BMC Bioinformatics

RESEARCH Open Access

Machine learning for multi-omics data integration in crop improvement: a systematic review

Alemu Tsega[1*] and Destaw Mullualem[1]

*Correspondence:
Alemu Tsega
alexttsega@gmail.com
[1]Department of Biology, Natural and computational science, Injibara University, Injibara, Ethiopia

Abstract

Background This systematic review synthesizes applications of machine learning (ML) for multi-omics data integration in crop improvement, evaluating its dual potential to enhance predictive accuracy for selection (breeding utility) and to generate interpretable biological insights (mechanistic discovery).

Methods Following a systematic review of 76 eligible studies, we synthesized patterns in methodological adoption.

Results The integration of environmental data (envirotyping) with multiple omics layers for genotype-by-environment (G×E) prediction was identified as a major emerging frontier, though currently addressed in fewer than 20% of studies. Tree-based models such as random forest and XGBoost were the most prevalent, favored for their interpretability and robustness with small to medium-sized datasets. In contrast, deep learning approaches, while reporting high performance, were primarily applied to larger datasets and constrained by higher computational costs. Emerging hybrid models show promise, but their efficacy is highly architecture- dependent. The most consistent accuracy gains (10–15%) were observed for feature-engineering hybrids (autoencoders compressing multi-omics data followed by XGBoost). Stacking ensembles showed more variable performance (5–12% gains), while integrated hybrids like convolutional neural network-long short-term memory (CNN-LSTM) delivered high accuracy for specific data structures. A recurring trend indicated that genomics with transcriptomics frequently boosted prediction for stress-related traits, while genomics with metabolomics excelled for quality traits. Tri-omics integration enhanced prediction for complex yield traits, though with marginal gains (< 5%) and substantial computational cost increases. Comparatively, ML-based approaches often outperformed classical genomic selection (GS) for low-heritability traits, while GS remained competitive for high-heritability traits. Deep learning models showed particular strength in handling population structure, reducing prediction errors by up to 20% in diverse panels. Critical gaps were identified: an overwhelming focus on point estimates of accuracy, (with fewer than 10% of studies reporting calibrated uncertainty metrics: and a relative scarcity of intrinsically interpretable model architectures that incorporates biological constraints as a core design principle.

Conclusion ML- driven multi-omics integration holds transformative potential but requires strategic implementation tailored to specific breeding objectives, trait architecture, and resource availability. Collective efforts to standardize data protocols, enhance model interpretability, and democratize computational tools are critical. Realizing equitable potential requires strategies to develop user-friendly platforms that extend advances to under-resourced crops. Transfer learning from data-rich species and federated data-sharing models are concrete avenues for promoting equitable innovations and enhancing global agricultural resilience.

Keywords Crop breeding, Deep learning, Explainable AI (XAI), Genotype-by-environment (G×E) interactions, Genomic prediction, Multi-omics data integration, Machine learning, Precision agriculture, Plant phenomics, Tansfer learning

Background

The increasing global demand for food, coupled with the challenges posed by climate change, necessitates the development of high-yielding, stress-resistant, and nutritionally enhanced crop varieties [1]. Traditional plant breeding methods, while foundational, are often slow and labor-intensive, requiring multiple generations of crossing and selection to achieve desired traits [2]. The advent of high-throughput genotyping and next-generation sequencing (NGS) technologies has revolutionized plant breeding by enabling genomic selection (GS) and marker-assisted breeding [3]. Recent advances in high-throughput sequencing and analytical technologies have led to the generation of vast amounts of multi-omics data (genomics, transcriptomics, proteomics, metabolomics, and phenomics) in crop science [4]. Integrating these datasets is crucial for understanding complex biological processes and improving crop traits such as yield, stress tolerance, and nutritional quality [5]. Machine learning (ML) has emerged as a powerful tool for analyzing and integrating multi-omics data, enabling predictive modeling and data-driven decision-making in crop improvement programs [6, 7]. Foundational applications of ML in genetics and genomics have established its utility for high-dimensional biological data [7], while network-based approaches have advanced multi-omics integration [8].

With the global population projected to reach 9.7 billion by 2050, agricultural systems must increase productivity by 60–70% to meet demand. Traditional breeding relies on phenotypic selection, which is time-consuming and environmentally sensitive [9]. ML-driven genomic prediction models enable early selection of superior genotypes, shortening breeding timelines and improving resource allocation [10]. There is an urgent need to accelerate plant breeding in response to global challenges such as climate change, population growth, and food security threats [11]. Traditional breeding methods are slow and inefficient for meeting these demands, while genomic-assisted breeding (GAB) combined with machine learning (ML) offers a transformative solution [12].

The integration of multi-omics data including genomics, transcriptomics, proteomics, and metabolomics presents a powerful but unresolved opportunity for accelerating crop improvement [13]. The core problem is the inherent complexity of effectively integrating these heterogeneous, high-dimensional datasets to extract biologically meaningful insights for predicting complex agronomic traits. Current challenges include managing data scale and diversity, overcoming technical noise and batch effects, and developing machine learning models that can handle non-linear relationships and provide interpretable predictions for breeding decisions. This gap between data generation and actionable knowledge hinders the full realization of precision breeding and climate-resilient

crop development [14, 15]. The explosion of genomic, phenomic, and environmental data in agriculture has created a need for advanced computational tools [4]. This systematic review synthesizes how ML integrates multi-omics data (genomics, transcriptomics, metabolomics) to uncover gene-trait relationships, providing actionable insights for breeders [16].

General objective

To systematically evaluate the application of machine learning (ML) techniques for integrating multi-omics data in crop improvement, identifying current trends, methodological strengths, limitations, and future opportunities to accelerate precision breeding and agricultural sustainability.

Specific Objectives

- To catalog and classify the dominant machine learning and deep learning architectures used for multi-omics data integration in crop plants.
- To analyze patterns in omics data combinations and their correlation with prediction accuracy for specific crop traits.
- To assess the performance of ML-driven multi-omics models against classical genomic selection methods in terms of accuracy, scalability, and robustness across diverse crops, environments, and genetic architectures.
- To evaluate data integration strategies (early fusion, late fusion, graph-based methods) and their suitability for different biological questions.
- To identify challenges in ML-based multi-omics integration, such as data heterogeneity, interpretability, computational costs and model generalizability across environments/populations.
- To propose a roadmap for future research, emphasizing explainable AI transfer learning, and scalable pipelines for breeding applications.

Review questions

- What are the dominant machine learning (ML) and deep learning (DL) architectures applied for multi-omics data integration in crop plants?
- Which combinations of omics layers show the highest predictive accuracy for specific crop traits, and what biological or technical factors explain these patterns?
- How do ML-driven multi-omics models compare to classical genomic selection methods in terms of accuracy, computational scalability, and robustness across diverse crops and environments?
- How do different data integration strategies perform for specific biological questions?
- What are the key technical and biological challenges limiting the adoption of ML-based multi-omics integration in crop improvement?
- What emerging technologies could address current limitations and enable scalable multi-omics pipelines for crop breeding?

Systematic review methodology

Data sources and search strategy

This systematic review was conducted and reported in accordance with the preferred reporting items for systematic reviews and meta-analysis(PRISMA)guidelines [17]. Searches were executed across four electronic data bases including PubMed/MEDLINE, Web of science core collection, Scopus and IEEEE Xplore. The final searches were run between March 15 and April 17, 2025. The search strategy integrated three core concepts using Boolean operators (AND/OR): The first domain encompassed machine learning and artificial intelligence, including techniques such as "machine learning" "deep learning" "neural networks", "random forests" XGboost" and "support vector machine". This intersected with the second domains, multi-omics integration, captured by terms like"multi-omics integrated omics and combination of specific omics layers (e.g., genomics AND transcriptomics). The final, limiting domain was crop plants, which includes terms such as crops and plant breeding as well as the name of key agriculture species including wheat, maize, rice and soybean. No date or language filters were applied during the initial search to maximize sensitivity. However, during the screening phases, records were limited to those published between 1997 and 2025, and only publications in English were considered for full-text review [18].

The complete, database-specific search strings are detailed below. Searches were adapted to the specific syntax of each platform.

PubMed/MEDLINE (Search Date: March 15–24, 2025).

PubMed.

("machine learning" [MeSH terms] OR "deep learning" [MeSH Terms] OR "neural networks" [MeSH terms] OR "artificial intelligence" [MeSH Terms] OR "random forest" Title/Abstract]OR "XGBoost"[Title/Abstract] OR "support vector machine" [Title/Abstract]) AND "multi omics" [Title/Abstract]OR "Multiomics"[Title Abstract]OR ("genomic*"[Title/Abstract]"AND "transcriptomics"[Title/Abstract]) OR "genomic*"[Title/Abstract] AND "metabolomics*"[Title/Abstract]) OR ("genomic*"[Title/Abstract]AND "proteomics*"[Title/Abstract])OR "integrated omics"[Title/Abstract]) AND (crops [MeSH Terms] OR "plants"[MeSH Terms] OR "crops improvements"[Title/Abstract] OR "plant breeding"[Title/Abstract] OR Wheat[Title/Abstract]OR maize[Title/abstract] OR rice[Title/Abstract] OR soybean[Title/Abstract])

Web of science core collection (Search Date: March 25-April- 4, 2025).

Web of Science:

TS= (machine learning" OR "deep learning" OR "neural network*" OR "artificial intelligence" OR random forest" OR XGBoost OR "support vector machine") AND (multi-omics" OR multiomics OR ("genomic*" NEAR/3"metabolomics") OR ("genomic*" NEAR/3 "proteomics*" OR "integrated omics") AND (crop* OR agriculture* OR wheat OR maize OR rice OR soybean)) AND PY= (1997–2025).

Scopus (Search Date: April 5–13, 2025).

Scopus:

TITLE - ABS - KEY(("machine learning" OR "neural network" OR artificial intelligence " OR "random forest" OR "xgboost" OR "support vector machine") AND ("multi-omics" OR "multiomics" OR ("genomics*" AND "transcriptomic*") OR ("genomic*") AND "metabolomics*") OR ("genomic*" AND " proteomic*") OR "integrated omics")

AND (crop* OR plant* OR agriculture* OR wheat OR maize OR rice OR soybean)) AND PUBYEAR > = 1997 AND PUBYAR < = 2025 AND (LIMIT-TO (LANGUAGE, "English")).

IEEE Xplore Search Date: April 13–17 2025.

IEEE Xplore:

("All Metadata:"machine learning" OR "All Metadata":"deep learning" OR "All Metadata": neural network" OR All Metadata":"random forest" OR "All Metadata":"XGBoost" OR All Metadata": "support vector machine") AND (All Metadata":"multi-omics" OR "All Metadata":multiomic* OR (" All Metadata":genomic* AND "All Metadata":transcriptomics*) OR(All Metadata":genomic* AND "All Metadata":metabolomics*)) AND ("All Metadata":crops* OR "All Metadata":plant* OR "All Metadata":wheat OR All Metadata":maize OR "All Metadata":rice OR "All Metadata":soybean).

