Received: July 30, 2026. Revised: September 24, 2026. 1116

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Multi-validation Benchmark of Selected SNP Representation-model Configurations for Maize Genomic Prediction

Rizki Darmawan[1]* Heru Agus Santoso[1] Abdul Syukur[1] Affandy[1]

[1]*Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia*
* Corresponding author's Email: P41202400082@mhs.dinus.ac.id

Abstract: Maize genomic prediction can accelerate breeding by estimating phenotypic performance from genome-wide markers, but conclusions can depend on the paired choice of learning model, SNP input, phenotype construction, and validation design. This study benchmarks selected model-representation configurations rather than estimating independent representation effects. The initial Genomes to Fields (G2F) genotype resource contained 5,733 hybrids and 2,156 SNPs, while the primary pollen days-after-planting (DAP) benchmark used 4,767 phenotype-matched hybrids. Eight representation families generated nine evaluated feature matrices because PCA100 and PCA300 were treated as distinct inputs, and all 1,945 planned runs were completed. Across ten common 80/20 holdouts, RF-PCA100 achieved the lowest mean RMSE of 2.3982 (R2 = 0.6411; Pearson = 0.8019), but corrected resampled tests with Holm adjustment found no significant RMSE difference from RF-additive, RF-MAF-weighted, or RF-recessive. A fixed-effect phenotype sensitivity analysis changed the ranking: GBLUP-additive led on adjusted BLUEs (RMSE = 1.0282; R2 = 0.8516), demonstrating sensitivity to target construction. Mean R2 was negative for every evaluated parent-disjoint configuration. HGB genomic-plus-environmental prediction achieved positive R2 in only 6/10 held-out years, 17/41 locations, 64/217 location-years, and 2/5 climatic regimes. On the same 2024 yield test records, a historical-location mean baseline (R2 = 0.0761; unseen hybrids = 0.0700) outperformed the best learned configuration (-0.0244 and -0.0451). These results support configuration-level claims and emphasize phenotype sensitivity and distribution-shift validation before operational use.

Keywords: Genomic prediction, Model-representation configuration, SNP encoding, Maize, Phenotype adjustment, Distribution shift, Multi-validation.

1. Introduction

Maize supports food, feed, and industrial production worldwide. As demand rises and climate conditions become more variable, breeding programs need faster ways to achieve genetic improvement [1]. Genomic prediction addresses this need by estimating phenotypic or breeding values from genome-wide markers before multi-season evaluation is complete [2-4]. The availability of dense single nucleotide polymorphism (SNP) data and Genomes to Fields (G2F) resources now makes it possible to study maize genotype-to-phenotype prediction across many hybrids, traits, environments, and years [5-7].

Genomic prediction draws on classical genomic models, machine learning, and deep learning. RR-BLUP, GBLUP, and kernel methods provide well-established shrinkage- or relationship-based baselines, whereas random forests and boosting can capture nonlinear marker patterns [4]. Deep networks offer greater representational flexibility, but their performance depends strongly on sample size, tuning, architecture, and input representation

[8-10]. Model choice and SNP representation therefore need to be considered together.

Different SNP representations emphasize different aspects of genetic variation. Additive dosage is simple and easy to interpret, while dominant, recessive, and heterozygous encodings highlight specific genetic contrasts. MAF-weighting adds information about allele frequency. One-hot and presence-absence-aware encodings preserve genotype categories but expand the feature space

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

[11, 12], whereas PCA compresses correlated markers at the cost of direct biological interpretability [13]. Recent systems, including Cropformer, AutoGP, WheatGP, DPCformer, and GP-WAITER, also suggest that architecture and marker representation should be evaluated together [14-18]. Their relative performance can change across traits, environments, and input types [19-22].

A recurring limitation in the literature is that many representation-model comparisons use only a small model set, a single trait, or one fixed holdout split [23]. This becomes important when the leading configurations differ only slightly, because their ranking may change with the data split. Such designs also provide limited evidence about transfer to unseen parental backgrounds, unseen environments, or future years, even though these are central challenges in practical breeding [24-26].

