet GT, York LM (2021) RhizoVision Explorer: open-source software for root image analysis and measurement standardization. AoB Plants. https://doi.org/10.1093/aobpla/plab056

Selby P, Abbeloos R, Backlund JE, Salido MB, Bauchet G, Benites-Alfaro OE, Birkett C, Calaminos VC, Carceller P et al (2019) BrAPI—An application programming interface for plant breeding applications. Bioinformatics 35(20):4147–4155. https://doi.org/10.1093/bioinformatics/btz190

Sharma S, Partap A, de Luis Balaguer MA, Malvar S, Chandra R (2022) DeepG2P: Fusing multi-modal data to improve crop production. arXiv. https://arxiv.org/abs/2211.05986

Sheikh M, Iqra F, Ambreen H, Pravin KA, Ikra M, Chung YS (2024) Integrating artificial intelligence and high-throughput phenotyping for crop improvement. J Integr Agric 23:1787–1802. https://doi.org/10.1016/j.jia.2023.10.019

Shi L, Marti Ferrando T, Landeo Villanueva S, Joosten MH, Vleeshouwers VG, Bachem CW (2023) Protocol to identify protein-protein interaction networks in *Solanum tuberosum* using transient TurboID-based proximity labeling. STAR Protoc 4:102577. https://doi.org/10.1016/j.xpro.2023.102577

Sinha D, Maurya AK, Abdi G, Majeed M, Agarwal R, Mukherjee R, Ganguly S, Aziz R, Bhatia M, Majgaonkar A, Seal S, Das M, Banerjee S, Chowdhury S, Adeyemi SB, Chen T (2023) Integrated genomic selection for accelerating breeding programs of climate-smart cereals. Genes 14:731–738. https://doi.org/10.3390/genes14071484

Tavis S, Hettich RL (2024) Multi-omics integration can be used to rescue metabolic information for some of the dark region of the Pseudomonas putida proteome. BMC Genomics 25:267. https://doi.org/10.1186/s12864-024-10082-y

Thapa S, Gill HS, Halder J, Rana A, Ali S, Maimaitijiang M, Gill U et al (2024) Integrating genomics, phenomics, and deep learning improves the predictive ability for Fusarium head blight–related traits in winter wheat. Plant Genome. https://doi.org/10.1002/tpg2.20470

Thompson AL, Thorp KR, Conley MM, Roybal M, Moller D, Long JC (2020) A data workflow to support plant breeding decisions from a terrestrial field-based high-throughput plant phenotyping system. Plant Methods 16(1):97. https://doi.org/10.1186/s13007-020-00639-9

Togninalli M, Wang X, Kucera T, Shrestha S, Juliana P, Mondal S, Pinto F, Govindan V, Singh RP, Borgwardt K, Poland J (2023) Multi-modal deep learning improves grain yield prediction in wheat breeding by fusing genomics and phenomics. Bioinformatics. https://doi.org/10.1093/bioinformatics/btad336

Tong H, Nikoloski Z (2021) Machine learning approaches for crop improvement: leveraging phenotypic and genotypic big data. J Plant Physiol 257:153354. https://doi.org/10.1016/j.jplph.2020.153354

Trippa D, Scalenghe R, Basso MF, Panno S, Davino S, Morone C, Giovino A, Oufensou S, Luchi N, Yousefi S, Martinelli F (2024) Next-generation methods for early disease detection in crops. Pest Manag Sci 80:245–261. https://doi.org/10.1002/ps.7733

Tyagi A, Mir ZA, Almalki MA, Deshmukh R, Ali S (2024) Genomics-assisted breeding: A powerful breeding approach for improving plant growth and stress resilience. Agronomy 14:1128. https://doi.org/10.3390/agronomy14061128

Ubbens J, Stavness I, Sharpe AG (2023) GPFN: Prior-data fitted networks for genomic prediction. bioRxiv: 2023.09.20.558648. https://doi.org/10.1101/2023.09.20.558648

van Hilten A, Kushner SA, Kayser M, Ikram MA, Adams HHH, Klaver CCW, Niessen WJ, Roshchupkin GV (2021) GenNet framework: Interpretable deep learning for predicting phenotypes from genetic data. Commun Biol 4:1094. https://doi.org/10.1038/s42003-021-02622-z

Volk GM, Byrne PF, Coyne CJ, Flint-Garcia S, Reeves PA, Richards C (2021) Integrating genomic and phenomic approaches to support plant genetic resources conservation and use. Plants 10(11):2260. https://doi.org/10.3390/plants10112260