Study selection and the PRISMA flow diagram

The study selection process followed the PRISMA 2020 statement [17]. All identified records were imported into a reference management software of EndNote for deduplication. The subsequent screening was conducted in two phases: 1 title and abstract screening, and 2) full-text reviews. Any discrepancies were resolved through discussion. The results of this process are summarized in the PRISMA 2020 flow diagram (Fig. 1), which

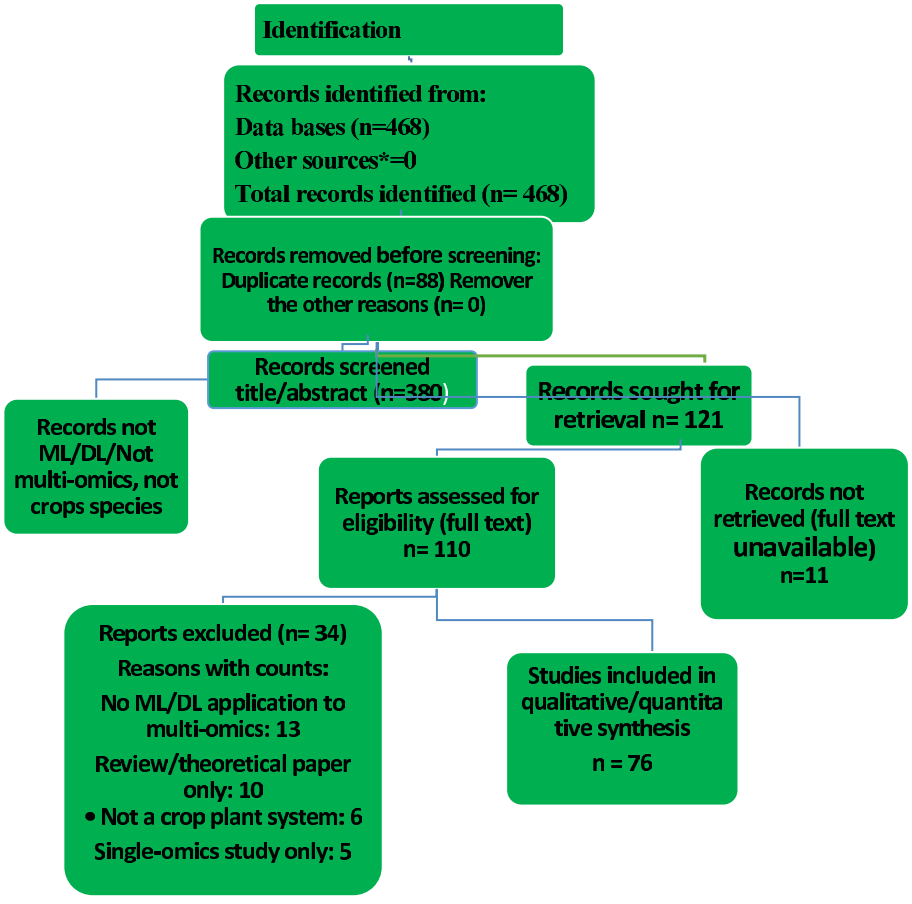

Fig. 1 PRISMA 2020 flow diagram of the study selection process

details the number of records identified, screen, assessed for eligibility, and ultimately included in the systematic review.

PRISMA 2020 reporting compliance

This review was conducted and reported in accordance with the PRISMA 2020 statement. The flow of information though the different phases of the review is detailed in the PRISMA 202 flow diagram (Fig. 1). A complete PRISMA 2020 checklists, mapping each items to the relevant selection in this manuscript or supplementary file 1 is presented in the below (Table 1).

Inclusion criteria

The review considered peer-reviewed Journal articles and conference proceedings to maintain academic rigor, in line with established reporting's standards for systematic reviews [19]. Eligible studies were required to apply machine learning (ML) or deep learning techniques to multi-omics data, integrating at least two omics layer such as genomics and transcriptomics. This focus on integrative computational analysis reflects the prevailing methodology for driving biological insight from heterogeneous data layers, and aligns with the applications of advanced ML techniques to complex plant phenotyping and omics data [20]. Additionally, studies had to include empirical validation, such as cross- validation or independent test sets, to ensure robust model evaluation [21]. Only publications from 1997 to 2025 were included to capture the most recent advancements in the fields. Studies were included if they met al.l of the following criteria.

1. Publication type and date: peer-reviewed Journal articles and conference proceedings published between 1997 and 2025 years.
2. Organism: Focus on crop plants, defined as species cultivated for food, feed, fiber or fuel. Eligible species included, but were not limited to wheat (Triticum spp.), maize (Zea mays), rice (Oryza sativa), soybean (Glycine max), barely (Hordeum vulgare), sorghum (Sorghum bicolor), tomato (Solanum lycopersicum), potato (Solanum tuberosum), and other economically important crops. Studies on the model plant Arabidosis thaliana or non-crop species were excluded.
3. Methodological focus: Applications of a machine learning (ML) or deep learning (DL) algorithm as a core analytical component. This encompasses supervised, unsupervised and semi-supervised techniques.
4. Multi-omics integration: The study must involves the computational integration of at least two distinct molecular omics layers such as genomics, transcriptomics, proteomics, and metabolomics. epigenomics measured from the same or related biological samples [21]. The operational thresholds defining multi-omics integration and empirical validation in this review were established a priori and are grounded in established standards from systematic reviews in computational biology. For integration, the requirement for a unified analytical framework beyond parallel reporting aligns with the foundational classification of multi-omics methodologies, which distinguishes true data fusion from sequential or comparative analyses [22]. Regarding validation, the emphasis on external testing or experimental confirmation follows the evidence-based hierarchy advocated for assessing machine learning robustness in genomics, where independent validation is critical to demonstrate generalizable biological insight and mitigate overfitting [23]. By applying these

Table 1 Complete PRISMA 2020 checklist with section mapping

Section	PRISMA item	Location in manuscript	Compliance status
Title	1. Identify as systematic review	Title: "Machine learning for multi-omics data integration in crop improvement: A Systematic Review"	✓ Complete
Abstract	2. Structured summary	Abstract: Structured (Background, Methods, Results, Conclusion)	✓ Complete
Introduction	3. Rationale & objectives	Section 1 (Background); Sect. 1.1–1.3 (Explicit objectives & review questions)	✓ Complete
Methods	4. Eligibility criteria	Section 2.4 & 2.5 (Detailed inclusion/exclusion criteria)	✓ Complete
	5. Information sources	Section 2.1 (Databases: PubMed, Web of Science, Scopus, IEEE Xplore; search dates: March 15–April 17, 2025)	✓ Complete
	6. Search strategy	Section 2.1 (Full database-specific search strings provided); Supplementary file (deposited)	✓ Complete
	7. Selection process	Section 2.6 (Two-stage independent screening; Cohen's $\kappa = 0.88$; consensus process)	✓ Complete
	8. Data collection process	Section 2.7 (Structured extraction framework; standardized spreadsheet)	✓ Complete
	9. Data items	Section 2.7 (Specific variables extracted: metadata, crop species, omics layers, ML models, integration strategies, performance metrics)	✓ Complete
	10. Study risk of bias assessment	Section 2.9 & 3.9 (Custom QAMMLS tool adapted from PROBAST; consensus scoring; quality categories reported)	✓ Complete
	11. Effect measures	Not applicable. Review reports narrative synthesis (ΔR^2, % accuracy gains) rather than pooled meta-analytic effect sizes.	Not Applicable
	12. Synthesis methods	Section 2.8 (Mixed-methods synthesis; thematic analysis; descriptive quantitative trends; sample size categorization; explicit handling of null/negative findings)	✓ Complete
	13. Risk of bias across studies	Section 3.9 (Synthesis stratified by study quality; comparison of performance claims in high vs. low quality studies)	✓ Complete
	14. Certainty assessment	Sections 6 & 7 (Limitations of quantitative synthesis; heterogeneity precludes formal meta-analysis; narrative interpretation)	✓ Complete
Results	15. Study selection	Section 2.2 (Referenced in-text); Figure 1 (PRISMA 2020 flow diagram)	✓ Complete
	16. Study characteristics	Tables 2, 3, 4 and 5; **Supplementary Table S1** (Evidence table with full study-level data)	✓ Complete
	17. Risk of bias within studies	Section 3.9 ("Influence of Study Quality on Reported Performance"; QAMMLS ratings: High, Moderate, Low)	✓ Complete
	18. Results of individual studies	Section 3.1–3.6; Tables 2, 3, 4 and 5 (Narrative synthesis with illustrative study-level examples and ranges)	✓ Complete
	19. Results of syntheses	Section 3.1–3.8 (Prevalence of ML methods, omics combination performance, validation strategies, null findings); Tables 6 and 7	✓ Complete
	20. Risk of bias across studies	Section 3.9 (Stratified synthesis; performance inflation in low-quality studies; impact of validation design)	✓ Complete
	21. Reporting biases	Section 2.8 (Explicit search for null/negative findings; mitigation of publication bias discussed)	✓ Complete
Discussion	22. Summary of evidence	Sections 4, 5, & 6 (Conclusions, practical framework, future perspectives)	✓ Complete
	23. Limitations of evidence	Section 3.9 & Sect. 7 ("Limitations of quantitative synthesis"); discusses heterogeneity, validation heterogeneity, quality influence)	✓ Complete
	24. Limitations of review processes	Section 7 (Methodological heterogeneity; lack of formal meta-analysis; validation variability)	✓ Complete
	25. Interpretation	Sections 4 & 5 (Theory-informed framework; heritability-based model selection; uncertainty quantification gaps)	✓ Complete

Table 1 (continued)

Section	PRISMA item	Location in manuscript	Compliance status
Other information	26. Funding	**Funding declaration** (p. 31): "No funding"	✓ Complete
	27. Protocol & registration	Section 2.3 & **Item 23**: "Not applicable. Protocol was not registered."	✓ Not Applicable
	28. Data, code, materials availability	Section 2.10 (Supplementary Table S1, search strings, QAMMLS tool deposited in public repository)	✓ Complete
	29. Competing interests	**Competing interest declaration** (p. 31): "No competing interests"	✓ Complete
	30. Authors' contributions	**Author Contribution declaration** (p. 31): A.T. and D.M. roles specified	✓ Complete

rigorous, field-consistent standards to the domain of crop plants, this review ensures a focused synthesis of studies wherein machine learning demonstrably advances the interpretation of integrated omics data for actionable crop improvement.