To provide a more reliable empirical comparison, this study established a multi-validation benchmark of predefined model-representation configurations. Eight SNP representation families-additive, dominant, recessive, heterozygous, MAF-weighted, one-hot, PCA-based, and nucleotide-aware DPC8-generated nine evaluated feature matrices because PCA100 and PCA300 were treated as distinct inputs. These inputs were paired with selected classical genomic, machine-learning, and deep-learning models according to a predefined experiment plan. The design was not a complete factorial crossing and therefore does not estimate independent representation, model, or interaction effects.

The principal contribution is a common multi-validation framework for comparing selected model-representation configurations across random, trait-specific, parental, environmental, and temporal settings. The framework combines predictive performance, dimensionality, runtime, phenotype-construction sensitivity, corrected statistical inference, explicit baselines, and distribution-shift evaluation across 1,945 completed runs.

The study aimed to identify competitive evaluated configurations, quantify their accuracy-efficiency trade-offs, determine whether their relative rankings remain stable across traits and phenotype definitions, and test whether strong within-distribution results transfer to unseen parents, environments, locations, and years. All conclusions are therefore restricted to the configurations and prediction tasks evaluated here.

2. Related works

Classical genomic prediction models remain important references because their shrinkage and relationship-based estimates are transparent and well understood [2, 3]. Tree-based ensembles can capture nonlinear thresholds and interactions without requiring the large parameter counts of deep networks. In moderate-sized tabular genomic datasets, random forests and boosting have often matched or outperformed deep-learning models [19].

Dimensionality reduction can improve computational efficiency and reduce collinearity [13]. Dominance-aware and environmental kernels can also improve prediction in multi-environment settings [27]. Results from multi-trait and multi-environment deep learning [20], local convolutional networks [21], and studies that combine structural and single-nucleotide variants [22] show that model rankings can change with the trait, input representation, and environmental structure. Several other approaches provide useful points of comparison, including GBLUP and Bayesian neural models [28], kernel regression [29], classification algorithms [30], and radial-basis-function networks [31] . Studies using principal-component integration [13], multi-trait prediction [32], and transcriptome-based prediction in maize [33] further support the need to evaluate representation and model choice together.

More recent architectures, such as Cropformer [14], WheatGP [16], DPCformer [17], and GP-WAITER [18], combine local convolution with mechanisms for learning longer-range sequence relationships. DPCformer is especially relevant here because its eight-dimensional representation encodes the two alleles reconstructed from REF/ALT metadata using nucleotide-level channels.

Table 1 positions the present work against classical genomic prediction, dimensionality reduction, machine-learning, deep-learning, and representation-specific studies. The final column now states the method-specific limitation that motivates the present configuration-level and multi-validation comparison.

3. Materials and methods

3.1 Dataset and phenotype construction

This study used publicly available maize genotype-by-environment resources from the Genomes to Fields (G2F) Initiative [34]. The initial genotype resource contained 5,733 hybrids and 2,156 chromosome-position-ordered SNP markers. The raw 2014-2023 phenotype file contained 173,960 plot records. After genotype alignment, non-missing trait filtering, and averaging repeated measurements within each hybrid-environment-trait combination, the nine retained traits comprised 801,164

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Table 1. Positioning of prior genomic-prediction approaches and method-specific limitations motivating the present benchmark

Study (year)	Representation / model	Method-specific limitation / gap motivating this study
Meuwissen et al. [2] (2001)	Genomic selection / marker-based prediction	Linear/shrinkage baseline; no modern nonlinear-model or SNP-input comparison.
Crossa et al. [3] (2017)	Genomic selection models	Review article; not a controlled encoding/model experiment.
Costa-Neto et al. [27] (2021)	Kernels, dominance, envirotyping	GxE/envirotyping focus; not a broad configuration benchmark.
Manthena et al. [13] (2022)	Dimensionality reduction / PCA	PCA focus; limited biological or nucleotide-aware input comparison.
Gill et al. [19] (2022)	Machine learning and deep learning	Soybean/model-family focus; no maize multi-validation benchmark.
Mora-Poblete et al. [20] (2023)	Multi-trait and multi-environment DL	Maize GxE DL; limited input and model-family breadth.
Wang et al. [11]; Krusenbaum and Wissuwa [12] (2025)	Expanded one-hot / PAV-aware encoding	Representation-specific studies in other species/crops; no common maize protocol.
Cropformer [14]; WheatGP [16] (2025)	CNN-attention / CNN-LSTM	Architecture-specific pipelines confound model, preprocessing, and input effects.
DPCformer [17]; GP-WAITER [18] (2025-2026)	8D nucleotide / weighted-embedding Transformer	Feature-selection/embedding is integral; representation is not independently isolated.
This study (2026)	Eight representation families; classical, ML and DL panel; multi-validation	Selected non-factorial configurations; claims restricted to evaluated pairings.