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 K, Abid MA, Rasheed A, Crossa J, Hearne S, Li H (2023a) 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 X, Han L, Li J, Shang X, Liu Q, Li L, Zhang H (2023b) Next-generation bulked segregant analysis for Breeding 4.0. Cell Rep 42:9. https://doi.org/10.1016/j.celrep.2023.113039

Wang N, Cao H, Huang X, Ding M (2024a) Rapeseed flower counting method based on GhP2-YOLO and StrongSORT algorithm. Plants 13:2388. https://doi.org/10.3390/plants13172388

Wang XY, Ren CX, Fan QW, Xu YP, Wang LW, Mao ZL, Cai XZ (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 25:6932. https://doi.org/10.3390/ijms25136932

Wang AX, Tran C, Desai N, Lobell D, Ermon S (2018) Deep transfer learning for crop yield prediction with remote sensing data. In: Proceedings of the 1st ACM SIGCAS Conf Comput Sustainable Soc (COMPASS'18) 5:1–5. https://doi.org/10.1145/3209811.3212707

Washburn JD, Whitmire DK, Murray SC, Burson BL, Wickersham TA, Heitholt JJ, Jessup RW (2013) Estimation of rhizome composition and overwintering ability in perennial *Sorghum* spp. using near-infrared spectroscopy (NIRS). Bioenerg Res 6:822–829. https://doi.org/10.1007/s12155-013-9305-8

Washburn JD, Schnable JC, Conant GC, Brutnell TP, Shao Y, Zhang Y, Ludwig M, Davidse G, Pires JC (2017) Genome-guided phylotranscriptomic methods and the nuclear phylogenetic tree of the Paniceae grasses. Sci Rep 7:13528. https://doi.org/10.1038/s41598-017-13236-z

Washburn JD, Katherine M, Ramstein G, Kremling KA, Valluru R, Buckler ES, Wang H (2019) Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. Proc Natl Acad Sci USA 116:5542–5549. https://doi.org/10.1073/pnas.1814551116

Washburn JD, Burch MB, Valdes Franco JA (2020) Predictive breeding for maize: making use of molecular phenotypes, machine learning, and physiological crop models. Crop Sci 60:622–638. https://doi.org/10.1002/csc2.20052

Washburn JD, Cimen E, Ramstein G, Reeves T, O'Briant P, McLean G, Cooper M, Hammer G, Buckler ES (2021) Predicting phenotypes from genetic, environment, management, and historical data using CNNs. Theor Appl Genet 134:3997–4011. https://doi.org/10.1007/s00122-021-03943-7

Washburn JD, LaFond HF, Lapadatescu MC, Pereira AE, Erb M, Hibbard BE (2023) GWAS analysis of maize host plant resistance to western corn rootworm (Coleoptera: Chrysomelidae) reveals candidate small effect loci for resistance breeding. J Econ Entomol 116:2184–2192. https://doi.org/10.1093/jee/toad181

Washburn JD, Varela JI, Xavier A, Chen Q, Ertl D, Gage JL, Holland JB, Lima DC, Romay MC, Lopez-Cruz M et al (2024) Global genotype by environment prediction competition reveals that diverse modeling strategies can deliver satisfactory maize yield estimates. bioRxiv.2024.2009.2013.612969. https://doi.org/10.1101/2024.09.13.612969

Westhues CC, Mahone GS, da Silva S, Thorwarth P, Schmidt M, Richter J-C, Simianer H, Beissinger TM (2021) Prediction of maize phenotypic traits with genomic and environmental predictors using gradient boosting frameworks. Front Plant Sci 12:699589. https://doi.org/10.3389/fpls.2021.699589

Winn ZJ, Amsberry AL, Haley SD, DeWitt ND, Mason RE (2023) Phenomic versus genomic prediction—A comparison of prediction accuracies for grain yield in hard winter wheat lines. Plant Phenome J. https://doi.org/10.1002/ppj2.20084

Wu H, Wiesner-Hanks T, Stewart EL, DeChant C, Kaczmar N, Gore MA, Nelson RJ, Lipson H (2019) Autonomous detection of plant disease symptoms directly from aerial imagery. Plant Phenome J 2(1):1–9. https://doi.org/10.2135/tppj2019.03.0006

Wu C, Luo J, Xiao Y (2024) Multi-omics assists genomic prediction of maize yield with machine learning approaches. Mol Breeding 44:14. https://doi.org/10.1007/s11032-024-01454-z

Xu Y, Liu X, Cao X, Huang C, Liu E, Qian S, Liu X, Wu Y, Dong F, Qiu C, Qiu J, Hua K, Su W, Wu J, Xu H, Han Y, Fu C, Yin Z, Liu M, Zhang J (2021) Artificial intelligence: a powerful paradigm for scientific research. The Innovation 2:100179. https://doi.org/10.1016/j.xinn.2021.100179