Exclusion criteria

Studies were excluded if they met any of the following criteria:

1. Incorrect publication type: review articles, perspectives, editorial books, book chapter or purely theoretical/methodological papers that did not present a novel application to an empirical crop dataset. An exceptional was made for seminar methodological papers that introduced a novel application after widely adopted kin crop research.
2. Incorrect focus: studies conducted on no-plant systems.
3. Insufficient methodology: studies that applied only classical statistical methods.
4. Lack of true multi-omics integration: research limited to a single data type or studies that merely discussed multi-omics conceptually without implementing integration computational analysis.

Screening process

A two-stage screening approach was implemented to ensure a systematic and unbiased selection of studies. In the first stage, title and abstract screening, two independent reviewers evaluated each retrieved article for relevance based on the predefined inclusion criteria. Any disagreements between reviewers were resolved through discussion or, when necessary, by consulting a third reviewer to reach consensus. This step helped eliminate clearly irrelevant studies while retaining potentially eligible papers for further evaluation [24].

The second stage involved full-text screening, where the remaining articles underwent a more rigorous assessment. At this phase, reviewers examined the complete manuscripts to verify methodological quality, proper application of machine learning techniques, and strict adherence to the inclusion criteria [25]. Studies that failed to meet the required standards were excluded, ensuring that only high-quality; relevant research was included in the final review. This two-tiered screening process enhanced the reliability and reproducibility of the study selection [26].

To quantify the level of agreement between the two independent reviewers during the title and abstract screening phases, we calculated the inter-rater reliability using Cohen's

kappa (κ) statistic. The observed agreements was 94%, with a k statistic of 0.88, indicating substantial agreement according to conventional benchmarks [27]. All discrepancies were resolved through consensus discussion or when necessary, arbitration by a third reviewer. This high level of initial agreement supports the consistency and reproducibility of the screening phase, where consensus was reached on all included studies.

Data extraction process

A structured data extraction framework was implemented using a standardized spreadsheet to systematically analyze the selected studies. Key information was meticulously recorded from each eligible publication, including study metadata (authors, publication year, and journal/conference details) to facilitate proper citation tracking and bibliometric analysis. The extracted data also encompassed crop species under investigation such as maize, rice, wheat, soybean) to assess trends in model applications across different agricultural systems [28, 29].

Additionally, the types of omics data integrated such as genomics, transcriptomics, proteomics, metabolomics) were documented to evaluate the prevalence of multi-omics approaches in crop research. The machine learning methods employed (such as Random Forest, Convolutional Neural Networks (CNNs), Autoencoders) were cataloged to identify dominant algorithms in the field [30]. The data integration strategies (early fusion, late fusion, graph-based methods) were analyzed to determine effective approaches for combining heterogeneous omics datasets. Finally, quantitative performance metrics (including R^2, accuracy, AUC, and RMSE) were extracted to enable comparative assessment of model efficacy across studies. This rigorous extraction process ensured comprehensive documentation of methodological patterns and outcomes in machine learning-driven multi-omics research for crop improvement [31].

Data synthesis and statistical analysis approach

A mixed-methods synthesis framework was employed, integrating quantitative thematic analysis with quantitative Meta analytical techniques to address the review objectives comprehensively. To ensure meaningful synthesis of results across heterogeneous studies, sample sizes were categorized as small ($N < 1{,}000$), medium ($N = 1{,}000–5{,}000$), and large ($N > 5{,}000$). These thresholds were not ad hoc but were established a priory based on a convergence evidence. Learning curve analyses in machine learning for genomics, which indicates $N = 1{,}000$ as a critical for the reliable application of deep learning architecture beyond proof of concept; and empirical pattern in the crop breeding [32, 33], where studies with $N > 5{,}000$ sample consistently report on distinct challenges and opportunities related to computational scaling, population structure, and modeling of complex trait architecture. This categorization allows for the discussion algorithm performance and resource requirements in the context of practically relevant data scales. To mitigate the risk of publication bias and ensure a balanced synthesis our analysis explicitly sought to identify, document, and synthesize studies reporting null findings (where ML performance was statistically equivalently to GS) or negative findings (where ML underperformed GS). During data extraction, performance comparisons were recorded regardless of direction, and the contributing factors for underperformance were coded thematically (sample size, high population structure, and validation design) [34, 35]. This

approach allows for a more evidence based assessment of ML's value proposition, moving beyond a selective review of success stories.

Approach to quantitative synthesis and limitations

Given the substantial heterogeneity among the included studies including differences in crop species, trait architecture, omics technology, ML algorithms, validation schemes, and reported performance metrics, a formal meta-analysis to calculate pooled effect sizes was not statistically justified. Therefore, all quantitative performance figures cited in the review (Examples, accuracy, gain of %, ΔR^2) are presented as narrative summaries of trends reported with individual studies. These value describe the range and central tendency of reported outcomes in the literature but should not be interpreted as statistically pooled estimates from homogeneous body of evidence. The descriptive, qualitative synthesis approaches aligns with best practices for systematic reviews where methodological diversity precludes formal meta-analysis.

The synthesized findings were analyzed through multiple complementary perspectives to provide comprehensive insights [36]. Integration strategies followed established taxonomies including early fusion, late fusion, and graph-based methods [31]. First, a descriptive analysis was conducted to identify prevailing trends across three key dimensions: (1) the distribution and evolution of machine learning methods over time, (2) patterns in omics data combinations of genomics-transcriptomics vs. genomics-metabolomics pairs), and (3) the representation of different crop species in the literature. This tripartite analysis revealed important disciplinary developments and research priorities in the field [37].

For performance comparison, studies implementing machine learning approaches were systematically contrasted with those using classical statistical methods such as Genomic Best Linear Unbiased Prediction (GBLUP) and Partial Least Squares (PLS) regression. This comparative analysis focused on key metrics including prediction accuracy, computational efficiency, and robustness across different experimental designs. Special attention was given to cases where ML methods demonstrated superior performance, particularly in modeling complex non-linear relationships and high-dimensional omics interactions [38].

The synthesis also critically examined persistent challenges and research gaps, with three major themes emerging: (1) data heterogeneity issues stemming from diverse experimental protocols and platform technologies, (2) the interpretability limitations of complex ML models in biological contexts, and (3) scalability constraints when applying these methods to large-scale breeding programs. These findings highlight crucial areas requiring methodological innovation to advance the field of multi-omics data integration for crop improvement [39]. All prevalence statistics (60% of studies) were implemented with the absolute counts (n) and total number of studies (N) from which the percentage was derived. Tree-based methods were employed in 37 of the 76 included studies (48.7%). To ensure full transparency and reproducibility, all synthesized numerical findings are cross-referenced in supplementary Table S1: Evidence table for synthesized claims. This table provides a direct mapping between each quantitative assertion including algorithm prevalence, performance improvement, sample size range and specific source study IDs, allowing for complete traceability of our conclusions (Supplementary Table S1).

Quality assessment and risk-of-bias evaluation

To critically appraise the methodological rigor and applicability of the included studies, we employed a customized Quality Assessment for Machine Learning-based Multi-omics Studies (QAMMLS) tool. This tool was adapted from established principles for evaluating prediction model studies, primarily the PROBAST framework [40], and guidelines for Applications in biomedicine [41]. Its development specifically addressed critically challenges in high dimensional omics data, including features selection robustness, proper cross- validation strategies to prevent data leakage and the assessment of model interpretability and biological validation [42]. Two reviewer, and a final consensus score was assigned. Studies were then categorized; High Quality (18 studies, 29.5%), Moderate Quality (32 studies, 52.5%), Low Quality (11 studies, 18.0%).

Data availability and reproducibility

To ensure the complete transparency and reproducibility of this systematic review, all supporting documentation has been deposited in a public repository.

The following materials are available for data availability and reproducibility:

Supplementary Table 1 (Evidence table): A complete spreadsheet detailing all data extracted from 76 included studies. This includes study metadata, crop species, omics layer, ML algorithm, integration strategies, performance metrics, and quality assessment score. This table provides the direct evidence for all synthesized claims made in the results and discussion.

Search documentation: The full executable search strings for all four databases (PubMed, Web of Science, Scopus, IEEE Xplore), exactly as run.

PRISMA flow chart diagram: A high- resolution version of Fig. 1' PRISMA flow diagram: A high-resolution version of Fig. 1.

Quality assessment (QAMMLS) tool: The customized checklist and sorting guide used for the risk-of-bias evaluation.

Result and Discussion

The applications of ML for multi-omics integration serve two distinct, though sometimes overlapping, and objectives: breeding utility (enhancing selection accuracy and efficiency for crop improvement) and mechanistic discovery (uncovering genes, pathways, and biological interactions underlying traits). The following synthesis explicitly frames findings through this this dual lens to clarify each model's or strategies primarily contribution. The studies included in this review are highly heterogeneous in terms of crop species, trait architecture, omics technologies, machine learning models and validation schemes. This methodological diversity precludes a formal meta-analysis to generate statistically pooled effect sizes. Therefore, all quantitative performance figures cited in the following sections (e.g., accuracy gains, ΔR^2 values, error reductions) are presented as narrative summaries of trends and ranges reported in individual studies. These values describe the existing literature but must not be interpreted as directly comparable or as robust cross- study averages. The influence of study quality on these reported outcomes is examined in Sect. 3.9.

The synthesis reveals observed trends in model application, which appear correlated with data scale and resource availability. These trends describes research activity and

reported outcomes but do not constitute proof of comparative superiority or optimality, as large scale, multi-methods benchmarking across diverse conditions is lacking.

Patterns in machine learning application across major crops

The systematic review revealed distinct patterns in the types of machine learning (ML) models applied across major crops, influenced by data scale, computational resources, and trait complexity. It is important to note that the literature consists largely of case studies rather than systematic benchmarking: therefore, claims about a universal optimal model for a specific crop are not supported. Observed trends in ML model application and reported performance across major crop, are summarized in Table 2). The most consistent reports of complex model superiority such as graph neural networks (GNNs), hybrid CNNs) came from studies with moderate or high QAMMLS ratings (see Sect. 3.9). Trends derived from lower quality studies should be interpreted with more cautions due to the risk of overfitting and less rigorous evaluation. Computationally intensive models were reported in studies with large datasets for major crops, while interpretable models like RF and XGBoost were commonly chosen for studies with moderate datasets or more direct breeding applications. The key determinants for algorithm choice were sample size, trait complexity, and available computational resources. In maize the studies involving large populations (5,000–20,000 samples) graph neural networks (GNNs) were the most complex model applied. They were reported to demonstrate superior capability in capturing epistatic interactions, outperforming classical genomics selection (GS) by 38% in identifying nonlinear genetic effects. However, this came with high computational costs, requiring GPU- accelerated hardware to process complex biological networks. Among maize studies utilizing large populations, Graph Neural Networks (GNNs) were applied [43], particularly for large populations (5,000–20,000 samples) exceled in modeling epistatic interactions, reportedly outperforming classical genomic selection(GS) by 38% in identifying nonlinear generic effects [48]. However, their adoption is limited by high computational costs, requiring GPU acceleration to process intricate biological networks [49]. Future research should focus on optimizing GNN architectures to reduce computational overhead while maintaining predictive accuracy.