Table 2. Trait inventory after genotype alignment and within-environment aggregation

Trait	Hybrid-environment observations	Hybrids	Environments	Mean ± SD
Pollen_DAP_days	85113	4767	229	66.372 ± 8.060
Ear_Height_cm	95721	4771	253	107.477 ± 24.107
Grain_Moisture	102933	4772	272	18.338 ± 5.165
Plant_Height_cm	94677	4771	250	228.961 ± 33.201
Root_Lodging_plants	71728	4750	210	3.787 ± 10.435
Silk_DAP_days	83418	4763	225	67.693 ± 8.176
Stalk_Lodging_plants	79158	4763	227	4.482 ± 8.959
Twt_kg_m3	85404	4767	225	732.509 ± 44.472
Yield_Mg_ha	103012	4772	272	9.590 ± 2.982

hybrid-environment observations. Pollen DAP contributed 85,113 hybrid-environment observations from 4,767 hybrids and 229 environments; other trait-specific analytical populations ranged from 4,750 to 4,772 hybrids and 210 to 272 environments. Thus, 5,733 denotes the initial genotyped population, not the sample size used directly in every benchmark.

For the primary pollen-DAP benchmark, each hybrid-environment value was first formed by averaging available plot records within that environment, and the primary hybrid-level target was then the equal-weight mean across the environments available for that hybrid. Missing phenotype values were excluded trait by trait rather than imputed. Genotype preprocessing was estimated within the training partition: samples with more than 50% missing markers and SNPs with more than 20% missingness or MAF below 0.01 were excluded, remaining missing dosages were imputed by the training-marker mean, and allele frequencies, standardization parameters, and PCA loadings were fitted only on training data. The 2024 yield analysis remained a separate temporal stress test and was never pooled with the 2014-2023 target construction.

To test whether unequal environmental sampling affected the primary target, a prespecified phenotype sensitivity analysis was added using 128,401 genotype-aligned plot-level pollen-DAP records.

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

A fixed-effect model y = mu + Hybrid + Environment + Replicate (Environment) + Block (Environment, Replicate) + error was solved by alternating least squares under weighted sum-to-zero constraints. It converged after 1,640 iterations (residual SD = 1.815) and produced adjusted hybrid BLUEs for the same 4,767 hybrids. The environment-equal primary target and adjusted BLUEs correlated at r = 0.4455. Approximate inverse-sampling-variance weights were also archived for 4,453 hybrids with at least five observations from at least three environments; these are precision weights, not formal mixed-model reliabilities.

Table 3 consolidates the analytical unit, partition size, and source structure for each validation setting so that results from different prediction tasks are not interpreted as directly interchangeable.

3.2 SNP quality control and representation

Let X = [*xij*] denote the genotype matrix, where *xij* is the allele dosage of SNP j for individual i and *xij* in {0, 1, 2} represents REF/REF, REF/ALT, and ALT/ALT. The study distinguishes eight representation families from nine evaluated feature matrices: additive, dominant, recessive, heterozygous, MAF-weighted, one-hot, PCA100, PCA300, and DPC8. PCA is one representation family but contributes two matrices. To prevent leakage, all data-dependent preprocessing parameters-including missing-value imputation, allele-frequency estimation, standardization, environmental preprocessing, and PCA loadings-were estimated from training data and then applied unchanged to the corresponding test data.

3.2.1. Additive encoding [2, 3, 5]

Additive encoding was used as the baseline dosage representation because allele dosage is the conventional input form in genomic selection and genome-wide marker-based prediction. It preserves the original dosage value

$$\textit{Additive}: \qquad z_{ij}^{add} = x_{ij}, \;\; x_{ij} \in \{0,1,2\} \qquad (1)$$

where x_{ij} is the allele dosage of SNP j for individual i. REF/REF, REF/ALT, and ALT/ALT are therefore coded as 0, 1, and 2. This representation assumes an ordered relationship among genotype states, with the encoded value increasing as the number of alternative alleles rises.