Xu Y, Zhang X, Li H, Zheng H, Zhang J, Olsen MS, Varshney RK, Prasanna BM, Qian Q (2022) Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. Mol Plant 15:1664–1695. https://doi.org/10.1016/j.molp.2022.09.001

Yan K, Guo X, Ji Z, Zhou X (2021) Deep transfer learning for cross-species plant disease diagnosis adapting mixed subdomains. IEEE/ACM Trans Comput Biol Bioinform 20:2555–2564. https://doi.org/10.1109/TCBB.2021.3135882

Yang W, Guo T, Luo J, Zhang R, Zhao J, Warburton ML, Xiao Y, Yan J (2022) Target-oriented prioritization: targeted selection strategy by integrating organismal and molecular traits through predictive analytics in breeding. Genome Biol 23:80. https://doi.org/10.1186/s13059-022-02650-w

Yang Z, Luo C, Pei X, Wang S, Huang Y, Li J, Liu B, Kong F, Yang QY, Fang C (2024) SoyMD: A platform combining multi-omics data with various tools for soybean research and breeding. Nucleic Acids Res. https://doi.org/10.1093/nar/gkad786

Yao Q, Zheng X, Zhou G, Zhang J (2024) SGR-YOLO: A method for detecting seed germination rate in wild rice. Front Plant Sci 14:1305081. https://doi.org/10.3389/fpls.2023.1305081

Yin Q, Wu T, Gao R, Wu L, Shi Y, Wang X, Wang M, Xu Z, Zhao Y, Su X, Su Y, Han X, Yuan L, Xiang L, Chen S (2023) Multi-omics reveal key enzymes involved in the formation of phenylpropanoid glucosides in *Artemisia annua*. Plant Physiol Biochem 201:107795. https://doi.org/10.1016/j.plaphy.2023.107795

Yoosefzadeh-Najafabadi M, Hesami M, Eskandari M (2023) Machine learning-assisted approaches in modernized plant breeding programs. Genes 14:777. https://doi.org/10.3390/genes14040777

Yunbi X, Zhang X, Li H, Zheng H, Zhang J, Olsen MS, Varshney RK, Prasanna BM, Qian Q (2022) Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. Mol Plant 15:1664–1695. https://doi.org/10.1016/j.molp.2022.09.001

Zeng S, Mao Z, Ren Y, Wang D, Xu D, Joshi T (2021) G2PDeep: a web-based deep-learning framework for quantitative phenotype prediction and discovery of genomic markers. Nucleic Acids Res. https://doi.org/10.1093/nar/gkab407

Zhai J, Gokaslan A, Schiff Y, Berthel A, Liu ZY, Miller ZR, Scheben A, Stitzer MC, Romay MC, Buckler ES, Kuleshov V (2024) Cross-species modeling of plant genomes at single nucleotide resolution using a pre-trained DNA language model. bioRxiv 2024.06.04.596709. https://doi.org/10.1101/2024.06.04.596709

Zhang P, Li D (2022) EPSA-YOLO-V5s: A novel method for detecting the survival rate of rapeseed in a plant factory based on multiple guarantee mechanisms. Comput Electron Agric 193:106714. https://doi.org/10.1016/j.compag.2022.106714

Zhang F, Song H, Zeng M, Li Y, Kurgan L, Li M (2019) DeepFunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions. Proteomics 19:1900019. https://doi.org/10.1002/pmic.201900019

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 E, Dong L, Zhao H, Zhang H, Zhang T, Yuan S, Jiao J et al (2023) A relationship prediction method for *Magnaporthe oryzae*–rice multi-omics data based on WGCNA and graph autoencoder. J Fungi 9:1007. https://doi.org/10.3390/jof9101007

Zhao T, Wang F, Mott R, Dekkers J, Cheng H (2024) Using encrypted genotypes and phenotypes for collaborative genomic analyses to maintain data confidentiality. Genetics. https://doi.org/10.1093/genetics/iyad210

Zhou DX (2020) Universality of deep convolutional neural networks. Appl Comput Harmon Anal 48:787–794. https://doi.org/10.1016/j.acha.2019.06.004

Zhu X, Leiser WL, Hahn V, Würschum T (2021) Phenomic selection is competitive with genomic selection for breeding of complex traits. Plant Phenome J 4(1):e20027. https://doi.org/10.1002/ppj2.20027

Zulfiqar M, Singh V, Steinbeck C, Sorokina M (2024) Review on computer-assisted biosynthetic capacities elucidation to assess metabolic interactions and communication within microbial communities. Crit Rev Microbiol. https://doi.org/10.1080/1040841X.2024.2306465