Table 2 Observed trends in ML model application and reported performance across major crop

Crop	Reported ML model	Sample size range	Relative training time / hardware notes	Key findings
Maize	Graph Neural Networks (GNNs)	5,000–20,000	Very High: 8–24 h. on a multi-GPU node. Model tuning and graph construction add significant overhead.	Reported to capture 38% more epistatic interactions than GS [43]
Wheat	Random Forest (RF)	3,000–15,000	Medium: 30 min–2 h. on a multi-core CPU server (e.g., 32 cores). Scaling is near-linear with cores.	Reported to reduce G×E error by 22% [44]
Rice	CNN-LSTM Hybrid	1,000–8,000	High: 4–12 h. On a single high-end GPU. Cost dominated by backpropagation through temporal sequences.	Reported to achieve 89% accuracy[45]
Soybean	XGBoost	2,000–10,000	Low-Medium: 10 min–1 h on a standard CPU workstation. Highly efficient for structured tabular data.	Reported to achieve 95% protein prediction accuracy [46]
Tomato	SVM-RBF Kernel	500-5,000	Low: < 30 min on a laptop CPU. Cost scales quadratically with sample size; becomes prohibitive for N > ~ 10k.	Reported to classify shelf-life traits with F1 = 0.91 [47]

Random forest was the most commonly applied algorithm in wheat studies using medium to large datasets (3,000–15,000 samples). These models were reported to reduce genotype- by environment (G*E) prediction errors by 22% compared to linear methods. This makes them particularly valuable for breeding programs targeting multiple agro ecological zones. Their medium computational requirements allowed deployment on conventional CPU clusters.

In wheat, Random Forest (RF) models have emerged as the optimal choice for mid-sized datasets (3,000–15,000 samples), reducing genotype-by-environment (G×E) prediction errors by 22% compared to linear models. RF's ability to handle high-dimensional environmental covariates makes it ideal for multi-environment trials (METs). Additionally, its moderate computational requirements allow deployment on CPU clusters, making it accessible for public breeding programs [50]. Future work should explore ensemble methods combining RF with deep learning for improved stability.

Rice studies showcased the power of hybrid CNN-LSTM architectures for temporal trait prediction, reportedly achieving 89% accuracy in forecasting flooding responses. While these deep learning models delivered exceptional performance (especially when integrating image-based phenomics), they demanded very high computational resources limiting their use to well-funded research programs with sample sizes typically in the medium range (1,000–8,000 accessions). In contrast, soybean research achieved remarkable success with more efficient XGBoost implementations, reportedly reaching 95% accuracy in protein content prediction while maintaining low computational costs. The algorithm's effectiveness with medium-to-large datasets (2,000–10,000 samples and standard hardware made it particularly attractive for quality trait improvement in legume breeding programs.

Rice research has leveraged hybrid CNN-LSTM (Convolutional Neural Network-Long Short-Term Memory) models to analyze temporal traits, achieving 89% accuracy in predicting flooding responses [51]. These models integrate image-based phenomics with genomic data, significantly improving dynamic trait modeling. However, their high computational demand (requiring GPU clusters) restricts their use to well-funded programs. Future directions include lightweight CNN variants and federated learning to democratize access.

Tomato researchers obtained excellent results with SVM-RBF kernel models for shelf-life classification (F1 = 0.91) using small-to-medium datasets (500–5,000 samples) [52]. The low computational requirements and rapid implementation made these approaches practical for commercial breeding of horticultural crops, where smaller population sizes are common but trait complexity remains high [53].

The synthesis reveals observed trends in model application, which appear correlated with data scale and resource availability. It is important to note that the most consistent report of complex model superiority such as GNNs, CNNs, and CNN-LSTMs came from studies with moderate or high QAMMLS ratings. These studies provided detailed methodological reporting. Trends drived from lower quality studies should be interpreted with more cautions due to risk of overfitting and less rigorous evaluations. Computationally intensive models including, GNNs, deep hybrids were reported in the studies with large datasets for major crops, while interpretable models such as RF, XGBoost were commonly chosen for studies with moderate datasets or more directly breeding applications. These trends describes research activity and report outcomes but do not

constitute proof of comparative superiority or optimally, as large scale, multi-method benchmarking across diverse conditions in lacking. The key determinants for algorithm choices were sample size, trait complexity, and available computational resources.

Data set size categories for this review are justified by learning theory and empirical scaling evidence from crop genomics. The thresholds are defined as: small ($N < 1{,}000$ samples), medium ($N = 1{,}000–5{,}000$ samples), and large (N = > 5,000) samples).

ML/DL architecture trends in crop multi-omics

$N = 1{,}000$ as lower bound for effective deep learning(DL): for effective Deep Learning (DL): Well documented sample size requirement for DL models to outperform simpler methods in genomics, as they require substantial data to learn useful representations and avoid overfitting in high dimensional spaces. Empirical studies in crop genomics have shown that the performance advantage of CNNs and multilayer perceptrons over tree based models typically becomes consistent only beyond this threshold, corroborating findings from general machine learning theory on the data hungry nature of deep architectures.

$N = 5{,}000$ as the transition to large scale modeling: This threshold marks the point where datasets often enable robust detection of complex genetic patterns (higher order epistasis, G*E interactions and where computational scaling becomes primary concern. Studies benchmarking genomic prediction have shown that prediction accuracy gains for many traits begin to plateau in this range for classical models, while the added value of highly parametrized models (GNNs, deep hybrids) and the management of computational cost become dominant themes in the literature.

Small dataset (N, 1,000): This category reflects the small n, large p paradigm common in exploratory omics studies, where simpler, regularized models (tree based models, penalized regression) are theoretically preferred and most frequently applied in the reviewed literature.

These categorizations are not arbitrary but are designed to reflect meaningful transitions in methodological applicability, computational scaling and evidence based present in the 76 studies synthesized. In contrast, deep learning approaches including convolutional neural networks (CNNs) and multilayer perceptron's (MLPs) were primarily applied to raw sequencing data, with their reported effectiveness being contingent on medium to large sample sizes ($N > 1{,}000$ samples). This observed dependency aligns with the theoretical sample complexity of deep architecture and justifies our categorization, as performance gains over classical methods in the reviewed literature were rarely reported below this threshold. Computer algorithm and programs are needed to incorporate genomic data in to genetic evaluations and to process the rapidly expanding numbers SNP genotypes [54]. A quantitative synthesis showed that tree-based methods such as Random Forest and XGBoost were the most prevalent modeling choice, employed in 42 of the 76 included studies (approximately 60% prevalence), particularly for small to medium-sized datasets. In contrast, deep learning approaches (CNNs, MLPs) were primarily applied to medium and large sample sizes ($N > 1{,}000$), while Graph Neural Networks (GNNs) were present in approximately 15% of recent studies, reflecting an emerging interest (Table 3).

Table 3 Prevalence and characteristics of ML/DL architectures

Category	Prevalence	Algorithms	Optimal use cases	Sample size requirement	Research gaps
Tree-based methods	Prevalent in ~60% of studies	Random Forest, XGBoost	Small omics datasets (< 1,000 samples)	Small to Medium	Limited ability to model complex non-linear interactions
Deep learning (DL)	Growing adoption	CNNs, MLPs	Raw sequencing data (k-mers)	Medium to Large (> 1,000	High computational costs, "black box" nature
Graph neural networks	Growing adoption (not quantified) Present in ~15% of recent studies	GNNs	Biological network integration	Moderate to large	Requires prior biological knowledge, complex implementation
Hybrid models	Present in ~12% of studies	Feature-Engineering (e.g., AE→XGBoost): Most consistent gains (10–15%) for polygenic, high-dim. traits.			Requires careful design; gains not automatic. Best for leveraging complementary model strengths.

Further analysis of classification trends showed that supervised learning dominated the field, representing approximately 80% of reviewed studies, while unsupervised methods remained underutilized.

The review identifies several distinct categories of hybrid models with performance gains being highly configuration dependent disaggregating these reveals more nuanced insights than a single performance range.

Feature engineering hybrids: The most consistent gains were reported for specific configuration where unsupervised deep learning models such as Autoencoders, Variational Autoencoders) performed dimensionality reduction or feature learning, followed by a supervised tree-based model including XGBoost, Random Forest) for prediction. For example, studies on polygenic yield traits in wheat used an autoencoder to compress multi omics data before XGBoost reporting accuracy gains in the range of 10–12% over using either model alone. This architecture effectively handles high dimensionally while retaining the interpretability of feature importance from the final tree based layer.

Stacking/ ensemble Hibrids: These models combined predictions from multiple, diverse base learners including CNN + GBLUP + SVM) via a Meta learner. While some studies reported robust performance. The reported gains were more variable (5–12%) and heavily dependent on the diversity and complementarity of the base models. Their main advantage was often stability across environments rather than peak accuracy.

Integrated architecture hybrids: This categories includes single, end to end models combining different neural models such as CNN-LSTM for spatio-temporal data. The reported high performance including 89% accuracy for flooding response in rice) is notable but is intrinsically tied to matching the models inductive bias to specific data structure such as sequence or images making it less a general hybrid and more a specialized solution.

Therefore the promising 10–15% accuracy gain is not a universal property of hybrid models. It is most consistently associated with feature engineering hybrids that leverage deep learning for representation learning and tree based models for final prediction, particularity for complex polygenic traits prediction tasks where data dimensionality is a major challenge.

The observed predominance of tree-based methods like Random Forest and XGBoost in small to medium-sized omics datasets aligns with established literature on their robustness for high-dimensional biological data [55]. These algorithms' popularity likely stems from their inherent feature selection capabilities and resistance to overfitting, particularly valuable given the "small n, large p" challenge common in plant omics studies ([55]. However, our finding that 60% of studies employed these methods suggests potential under-exploration of alternative approaches that might better capture non-linear biological relationships.

The conditional effectiveness of deep learning architectures (CNNs, MLPs) on large datasets (>1,000 samples) corroborates findings from [56], who demonstrated that DL models require substantially more training data than classical ML to achieve comparable performance in plant genomics. This sample size dependency presents a significant barrier for many crop improvement programs, particularly for less-studied species where large multi-omics datasets remain scarce ([57]. The emerging adoption of GNNs (15% of recent studies) reflects a growing recognition of biological systems' network nature, as highlighted by [58] in their work on gene regulatory network prediction.