3.2.2. Dominant encoding [36]

Dominant encoding was used to capture the presence of at least one alternative allele and is related to dominance-aware genomic prediction. It is defined as:

$$\textit{Dominant}: \qquad z_{ij}^{dom} = I(x_{ij} \geq 1) \qquad (2)$$

REF/REF is coded as 0, while REF/ALT and ALT/ALT are both coded as 1. The model can therefore test whether carrying one or two copies of the alternative allele provides similar predictive information.

Table 3. Consolidated validation settings and effective analytical sample sizes

Validation task / target	Split and train/test	Structure
Primary holdout - Pollen DAP hybrid mean	10 shared 80/20 splits; 3,800-3,817 / 950-955 hybrids	229 env; 2014-2023
Parent-disjoint - Pollen DAP hybrid mean	10 strict parent splits; 1,946-2,044 / 461-559 hybrids	0 shared parents; 2014-2023
Multi-trait - eight additional hybrid traits	Repeated holdouts; 4,750-4,772 hybrids before split	210-272 env; 2014-2023
Within-distribution GxE - hybrid-env records	Unseen-genotype split; 64,629-65,788 / 15,720-16,879 records	EC-supported 2014-2023 env
Legacy LOEO - hybrid-env records	8 held-out env x 10 seeds; 80,339-80,600 / 908-1,169 records	One env excluded per split
Grouped transfer - pollen-DAP plot records	Hold out 10 years, 41 locations, 217 location-years, 5 climate regimes	Test/group: 8,067-15,131; 87-9,290; 43-1,588; 3,678-46,115
Temporal 2024 yield - hybrid-env records	2014-2023 -> 2024; 91,135 / 9,064 records	21 test env; 1,063 hybrids
2024 unseen hybrids - yield records	2014-2023 -> unseen 2024; 91,135 / 7,963 records	21 env; 959 unseen hybrids
Phenotype sensitivity - original mean vs BLUE	10 shared 80/20 splits; ~3,813 / ~954 hybrids	Same 4,767 hybrids

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

3.2.3. Recessive encoding [36]

Recessive encoding was used to represent effects that appear only when two copies of the alternative allele are present. It is defined as:

$$\textit{Recessive:} \qquad z_{ij}^{rec} = I(x_{ij} = 2) \tag{3}$$

Only ALT/ALT is coded as 1; REF/REF and REF/ALT are coded as 0. This representation tests whether homozygous alternative genotypes contain predictive information that is not captured as clearly by additive dosage.

3.2.4. Heterozygous encoding [37]

Heterozygous encoding was used to isolate heterozygosity-specific information. It is defined as:

$$\textit{Heterozygous:} \qquad z_{ij}^{het} = I(x_{ij} = 1) \tag{4}$$

where only REF/ALT genotypes are coded as 1 and both homozygous states are coded as 0. This encoding evaluates whether the heterozygous state carries predictive information that is not captured by additive, dominant, or recessive encodings.

3.2.5. MAF-weighted encoding [38]

MAF-weighted encoding was used to standardize allele dosage according to allele frequency and marker variance. It is defined as:

$$z_{ij}^{maf} = \frac{x_{ij} - 2p_j}{\sqrt{2p_j(1-p_j)}} \tag{5}$$

where p_j is the allele frequency of SNP j estimated from the training data. The term $2p_j$ is the expected allele dosage, and $\sqrt{(2p_j(1 - p_j))}$ scales the marker by its expected variance. This transformation places markers with different allele frequencies on a comparable scale.

3.2.6. One-hot encoding [11, 12]

One-hot encoding represents each genotype category as a separate binary indicator. Similar expanded categorical representations have been explored in recent genomic prediction studies. It is defined as:
One-hot:

$$z_{ij}^{onehot} = \begin{bmatrix} I(x_{ij} = 0), I(x_{ij} = 1), \\ I(x_{ij} = 2) \end{bmatrix} \tag{6}$$

Unlike additive dosage, one-hot encoding does not assume that REF/REF, REF/ALT, and ALT/ALT lie on an ordered numeric scale. The trade-off is a threefold increase in dimensionality: 2,156 SNPs become 6,468 binary features.