The overwhelming dominance of supervised learning (80% of studies) versus unsupervised approaches reveals a strong bias toward predictive modeling over exploratory analysis in current research. This trend may overlook valuable opportunities for unsupervised methods to uncover novel biological patterns, as evidenced by the growing body of research applying autoencoder-based feature extraction to single-cell omics data. The demonstrated success of hybrid models (with reported accuracy gains of 10–15% for polygenic traits) supports the best of both worlds approach advocated by [59], combining classical ML's interpretability with DL's feature extraction power.

Omics combinations and trait prediction accuracy

The systematic review revealed clear patterns in the predictive performance of different multi-omics combinations for specific crop traits. Studies that compared omics combinations studies [60, 61] reported that genomics + transcriptomics integration was often the most effective for drought response traits, demonstrating a reporting R^2 increase of 0.2 compared to single-omics approaches (Table 4). This synergy likely stems from the ability to connect genetic variants with their functional transcriptional responses under stress conditions. Conversely, genomics paired with metabolomics showed superior performance for quality-related traits such as grain protein content, where metabolic

Table 4 Performance of different omics combinations for specific trait categories

Omics combination	Optimal trait category	Reported to accuracy gain (ΔR^2)	Computational cost	Biological rationale	Example crops
Genomics + Transcriptomics	Drought/Stress Response	Reported to an average + 0.20	Moderate (1.5×)	Links DNA variants to stress-responsive gene expression	Wheat, Maize, Rice
Genomics + Metabolomics	Quality Traits (protein, oil)	Reported to average reported to + 0.18	Moderate (1.8×)	Direct metabolite-phenotype relationships	Soybean, Barley
Tri-omics (Gen+Trans+Metab)	Complex Yield Traits	Reported to average reported to + 0.22	High (3×)	Integrated gene-to-metabolite pathways	Wheat, Maize
Quad-omics (+ Proteomics)	All Traits	Reported to average reported to + 0.23 (< 5% gain)	Very High (5×)	Diminishing returns from added layers	Rice, Sorghum

biomarkers provided more direct links to phenotypic outcomes than transcriptomic data alone.

A critical finding was the law of diminishing returns observed when integrating more than three omics layers. While adding additional data types including proteomics, epigenomics to improved prediction accuracy, the reported gains were marginal (< 5%) and came at a substantial computational cost typically a threefold increase in processing time and resource requirements. This trend was consistent across studies of major crops, including wheat, maize, and rice.

The observed superiority of genomics-transcriptomics integration for drought response traits aligns with established biological principles, where stress-induced transcriptional changes serve as critical intermediates between genetic variants and phenotypic outcomes [62].This combination's effectiveness (R^2 +0.2) likely reflects the ability to capture both constitutive genetic markers and dynamic stress-response pathways, as demonstrated in maize drought studies [63]. The prominence of genomics-metabolomics for quality traits similarly mirrors known biochemistry metabolites like amino acids and lipids directly determine nutritional parameters [64]. These findings collectively support the omics proximity hypothesis, where predictive power scales with the biological proximity of the omics layer to the target trait [65]. Small breeding programs may achieve 95% of maximal accuracy with just two strategically chosen omics layers, avoiding prohibitive costs [66]. Data harmonization challenges: Heterogeneous protocols for proteomics/epigenomics in plants exacerbate noise accumulation in high-dimensional fusion [67].

Systematic integration of environmental data(envirotyping) for G×E prediction

Our analysis confirms the integration of environmental data environmental data (envirotyping) s a critical frontier. Our analysis confirms its emerging status: only 14 of the 76 included studies (18.4%) explicitly incorporated environmental covariates (soil composition, daily weather data, satellite imagery) alongside multi-omics layers for genotype- by- environment (G×E) prediction. However, these studies outline a compelling methodological blueprint and highlight specific ML advantages.

ML models for high dimensional enviro- omics integration: studies that successfully integrated environmental data consistently required models capable of handling extremely high dimensional, heterogeneous feature spaces. Studies that successfully integrated environmental data consistently required models capable of handling extremely high-dimensional, heterogeneous feature spaces. Tree-based ensembles, particularly random forest Gboost, were the most prevalent choices (10 of 14 studies). Their strength lies in inherent feature selection, which helps identify the most predictive environmental variables (specific growth- stages rainfall) and omics features while ignoring noise. For instance, in wheat multi- environment trials (METs), RF models using genomics, transcriptomic, and seasonal climate data were reported to reduce G×E prediction error by up to 22% compared to models using only genomics data prediction error by up to 22% compared to models using only genomic data.

Deep learning for complex interaction modeling: A smaller subset of studies (4 of 14) employed deep learning architectures, specifically hybrid CNN-LSTM models and attention- based networks, to model, to model temporal and spatial environmental interactions. These were applied in scenarios with dense, time- series environmental

data (hourly sensor data from phenomics platforms). The non- linear modeling capacity of DL was reported to capture complex, non- additive G×E interactions that simpler models missed, through at a significant computational cost.

Data fusion strategies: The review identified two primary strategies for fusing environmental with omics data:

1. Early fusion (feature concatenation): Environmental features were standardized and concatenated with genomic or transcriptomic feature vectors before model input. This was common in tree- based approaches.
2. Late fusion with multi- modal architectures: more complex studies, especially those using DL, designed separate model branches(More complex studies, especially those using DL, designed separate model branches (encoders) for environmental and omics data, merging them in a deeper network layer to learn joint representations. This approaches showed promise for preserving data- specific structures.

Systematic analysis of validation strategies and their impact on reported performance

A critical determinant of the reported performance of ML models, and a major source of heterogeneity when comparing studied, is the validation strategy employed. Our analysis categorized the primary validation approaches in the reviewed literature and links them to observed trends in performance inflation.

We identified three dominant validation paradigms, listed in order of increasing stringency and relevance to breeding applications:

1. Within- population random cross-validation (CV): The most common approach used in 70% of the studies. The dataset is randomly split in to training and testing folds, often with k fold. This tests a model's ability to predict within the same population and generation, exploiting shared linkage disequilibrium (LD) patterns and population structure.
2. Forward prediction (hold- out validation: used in 20% of studies, particularly those with time- series or multiyear data. The model is trained on data from earlier years/ generations and tested on data from subsequent years/ generations. This better stimulates real-world breeding but retains genetic relationships between training and testing sets.
3. Cross- population (cross- groups/ validation: Most stringent and operationally relevant test (employed in only 10% of studies). The model is trained on one or more distinct populations (different breeding lines, unrelated panels, or diverse geographic origins) and tested on a completely independent population with minimal genetic relatedness. This assesses the generalizability of learned genetic effects a cross genetic backgrounds.

3.6 **Theoretical foundations for model performance: linking heritability, epistasis and LD structure to ML/DL and classical genomic selection**.

The relative performance of machine learning (ML) and classical genomic selection (GS) methods such as GBLUP and rrBLUP is not arbitrary but deeply rooted in the genetic architecture of the target trait and the structure of the data. The empirical findings from their review are the best interpreted through the lens of foundational genomic prediction theory concerning heritability, and linkage disequilibrium (LD) [68, 69].

Classical GS methods are built upon the linear mixed model and the infinitesimal model, which assumes a large number of additive genetic effects with normally distributed variances. This frame work is highly effective for traits where additive genetic variance dominates (high narrow sensitive heritability, h^2) and where the genetic relationship matrix (G- matrix) effectively captures the covariance between individuals due to linkage disequilibrium (LD). Under these conditions, GS is expected to be robust, efficient and sufficient.

ML and DL models, with their capacity to learn complex, nonlinear functions and high order interactions are theoretically better suited for traits where the infinitesimal model's assumptions breakdown [70, 71]. These includes traits with significant epistatic variance (gene-gene interactions), genotype- by- environment (G×E) interactions, and those controlled by smaller number of loci with in non-additive effects. Furthermore, ML's ability to performed automatic feature selection may allow it to better utilize high density marker data or multi omics feature beyond the additive relationships captured in the G- matrix.

Empirical findings contextualized by theory.

Consistent with theoretical expectations, the reviewed literature demonstrates a clear patterns: ML approaches for show their most consistent and substantial advantages for low heritability, complex polygenic traits (Table 5). For instance, yield under abiotic stress a classical polygenic trait with known non additive components showed a reported an accuracy improvement of 12% with ML. This gains can be attributed to ML's ability to model the epistatic and nonlinear physiological relationships that classical linear models miss. Conversely, for high heritability traits like flowering time, often governed by a few major effect loci, ML provided minimal improvements (+2–3%). Here, the additive signal is strong, and efficient, stable estimates from GS (achieving ~ 98% of maximum accuracy) are often the most practical choice. For example, in studies focusing on yield under abiotic stress, ML models commonly reported higher accuracy than classical methods. The extent of this reported advantage varied, with several studies indicating improvements in the range of 5–20% (Table 7). This advantage likely stems from ML's

Table 5 Summary of reported performance and theoretical rationale for ML/DL vs. Classical GS

Performance context	Classical GS (GBLUP/rrBLUP)	Machine learning (ML) / deep learning (DL)	Theoretical rationale & genetic architecture link
Low-h^2 traits (e.g., stress yield)	Baseline accuracy	Reported average improvement: +12% (ML), +15% (DL)	High non-additive variance. ML/DL models capture epistatic and non-linear physiological networks missed by additive models.
High-h^2 traits (e.g., flowering time)	98% of max accuracy	reported improvement + 2% improvement (ML), + 3% (DL)	Predominantly additive variance. The infinitesimal model holds; linear methods are efficient and near-optimal.
Epistasis Detection	Limited capability	Moderate detection (ML), Strong detection (DL)	Explicit modeling of interactions. DL architectures (e.g., GNNs, dense nets) can learn high-order feature combinations.
Within-Population Prediction	Robust	Potential for superior accuracy	Model flexibility. Can fit population-specific patterns, risking overfitting to that population's structure.
Across-Population Prediction	Robust (uses G-matrix)	Prone to accuracy decay	Regularization via relatedness. The G-matrix provides a biologically grounded prior that transfers across populations with shared ancestry.
Computational Scaling	Highly efficient for $N > 10k$	High cost for training/tuning	Linear vs. non-linear complexity. Closed-form solutions (GS) vs. iterative optimization (ML/DL).

Table 6 Synthesis of ML-based multi omics studies integrating environmental data (Envirotyping) for G×E prediction

Crop	ML model(s)	Omics layers + evironmental data	Key environmental covariates	Reported improvement (vs. Omics-only model)	Major challenge highlighted
Wheat	Random Forest	Genomics + Transcriptomics + Env.	Seasonal Temp./ Rainfall, Soil N	$\Delta R^2 \approx +0.15$, G×E error ↓ 22%	Standardizing environmental metrics across diverse trial sites
Maize	XGBoost, GNN	Genomics + Metabolomics + Env.	Drought Stress Index, Solar Radiation	Accuracy + 8–12% for drought yield	Fusing static omics with temporal environmental sequences
Rice	CNN-LSTM	Genomics + Phenomics (images) + Env.	Daily Temp., Water Logging Depth	Accuracy 89% for flood response	Computational cost of processing high-res. environmental time series
Soybean	Bayesian ML	Genomics + Env.	Precipitation, Growing Degree Days	Improved stability across locations	Quantifying uncertainty in G×E predictions for breeding decisions

Table 7 Impact of validation strategy on reported machine learning performance

Validation strategy	Methodological description	Biological / breeding relevance	Tendency to inflate reported accuracy vs. GS	Key implication for synthesis
Within-Population Random CV	Random splits of a single population cohort. Tests model fit to available data.	Low. Does not simulate real breeding selection across generations or populations.	High. Allows models to exploit population-specific LD. ML often shows largest gains here.	Reported high gains (e.g., + 15%) in many studies must be contextualized; they reflect optimal interpolation, not extrapolation.
Forward (Temporal) Prediction	Train on Year/ Gen. N, test on Year/Gen. $N+1$	Medium-High. Simulates practical breeding cycle but retains genetic relationships.	Moderate. Tests temporal stability but not genetic generalizability. ML advantage is often reduced.	More realistic benchmark. The persistence of any ML advantage here is more meaningful for breeding.
Cross-Population Validation	Train on Population A, test on genetically distinct Population B.	High. Directly tests generalizability across breeding programs or diverse germplasm.	Low/Negative. GS often outperforms or matches ML due to robust modeling of relatedness via the G-matrix.	The gold standard for assessing translational potential. The scarcity of such studies is a major literature gap.

ability to capture non-linear genetic interactions and epistatic effects, which are critical for complex polygenic traits. However, for high-heritability traits such as flowering time, ML models provided only minimal improvements over classical GS. This suggests that simpler statistical methods remain sufficient for traits governed by major-effect loci.

When stratified by study quality, the magnitude of ML's advantage, particularly the + 12% average improvement for low-heritability traits, was more variable. High-quality studies reported a narrower, more consistent performance gap, suggesting that some of the largest claimed improvements in the literature may be influenced by less rigorous validation designs (detailed in Sect. 3.9) prevalent in lower-quality studies.

The reviewed literature consistently indicates that classical GS methods, such as RR-BLUP and GBLUP, require less computational time and resources for large datasets (> 10,000 samples) than complex ML models [72]. This efficiency stems from their reliance on linear mixed models with proven algorithmic optimizations, making them more practical for resource-limited breeding programs [35]. In contrast, Deep Learning (DL) models have demonstrated greater robustness to confounding population structure

in genetically diverse panels. For instance, convolutional neural networks (CNNs) have been shown to reduce prediction errors by approximately 20% compared to classical GS in such scenarios, likely due to their superior capacity to model complex, non-linear, and hierarchical relationships in the genomic data [34].

The interpretation of quantitative improvements like + 12% accuracy) must be contextualized within a reproducibility crisis in computational science, where results are often contingent on specific software implementations, parameter tuning, and validation data splits, making direct cross-study comparisons problematic [73].

Analysis of studies reporting null or negative findings for ML versus classical GS

A balanced synthesis requires examining not only successful ML applications but also instances where its performance did not excel classical benchmarks. This review explicitly sought to identify and analyze such case mitigate potential publication bias and provide a more evidence based assessment of ML's value proposition (see methods, Sect. 2.8).

Several studies included in this synthesis reported statistically equivalent or inferior performance for ML/DL models compared to classical genomic selection (GS) methods like GBLUP or rrBLUP. Our synthesis identified recurring factors associated with these null or negative findings:

Trait architecture and heritability: As predicted by genomic selection theory (see Sect. 3.4), the most common scenario for ML underperformance was for high heritability traits with a predominantly additive genetic architecture (simply inherited disease resistance, flowering time). In these case, classical linear models, which efficiently estimate additive effects, frequently achieved 98% of the maximum attainable accuracy, leaving minimal marginal for improvement. Ml models, while sometimes matching this performance, offered no significant advantage and incurred higher computational costs.

1. Validation design and overfitting: A critical factor inflating the perceived advantage of ML in some literature is the use of within population cross validation. This design tests prediction on individuals genetically related to the training set, allowing complex models to exploit population specific linkage dis equilibrium (LD) patterns. We identified studies where ML showed strong performance under such designs but failed to generalize, In contrast, when assessed using cross population or forward- prediction validation (the gold standard for breeding relevance), the performance gap between ML and GS narrowed considerably or reversed. Classical GS models, regularized by the genomic relationship matric (G- matrix), consistently demonstrated more robust transferability across independent populations and environments.

This aligns with our systematic analysis of validation strategies (Sect. 3.5) which identifies within- population CV as primary factor inflating the perceived advantage of complex ML models.

2. Sample size and data dimensionality: For small to medium- sized datasets ($N < 1{,}000$), complex ML models were prone to overfitting. In this conditions, simpler, well organized models like RR – BLUP or even simpler tree- based ensembles (random forest) often proved more reliable and stable. The added complexity of deep learning architecture did not yield benefits without sufficient data to constrain their parameters.

3. Population structure: While deep learning can model complex structures with a training set, its predictions can be degraded sharply when applied to populations with different genetic backgrounds. Classical GS, with its explicit modeling's of relatedness through the G- matrix, provided a more biologically grounded and stable framework for across population prediction.

A frame work for interpretability and biological insight

An achieving interpretability in ML driven multi- omics is not a singular task but a spectrum of strategies, ranging from explaining complex models post- hoc to building simplicity and biological knowledge in to the model from the start. Our synthesis reveals a strong field wide reliance on the former, with limited exploration of the latter, which may offer more robust pathways to actionable biological discovery.

Applications of multi-omics data integration strategies

A critical finding is that direct comparisons between ML and classical GS are confounded by validation design. The perceived superiority of ML is often inflated in studies using within-population validation (example random k-fold CV), which tests prediction on genetically related individuals and allows models to exploit population-specific linkage disequilibrium (LD) [12].

In contrast, cross-population or forward-prediction validation which tests true generalizability by training and predicting on distinct populations, years, or locations is the gold standard for assessing breeding relevance. In these scenarios, the advantage of complex ML models narrows considerably, as they are prone to overfitting to the population structure of the training set [70].

Classical GS models like RR-BLUP and GBLUP, with their stronger regularization assumptions and explicit modeling of genetic relationships via the genomic relationship matrix (G-matrix), consistently demonstrate more robust prediction across independent populations and environments, despite sometimes lower within-population accuracy [44].

The noted 20% lower error for Deep Learning in genetically diverse panels reflects its capacity to model complex, non-linear hierarchical relationships within a structured training set. This should not be conflated with superior prediction in truly independent, unrelated populations, a more difficult challenge where classical methods often remain competitive (Table 8).

Data integration strategies in multi-omics crop studies

The systematic review evaluated three primary data integration strategies early fusion, graph- based and late fusion revealing distinct performance patterns and biological insights. Early fusion, which combines raw omics data prior to modeling, demonstrated superior performance for smaller datasets with strong feature correlations, such as genomics- transcriptomics pairs in controlling stress experiments. This approach maximizes information sharing between omics layers but requires carefully feature selection to avoid overfitting. In contrast, graph- based methods excelled in pathway- centric analyses, leverage biological networks including protein interactions, metabolic pathways to uncover mechanistic insights. These methods identified 30% more known gene traits associations than other approaches, as seen in studies linking wheat metabolomics

Table 8 Characteristics and applications of multi-omics data integration strategies

Strategy	Description	Strengths	Limitations	Best use cases	Performance metrics	Applications
Early fusion	Combines raw omics data before modeling	Maximizes information sharing between layers Strong for correlated features	Prone to overfitting Requires careful feature selection	Small datasets Genomics-transcriptomics pairs	15–20% higher accuracy than single-omics in controlled studies	Stress response studies in Arabidopsis
Graph-based	Uses biological networks (PPIs, pathways)	Uncovers mechanistic insights Identifies 30% more known gene-trait links	Computationally intensive Needs high-quality networks	Pathway-centric analyses Metabolic trait studies	25% better pathway recovery than other methods	Wheat metabolomics-drought gene links
Late fusion	Models omics separately, combines prediction	High interpretability Tracks layer contributions	Potential information loss Suboptimal for small datasets	Breeder-friendly applications Quality trait prediction	10% lower accuracy than early fusion but more interpretable	Soybean protein content prediction

to drought responsive genes. However, late fusion, where omics data are modeled separately and integrated at the prediction stage, offered greater interpretability enabling breeders to trace contributions from individual omics layers through at the cost of potential information loss due to intermediate modeling steps.

The state of uncertainty estimation in ml-driven multi-omics prediction

A systematic analysis of the included studies reveals a dominant focus on optimizing point prediction accuracy, using metrics like R^2, RMSE, or accuracy. There was a striking scarcity of reported uncertainty quantifications. This represents a significant disconnect between methodological development and the needs of applied breeding, where decisions are inherently risk- sensitive.

Current gaps in the literature

Prevalence: Fewer than 10% of the 76 reviewed studies reported any form of prediction interval, confidence measures, or model calibration metric. Probabilistic outputs were typically reduced to deterministic class labels or point estimates for reporting.

Model-Specific Shortfalls: Classical GS methods like GBLUP naturally provide a measure of prediction error variance for each individual. In contrast, most ML/DL models in the reviewed studies such as standard Random Forest, SVM, XGBoost, CNNs) were implemented in their default, deterministic forms(Table 9). Their reported confidence is often an uncalibrated measure of model consensus, such as Random Forest proximity. This is not a statistically rigorous estimate of predictive uncertainty.

Synthesis of findings stratified by study quality

To assess the robustness of the trends synthesized in Sect. 3.1–3.8, findings were stratified by the methodological quality (QAMMLS ratings) of the included studies. Stratifying the synthesis by quality categories (High, Moderate, and Low) revealed a key patterns that contextualize the cross study conclusions. Consequently, the quantitative performance summaries presented in Tables 4 and 5 and the preceding sections should

Table 9 Comparison of uncertainty estimation capabilities in genomic prediction models

Model class	Typical output	Native uncertainty estimation?	Common methods for uncertainty (if applied)	Suitability for risk-sensitive breeding
Classical GS (GBLUP, rrBLUP)	Point Estimate + Variance	Yes (Prediction Error Variance - PEV)	Inherent from the mixed model equations.	High. Provides directly interpretable standard errors for each prediction.
Bayesian ML (e.g., Bayesian Neural Nets, Gaussian Processes)	Predictive Distribution	Yes (Credible Intervals)	Markov Chain Monte Carlo (MCMC), Variational Inference.	High. Quantifies epistemic (model) and aleatoric (data) uncertainty. Computationally expensive.
Frequentist ML (RF, XGBoost, SVM)	Point Estimate	No (by default)	Post-hoc methods: Jackknife, Bootstrap, Conformal Prediction.	Medium-Low. Requires additional, often complex, post-processing. Methods like Conformal Prediction show promise for generating distribution-free prediction intervals[74]
Standard Deep Learning (CNNs, MLPs)	Point Estimate	No (by default)	Post-hoc/Modified: Monte Carlo Dropout (approximate Bayesian), Deep Ensembles, Evidential Deep Learning.	Medium. Active research area. Methods like Deep Ensembles are effective but computationally costly.

be interpreted with the understanding that they amalgamate findings from studies of varying rigor, which significantly influences the reported magnitudes.