3.2.7. PCA-based encoding [13]

PCA-based encoding projects the standardized additive marker matrix into a lower-dimensional orthogonal feature space, reducing redundancy among correlated SNPs while retaining most of the informative genomic variation for prediction. It is defined:

$$\textit{PCA:} \qquad Z_K^{PCA} = \tilde{X} W_K, \quad K \in \{100,300\} \tag{7}$$

where $\tilde{X}$ is the standardized additive genotype matrix, W_k is the PCA loading matrix, and K is the number of retained principal components. We used K = 100 and K = 300 to create PCA100 and PCA300. PCA100 reduced the original 2,156-marker space to 100 components, a dimensionality reduction of 95.36%.

3.2.8. DPC8 nucleotide-aware encoding [17]

DPC8 nucleotide-aware encoding was used as an input representation inspired by DPCformer [17]. It reconstructs the two alleles of each genotype using REF and ALT metadata. The evaluated DPCformer-inspired systems are controlled adaptations under the common protocol and are not claimed as exact reproductions of the original DPCformer pipeline.

$$\textit{DPC8:} \qquad z_{ij}^{DPC8} = [e(a_{ij1}), e(a_{ij2})] \tag{8}$$

where a_{ij1} and a_{ij2} denote the two reconstructed alleles of SNP j for individual i, and e(·) denotes a four-channel nucleotide one-hot vector for A, C, G, and T. Thus, each SNP is represented by eight values, corresponding to two alleles multiplied by four nucleotide channels. Since the dataset contained 2,156 SNPs, DPC8 produced 17,248 input values. Because the genotype data were unphased, heterozygous genotypes were encoded using a canonical REF/ALT order and were not interpreted as phased haplotypes.

3.3 Prediction models and fair training protocol

The benchmark evaluated predefined configurations drawn from RR-BLUP, GBLUP, LASSO, Elastic-Net, RF, HGB, three MLP depths, a residual MLP, a chromosome-ordered 1D-CNN, a Cropformer-inspired CNN-attention adaptation, and DPCformer-inspired dosage and 8D adaptations.

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

The Cropformer and DPCformer systems were architecture-level controlled adaptations: the original Cropformer optimization/interpretation workflow and the DPCformer MIC/PMF feature-selection pipeline were not reproduced. Accordingly, this paper makes no superiority or inferiority claim about the original published methods [14, 17].

All model-selection procedures were confined to each outer-training partition. RR-BLUP, LASSO, and Elastic-Net used three-fold inner cross-validation. GBLUP used K = ZZ^T/m after standardizing the selected feature matrix Z and selected lambda from {0.01, 0.1, 1, 10} on a 15% inner-validation subset. For additive GBLUP, Z was the standardized 0/1/2 dosage matrix; a one-hot diagnostic configuration used a separately standardized one-hot matrix and therefore a different K. Trait-level GBLUP rankings reported in this paper use the additive configuration. RF used 500 trees, max_features = 0.5, and minimum leaf size 1. HGB and neural settings are summarized in Table 4; the outer test set was never used for tuning or checkpoint selection.

The experiment did not use a full factorial crossing of every SNP representation with every learning model. Each evaluated pairing is therefore treated as a complete model-representation configuration. Differences between configurations that simultaneously change both components cannot be attributed exclusively to the representation or to the learning algorithm. Representation-specific statements are restricted to within-model comparisons using the same validation split. Independent model, representation, and interaction effects are outside the scope of this design.

The archived full experiment was executed in Google Colab with Python 3.12.13, NumPy 2.0.2, pandas 2.2.2, SciPy 1.16.3, scikit-learn 1.6.1, Matplotlib 3.10.0, and PyTorch 2.11.0+cpu. The archived environment reports CPU execution and no CUDA device; runtime values should therefore be interpreted as environment-specific CPU measurements rather than GPU benchmarks. Exact seeds, split identifiers, selected hyperparameters, software information, predictions, and per-run details are stored in the reproducibility archive.

Table 4 provides a self-contained summary of the principal hyperparameters and tuning rules. Full candidate grids, training histories, split identifiers, and per-run selected values are retained in the versioned experiment files.

3.4 Validation framework

The main comparison used ten shared 80/20 holdout splits with seeds 2, 42, 123, 777, 1001, 2024, 2025, 2026, 31415, and 27182. The earlier fixed split of 3,815 training and 952 test hybrids was retained only for continuity. The ten random test sets overlap because they are resampled from the same 4,767 hybrids; their mean pairwise Jaccard overlap was 0.112, and 4,254 hybrids appeared in at least one test set, with a maximum of seven appearances. Parent-disjoint validation assigned parents to only one side and retained only hybrids whose two parents belonged to the same partition; across all ten splits there were zero shared parents, while mixed-parent crosses were excluded. A selected panel was then evaluated across the other eight traits.

Table 4. Reproducibility settings for the evaluated model families

Model family	Principal settings / tuning
RR-BLUP	RidgeCV alpha {0.01, 0.1, 1, 10, 100, 1000}; 3-fold inner CV
GBLUP	lambda {0.01, 0.1, 1, 10}; 15% inner validation; conjugate-gradient solver
LASSO	16 log-spaced alphas 1e-4 to 10^0.5; 3-fold CV; max_iter 10,000
Elastic-Net	12 log-spaced alphas 1e-4 to 10^0.5; l1 ratios {0.1, 0.5, 0.9}; 3-fold CV
Random Forest	500 trees; max_features 0.5; min_samples_leaf 1
HGB	learning_rate 0.05; max_iter 300; max_leaf_nodes 31; L2 0.10; 15% internal validation; patience 30
MLP family	64 / 128-64 / 256-128-64; dropout 0.20; AdamW lr 1e-3; weight decay 1e-5; batch 64; max 300 epochs; patience 30
Residual MLP / 1D-CNN	ResMLP width 128, 3 blocks; CNN 32-64, kernel 7; common neural controls
Cropformer-inspired	hidden 64; 4 heads; kernel 3; dropout 0.50/0.20; Adam lr 1e-3; batch 32; max 150 epochs; patience 20
DPCformer-inspired	Res-CNN 32-64-64, kernel 5; 4 heads; 1 Transformer layer; FF 128; dropout 0.20; no MIC/PMF selection

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Genotype-by-environment analyses used hybrid-environment records and compared genomic-only, environment-only, genomic-plus-environmental, and explicit interaction inputs. The original GxE panel used training-fitted PCA summaries of genomic and 654 environmental covariates. Because environment covariates can also carry environment-specific signatures, targeted grouped transfer analyses were added using plot-level pollen-DAP records: leave-one-year-out, leave-one-location-out, leave-one-location-year-out, and leave-one-climatic-regime-out. The grouped models used 16 training-fitted genomic PCs and 16 training-fitted environmental PCs. Five climatic regimes were constructed by standardizing the 654 environmental variables, reducing them to at most 10 PCs, and applying K-means with random_state = 2 and 20 initializations. Environment-only and simple mean/history baselines were evaluated with the same held-out groups.

The validation groups use different targets and prediction units and are therefore interpreted separately. The primary benchmark uses an environment-equal hybrid-level pollen-DAP mean; GxE and grouped environmental analyses use pollen-DAP records; the phenotype sensitivity analysis uses adjusted hybrid BLUEs; and the 2024 temporal stress test uses yield. For the learned 2024 models, 91,135 EC-supported 2014-2023 hybrid-environment yield records were used for training and 9,064 observations from 21 2024 environments were available for testing; the strict unseen-hybrid subset contained 7,963 records from 959 hybrids. Table 3 reports the corresponding sample-size ranges and grouping structure.

3.5 Metrics, uncertainty, and statistical testing

RMSE and MAE were used as error-based metrics [39]. R2 was calculated independently within each test split as 1 - SSE/SST using that split's observed target mean; Pearson correlation was calculated separately and was not squared to obtain R2. RMSE, MAE, R2, and Pearson were first computed for each outer test split and then averaged in the same split order. Two-sided t-based 95% confidence intervals across the ten split-level values are retained as descriptive uncertainty summaries and are labeled explicitly as split-level t intervals.

Because the ten repeated holdouts are dependent, primary paired inference was revised to a corrected resampled procedure. For each prespecified pair, split-level RMSE differences d were formed on the ten common splits and the corrected standard error was computed using Eq. (9).

$$SE_c = \sqrt{(\frac{1}{k} + \frac{n_{test}}{n_{train}})s_d^2} \quad (9)$$

Here, k = 10 and the test-to-training ratio is 0.25; tests used 9 degrees of freedom. Five primary comparisons against RF-PCA100 were adjusted by Holm correction and accompanied by Cohen dz effect sizes. The previously archived 2,000-repetition hierarchical paired bootstrap is reported separately as a bootstrap interval and is not mixed with the descriptive t-based split interval. Undefined Pearson values caused by constant predictions are reported as undefined rather than replaced with zero.