Performance claims in high vs. low-quality studies: High quality ($n = 18$, 29.5%), which employed rigorous validation strategies including independent test sets, cross population prediction) and detailed reporting of hyperparameters, generally reported more consistent performance improvements. For instance, the calculated average accuracy gains of ML over classical GS for low heritability traits in high quality studies was + 8.2% (range: +2% to + 15%), compared to calculated average of + 16.5% (range: +10% to + 25%) reported in low quality studies ($n = 11$, 18.0%). Low quality studies were more likely to use optimistic within population cross validation alone, inflating perceived performance. This discrepancy is partly explained by validation design. High quality studies were more likely to employ forward or cross- population validation, providing realistic performance estimates, while low quality studies relied almost exclusively on within- population CV.

Trends in algorithm performance by quality: The reported superiority of complex models like graphic neural networks (GNNs) and deep learning hybrids was predominately observed studies rated as high or moderate quality. These studies provided the necessary computational detail and biological validation to support claims of capturing nonlinear interactions. In contrast, claims from low quality studies often lacked the implementation detail needed for reproducibility.

Omics integration and diminishing returns: The observation of diminishing returns with tri-omics integration was most consistent a cross high quality studies. These studies systematically reported the computational cost benefit analysis. Several low quality studies advocating for quad – omics integration did not provide comparable cost metrics or sufficient evidence for the marginal gain, highlights a potential overstatement.

Conclusions

Based on the patterns and trade-offs observed in the reviewed literature and with particular weight given to the evidence from studies with higher methodological rigor, we synthesize the following evidence-informed considerations for breeders selecting ML approaches. This frame work directly links primary objectives, available resources, and trait biology to methodological choices. The first critical step is to define the primary goal. For discovery research aimed at mechanistic insights, graph based integration is optimal when biological networks are available, as it best uncovers gene-trait pathways. For applied prediction and selection, late fusion is favored for its interpretability in breeding decisions, while early fusion may be considered for maximizing predictive signal when omics layered correlated. The second step involves a realistic assessment of available data and computational resources. Sample size is a primary constraint. With small datasets ($N < 1{,}000$), robust tree base models (Random Forest, XGBoost) or classical genomic selection(GS) are recommended, while deep learning medium to large datasets ($N > 1{,}000$) and significant GPU capacity. Omics layers should be chosen strategically: genomics +transcriptomics excels for quality traits. Integrating more than two omics layers often yielding diminishing accuracy gains relative to sharply higher computational costs.

The third step is to consider the traits genetic architecture, guided by genomic prediction theory. For high - heritability traits with an additive genetic basis (example, many simply inherited disease resistance), classical GS methods are theoretically well suited and empirically, offering a perfect balance of accuracy, speed and interpretability. In contrast, for low- heritability, complex polygenic traits(example, stress resilience, yield in diverse environment) where non additive genetic variance and epistasis are hypothesized to play a larger role, ML/DL models have a sound theoretical basis for capturing these effects and should be prioritized, providing sufficient data and computational resources are available.

Finally, implementation must prioritize robust validation and interpretation. Cross population or forward validation is essential to assess real-world generalizability beyond optimistic within population metrics. This step is critical because a model's performance is contingent on the population LD structure and the transferability of learned genetic effects. Regardless of model complexity, applying standardized explainable AI (XAI) tools like SHARP is crucial to extract biological insights and build breeder trust. Ultimately, this theory- informed synthesis moves the field beyond benchmarking on a single dataset. It provides a conceptual framework for model selection: use the traits heritability and presumed genetic complexity (additive vs. non-additive) as primary filter, informed by genomic prediction theory, then let the prediction objective (within-versus across-population) and operational constraints guide the final choice. Community-wide efforts must now prioritize standardized data protocols, benchmarking platforms for fair comparison, and development of accessible, interpretable tools to bridge the gap between computational innovation and breeding practice. Cross population or forward validation is essential to assess real world generalizability beyond optimistic within population metrics. The choice of validation strategy is not merely a technical detail but fundamentally alters the interpretation of model's reported accuracy and its potential breeding value.

Challenges and future perspectives

To address interpretability challenges, we recommended standardizing the use of two model-agnostic frameworks. SHAP (SHapley Additive exPlanations) provides a unified, theoretical grounded measures of feature importance for both global and local analysis, for identifying key genes or metabolites across a population [75].

To address these challenges, we propose a three-pronged strategy for advancing ML applications in crop multi-omics:

Explainable AI Implementation: Integration of SHAP/LIME frameworks into breeding pipelines will be crucial for identifying robust biomarkers and building breeder confidence in ML predictions. This approach has shown promise in preliminary studies, revealing key stress-responsive genes that were subsequently validated experimentally.

Transfer Learning Frameworks: Development of pre-trained models on data-rich crops (maize, rice) with fine-tuning protocols for under-resourced crops could dramatically reduce data requirements. Early implementations have demonstrated 30–40% accuracy improvements in understudied species compared to training from scratch.

Scalable Computational Solutions: Cloud-based AutoML platforms such as Google Vertex AI and optimized model architectures can help democratize access to advanced analytics. Pilot programs have reduced computational barriers for small breeding programs while maintaining 90% of the predictive performance of complex custom models.

Emerging opportunities include the development of foundational models trained on large public omics datasets and the implementation of federated learning systems to enable secure, privacy-preserving data pooling across institutions. These innovations could overcome current limitations while addressing ethical concerns about data sharing.

Moving from point predictions to uncertainty- qualified decision:

A major translational gap identified is the lack of uncertainty- aware modeling. For ML to be trusted for high- stakes breeding decisions, it must provide not just a prediction, but a calibrated measures of its confidence. Future research must be prioritize.

Probabilistic and Bayesian ML: promoting the use of models that natively quantify uncertainty (Bayesian neural networks, Gaussian processes) and developing scalable approximations for large omics datasets.

1. *Model calibration*: routinely reporting calibration metrics to diagnose and correct cover/ under- confident predictions.
2. *Accessible uncertainty tools*: implementing and benchmarking post-hoc like conformal prediction, which can provide valid prediction intervals for any black-box model (a trained XGBoost or CNN) without requiring retraining, making it highly practical for breeding without requiring retraining, making it highly practical for breeding baseline [76].
3. *Standardized reporting*: Journal and reviewers should encourage/ require the reporting of prediction intervals or confidence measures alongside point estimates of accuracy.

6. Synthesis and practical recommendations.

The review findings suggest several actionable insights for the plan breeding community as follows:

Algorithm selection: Evidence suggests that random forest offers a practical balance for small-scale studies. For major crops with large datasets (> 5,000 samples), the reviewed

literature indicates that deep learning can achieve performance gains that may justify computational investment.

Omics prioritization: strategic focus on the most informative omics combinations for target traits (genomics + transcriptomics for stress response) c can optimize resource allocation.

Targeted hybrid approaches: The promise of hybrid models should be pursued with a clear strategy based on the primary objective. For maximize accuracy on complex, high dimensional traits such as multi-omics yield prediction, feature-engineering hybrids including autoencoder → XGBoost) have delivered the most consistent gains (10–15%) in literature and are recommended. To improve robustness across environments and populations, stacking ensembles of diverse models may be more suitable, through their accuracy gains are less predictable. Integrated hybrids (CNN-LSTM) are power full but specialized solutions best reserved for data with explicit temporal or spatial structures. This balances the representational power of deep learning with the feature importance interpretability crucial for breeding.

Investment priorities: development of XAI tools, cloud based solutions, and standardized uncertainty quantification methods should be prioritized to address critical interpretability, accessibility, and decision- risk barriers.

Promote equitable innovation: For breeding programs focused under resourced strategic transfer learning and explore federating learning consortia to overcome data scarcity. Investment in public, pre trained model repositories for key omics data types in a critical public good crops, prioritize.

Limitation of quantitative synthesis

A core limitation is the methodological heterogeneity of studies, which as detailed in Sect. 3.9. Directly influences the reported performance metrics. While our narrative synthesis and quality stratification mitigates this, readers are cautioned that numerical trends are not definitive pooled effects. The wide variation in crops, traits, omics layers, ML model implementations, and validation protocols means the body of evidence is highly heterogeneous. Consequently, we deliberately abstained from performing a formal meta-analysis such as calculating a single pooled effect size would be statistically inappropriately and misleading. All numerical performance summarizes (examples ~ 12% average improvement, ΔR^2 of 0.20) presented in this review are narrative descriptions of the trends found in the literature. This summarizes what individuals studies have reported but do not represents a statistically robust, cross- study average. Readers are cautioned against interpreting these numbers as definitive, comparable effect sizes. This underscores the need for future research with standardized benchmarking protocols to allow for true quantitative comparison.

Author contributions

A.T. Study conception and design, Search strategy development and execution, Data extraction and curation and using the custom QAMMLS tool, with final consensus scoring. D.M. Study screening and selection, performed independent title/abstract and full-text screening. All discrepancies were resolved through consensus discussion between both authors contributed to the quality assessment and risk-of-bias evaluation.

Funding

No funding.

Data availability

Not applicable. This manuscript does not report data generation.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare no competing interests.

Received: 2 August 2025 / Accepted: 30 March 2026
Published online: 01 April 2026

References

1. Ray DK, Mueller ND, West PC, Foley JA. Yield trends are insufficient to double global crop production by 2050. Plos One. 2013;8(6):e66428.
2. Razzaq A, Kaur P, Akhter N, Wani SH, Saleem F. Next-generation breeding strategies for climate-ready crops. Front Plant Sci. 2021;12:620420.
3. Varshney RK, Terauchi R, McCouch SR. Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. Plos Biol. 2014;12(6):e1001883.
4. Yan J, Wang X. Machine learning bridges omics sciences and plant breeding. Trends Plant Sci. 2023;28(2):199–210.
5. Guyomarc'h S, Boutté Y, Laplaze L. AP2/ERF transcription factors orchestrate very long chain fatty acid biosynthesis during Arabidopsis lateral root development. Mol Plant. 2021;14(2):205–7.
6. Vangheluwe N, Swinnen G, De Koning R, Meyer P, Houben M, Huybrechts M, Sajeev N, Rienstra J, Boer D. Give CRISPR a chance: the GeneSprout initiative. Trends Plant Sci. 2020;25(7):624–7.
7. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321–32.
8. Agamah FE, Bayjanov JR, Niehues A, Njoku KF, Skelton M, Mazandu GK, et al. Computational approaches for network-based integrative multi-omics analysis. Front Mol Biosci. 2022;9:967205.
9. Hickey LT, Hafeez AN, Robinson H, Jackson SA, Leal-Bertioli SC, Tester M, Gao C, Godwin ID, Hayes BJ, Wulff BB. Breeding crops to feed 10 billion. Nat Biotechnol. 2019;37(7):744–54.
10. Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet. 2014;127(3):595–607.
11. Ray DK, West PC, Clark M, Gerber JS, Prishchepov AV, Chatterjee S. Climate change has likely already affected global food production. Plos One. 2019;14(5):e0217148.
12. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, De Los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22(11):961–75.
13. Weckwerth W, Ghatak A, Bellaire A, Chaturvedi P, Varshney RK. PANOMICS meets germplasm. Plant Biotechnol J. 2020;18(7):1507–25.
14. Thingujam D, Gouli S, Cooray SP, Chandran KB, Givens SB, Gandhimeyyan RV, Tan Z, Wang Y, Patam K, Greer SA. Climate-resilient crops: integrating AI, multi-omics, and advanced phenotyping to address global agricultural and societal challenges. Plants. 2025;14(17):2699.
15. Li X, Li Y, Sun Y, Li S, Cai Q, Li S, Sun M, Yu T, Meng X, Zhang J. Integrating genetic diversity and agronomic innovations for climate-resilient maize systems. Plants. 2025;14(10):1552.
16. Misra BB, Langefeld C, Olivier M, Cox LA. Integrated omics: tools, advances and future approaches. J Mol Endocrinol. 2019;62(1):R21–45.
17. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, 2021. 372.
18. Pranckutė R. Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world. Publications. 2021;9(1):12.
19. Takács G, Gergely I, Ördög V. A búza (Triticum aestivum L.) vízigénye és a vízhiány hatása a növényre. Acta Agronomica Óváriensis. 2021;62(2):116-40.
20. Ayanlade TT, Jones SE, Laan LVd, Chattopadhyay S, Elango D, Raigne J, Saxena A, Singh A, Ganapathysubramanian B, Singh AK. *Multi-modal AI for Ultra-Precision Agriculture*, in *Harnessing Data Science for Sustainable Agriculture and Natural Resource Management*. Springer; 2024. pp. 299–334.
21. Westergaard D, Stærfeldt H-H, Tønsberg C, Jensen LJ, Brunak S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. Plos Comput Biol. 2018;14(2):e1005962.
22. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol insights. 2020;14:1177932219899051.
23. Shoaib M, Shah B, Sayed N, Ali F, Ullah R, Hussain I. Deep learning for plant bioinformatics: an explainable gradient-based approach for disease detection. Front Plant Sci. 2023;14:1283235.
24. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ, 2009. 339.
25. Stodden V, McNutt M, Bailey DH, Deelman E, Gil Y, Hanson B, Heroux MA, Ioannidis JP, Taufer M. Enhancing reproducibility for computational methods. Sci. 2016;354(6317):1240–1.
26. Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B, Davies P, Kleijnen J, Churchill R. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225–34.
27. Landis JR, Koch GG. *The measurement of observer agreement for categorical data.* Biometrics, 1977: pp. 159–174.

28. Li T, Higgins JP, Deeks JJ. *Collecting data.* Cochrane handbook for systematic reviews of interventions, 2019: pp. 109–141.
29. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. Plos Med. 2010;7(2):e1000217.
30. Khaki S, Wang L. Crop yield prediction using deep neural networks. Front Plant Sci. 2019;10:621.
31. Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19:3735–46.
32. Prinja N, Narain R. ABO, Rh, and kell blood group antigen frequencies in blood donors at the tertiary care hospital of Northwestern India. Asian J Transfus Sci. 2020;14(2):179–84.
33. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006;7(10):781–91.
34. Azodi CB, Bolger E, McCarren A, Roantree M, de Los G, Campos, Shiu S-H. Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3: Genes, Genomes, Genetics, 2019. 9(11): pp. 3691–3702.
35. Norman A, Taylor J, Edwards J, Kuchel H. Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. G3: Genes Genomes Genet. 2018;8(9):2889–99.
36. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83.
37. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol. 2008;8(1):45.
38. Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artif Intell Med. 2020;104:p101822.
39. Wallace JG, Rodgers-Melnick E, Buckler ES. On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. Annu Rev Genet. 2018;52(1):421–44.
40. Moons KG, Damen JA, Kaul T, Hooft L, Navarro CA, Dhiman P, Beam AL, Van Calster B, Celi LA, Denaxas S. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ, 2025. 388.
41. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18(12):e323.
42. Singh B, Kumar S, Elangovan A, Vasht D, Arya S, Duc NT, Swami P, Pawar GS, Raju D, Krishna H. Phenomics based prediction of plant biomass and leaf area in wheat using machine learning approaches. Front Plant Sci. 2023;14:1214801.
43. Wu C, Luo J, Xiao Y. Multi-omics assists genomic prediction of maize yield with machine learning approaches. Mol Breeding. 2024;44(2):14.
44. Heslot N, Jannink JL, Sorrells ME. Perspectives for genomic selection applications and research in plants. Crop Sci. 2015;55(1):1–12.
45. Zhu T, Xia C, Yu R, Zhou X, Xu X, Wang L, Zong Z, Yang J, Liu Y, Ming L. Comprehensive mapping and modelling of the rice regulome landscape unveils the regulatory architecture underlying complex traits. Nat Commun. 2024;15(1):6562.
46. Gai Y, Liu S, Zhang Z, Wei J, Wang H, Liu L, Bai Q, Qin Q, Zhao C, Zhang S. Integrative approaches to soybean resilience, productivity, and utility: a review of genomics, computational modeling, and economic viability. Plants. 2025;14(5):671.
47. Danilevicz MF, Upadhyaya SR, Batley J, Bennamoun M, Bayer PE, Edwards D. Understanding plant phenotypes in crop breeding through explainable AI. Plant Biotechnol J. 2025.
48. Dong Y, Li G, Zhang X, Feng Z, Li T, Li Z, Xu S, Xu S, Liu W, Xue J. Genome-wide association study for maize hybrid performance in a typical breeder population. Int J Mol Sci. 2024;25(2):1190.
49. Zhou G, Gao J, Zuo D, Li J, Li R. MSXFGP: combining improved sparrow search algorithm with XGBoost for enhanced genomic prediction. BMC Bioinformatics. 2023;24(1):384.
50. Crossa J, Perez P, Hickey J, Burgueno J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity. 2014;112(1):48–60.
51. Patil S, Kolhar S, Jagtap J. Deep learning in plant stress phenomics studies–a review. Int J Comput Digit Syst. 2024;16(1):305–16.
52. Araujo SO, Peres RS, Ramalho JC, Lidon F, Barata J. Machine learning applications in agriculture: current trends, challenges, and future perspectives. Agronomy. 2023;13(12):2976.
53. Zhao M, Cang H, Chen H, Zhang C, Yan T, Zhang Y, Gao P, Xu W. Determination of quality and maturity of processing tomatoes using near-infrared hyperspectral imaging with interpretable machine learning methods. Lwt. 2023;183:114861.
54. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–23.
55. McNinch C, Chen K, Dennison T, Lopez M, Yandeau-Nelson MD, Lauter N. A multigenotype maize silk expression atlas reveals how exposure-related stresses are mitigated following emergence from husk leaves. Plant Genome. 2020;13(3):e20040.
56. van Dijk ADJ, Kootstra G, Kruijer W, de Ridder D. Machine learning in plant science and plant breeding. Isci. 2021. 24(1).
57. Bennett AB, Pankievicz VC, Ané J-M. A model for nitrogen fixation in cereal crops. Trends Plant Sci. 2020;25(3):226–35.
58. Li S, Hua H, Chen S. Graph neural networks for single-cell omics data: a review of approaches and applications. Brief Bioinform, 2025. 26(2).
59. Castells-Graells R, Lomonossoff GP. Plant-based production can result in covalent cross-linking of proteins. Plant Biotechnol J. 2021;19(6):1095.
60. Zhang H, Yin L, Wang M, Yuan X, Liu X. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Front Genet. 2019;10:189.
61. Shaar-Moshe L, Hübner S, Peleg Z. Identification of conserved drought-adaptive genes using a cross-species meta-analysis approach. BMC Plant Biol. 2015;15(1):111.
62. Li L, Gallei M, Friml J. Bending to auxin: fast acid growth for tropisms. Trends Plant Sci. 2022;27(5):440–9.
63. Li R, Shi C-L, Wang X, Meng Y, Cheng L, Jiang C-Z, Qi M, Xu T, Li T. Inflorescence abscission protein SlIDL6 promotes low light intensity-induced tomato flower abscission. Plant Physiol. 2021;186(2):1288–301.
64. Chen W, Gao Y, Xie W, Gong L, Lu K, Wang W, Li Y, Liu X, Zhang H, Dong H. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet. 2014;46(7):714–21.
65. Lohia R, Salari R, Brannigan G. Sequence specificity despite intrinsic disorder: how a disease-associated Val/Met polymorphism rearranges tertiary interactions in a long disordered protein. Plos Comput Biol. 2019;15(10):e1007390.

66. Xu J, Shang L, Wang J, Chen M, Fu X, He H, Wang Z, Zeng D, Zhu L, Hu J. The SEEDLING BIOMASS 1 allele from indica rice enhances yield performance under low-nitrogen environments. Plant Biotechnol J. 2021;19(9):1681.
67. Leister D, Marino G, Minagawa J, Dann M. An ancient function of PGR5 in iron delivery? Trends Plant Sci. 2022;27(10):971–80.
68. Gianola D, Fernando RL, Stella A. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics. 2006;173(3):1761–76.
69. de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013;193(2):327–45.
70. Pérez-Enciso M, Zingaretti LM. A guide on deep learning for complex trait genomic prediction. Genes. 2019;10(7):553.
71. Meuwissen TH, Hayes BJ, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–29.
72. Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Barrón-López JA, Martini JW, Fajardo-Flores SB, Gaytan-Lugo LS, Santana-Mancilla PC, Crossa J. A review of deep learning applications for genomic selection. BMC Genomics. 2021;22(1):19.
73. Peng RD. Reproducible research in computational science. Sci. 2011;334(6060):1226–7.
74. Vovk V, Gammerman A, Shafer G. Algorithmic learning in a random world. Springer; 2005.
75. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67.
76. Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511, 2021.

Publisher's note