3.6 Reproducibility and integrity safeguards

All 1,945 planned runs were completed and matched the predefined experiment plan. The archive contains no failed, missing, or duplicate run records, and all RMSE, MAE, and R2 values are finite. Eighty leave-one-environment-out environment-only HGB runs produced undefined Pearson correlations because predictions were constant or near-constant; these cases were explicitly recorded. Across the ten parent-disjoint splits, no parental overlap was observed. All 2,156 selected SNP markers passed REF/ALT parsing and PVAR consistency checks for DPC8 construction. The targeted revision archive additionally stores adjusted phenotype values, corrected resampled tests, grouped-transfer metrics, stress-test baselines, and the v5.1 sensitivity outputs.

Fig. 1 provides an overview of the workflow, from genotype–phenotype alignment and SNP representation to model training, statistical testing, and stress-test validation.

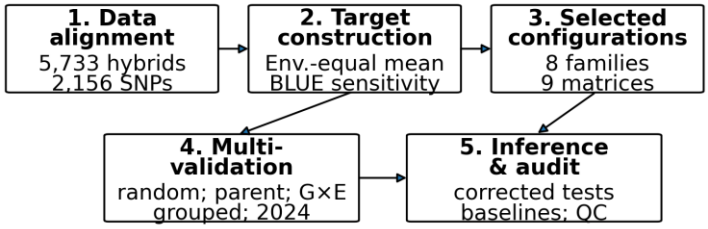


Figure. 1 Revised experimental workflow emphasizing target construction, selected configurations, multi-validation, baselines, and corrected inference

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

4. Results and discussion

4.1 Confirmatory repeated-holdout performance

Fig. 2 summarizes the primary repeated-holdout comparison for pollen DAP. RF-PCA100 achieved the lowest mean RMSE of 2.3982 (SD = 0.0822; descriptive 95% t interval 2.3394-2.4570), with mean MAE 1.5727, R2 0.6411, and Pearson 0.8019. RF-additive and RF-MAF-weighted had RMSE values of 2.3992 and 2.3993 and slightly higher Pearson correlations of 0.8021 and 0.8021, respectively. Thus, RF-PCA100 is the lowest-RMSE configuration rather than uniformly best across metrics. In the corrected resampled analysis, the RMSE differences relative to RF-additive, RF-MAF-weighted, and RF-recessive were 0.0010, 0.0011, and 0.0176 days, with Holm-adjusted p = 1.000 for all three. The leading RF configurations are therefore treated as statistically comparable.

4.2 Accuracy-dimensionality and runtime trade-off

The leading RF configurations were nearly identical in predictive accuracy but differed markedly in dimensionality and CPU runtime. In the primary benchmark, PCA100 reduced 2,156 SNP markers to 100 components (95.36% reduction) and required 26.46 s per run. RF-additive, RF-MAF-weighted, RF-recessive, and RF-one-hot required 78.48 s, 107.33 s, 57.39 s, and 180.03 s, respectively. One-hot expanded the input to 6,468 features without improving RMSE, while the DPC8 matrix expanded it to 17,248 values. Runtime is reported only as an environment-specific CPU measurement.

The corrected paired results support an accuracy-efficiency interpretation rather than a numerical winner. Fig. 3 shows that RF-PCA100 occupies the compact end of the accuracy-dimensionality trade-off, while expanded categorical and nucleotide-aware inputs do not improve random-holdout pollen-DAP performance in the evaluated configurations.

4.3 Biological contrast encodings and allele frequency

Within the RF model, where the learning algorithm and holdout splits were held constant, recessive encoding (RMSE 2.4158) was more accurate than dominant (2.4603) and heterozygous (2.4822) but did not outperform additive, MAF-weighted, or PCA100. RF-MAF-weighted and RF-additive were nearly identical. Increasing the MAF threshold from 0.01 to 0.03 or 0.05 removed only one to three markers and had little effect. These are within-model representation comparisons and do not establish an independent representation main effect.

4.4 Deep-learning depth and recent architectures

MLP-Medium with additive encoding produced the best plain-MLP result (RMSE = 2.4360), whereas deeper or residual configurations did not improve the primary target. These outcomes are configuration-level observations rather than evidence that one model family is universally preferable [40].

Earlier benchmarks likewise show that deep learning does not consistently outperform classical or machine-learning methods [41, 42]. In the present study, Cropformer-inspired configurations and DPCformer-inspired 1D/8D configurations were controlled architectural adaptations rather than faithful reproductions of all original preprocessing and feature-selection stages [14, 17].

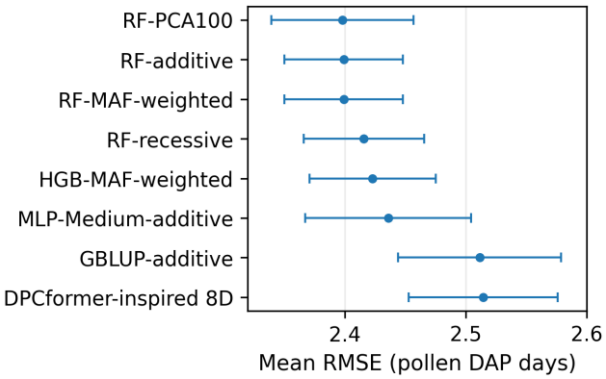


Figure. 2 Mean pollen-DAP RMSE for representative selected configurations across ten common holdout splits. Error bars are descriptive split-level t-based 95% confidence intervals

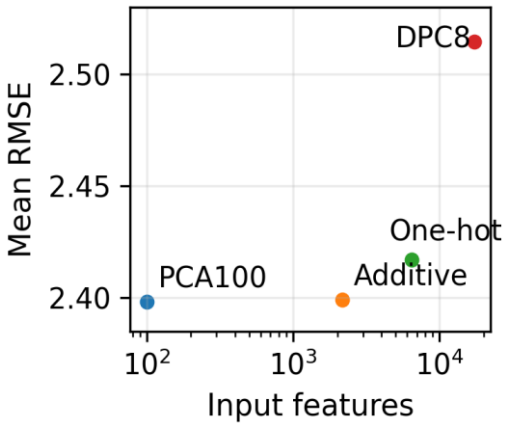


Figure. 3 Accuracy-dimensionality trade-off for representative configurations. PCA100 compresses 2,156 SNPs to 100 components, whereas one-hot and DPC8 expand the input

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

The DPCformer-inspired 8D input expanded 2,156 loci to 17,248 values without improving random-holdout pollen-DAP accuracy. These results apply only to the adaptations evaluated here and must not be interpreted as rankings of the original Cropformer or DPCformer methods.

4.5 Trait-dependent configuration ranking

No single evaluated configuration performed best for every trait. RF-PCA100 led for pollen and silk DAP; HGB-additive or HGB-MAF-weighted led for height, moisture, and yield; and GBLUP-additive led for lodging and test weight. These are rankings within the configuration panels evaluated for each trait, not independent estimates of model or representation effects.

4.6 Generalization to unseen parental backgrounds

Strict parent-disjoint validation produced a substantial decline relative to random holdout validation. DPCformer-inspired 8D + DPC8 had the lowest mean RMSE (3.3865; 95% CI 3.0971-3.6759), followed by MLP-Medium + PCA300 (3.4412) and RF + PCA100 (3.4848). Every evaluated configuration had a negative mean R2; the leading DPC8 configuration had mean R2 = -0.0124 (95% CI -0.1843 to 0.1595). A training-mean baseline also had R2 near zero across the ten strict splits, and the partitions contained 461-559 test hybrids, 941-998 test parents, and zero shared parents. Strong random-holdout accuracy therefore did not establish reliable transfer to new parental backgrounds.

4.7 Environmental covariates, transfer, and 2024 temporal validation

For within-distribution pollen-DAP prediction, HGB genomic-plus-environmental (G+E) achieved RMSE 2.1404 and R2 0.9198, compared with RMSE 7.1285 and R2 0.1113 for genomic-only HGB. Environment-only HGB also achieved R2 0.8845, indicating that much of the within-distribution signal was environmental. In the original eight-environment leave-one-environment-out analysis, mean R2 values became strongly negative, including -5.6073 for HGB G+E and -5.0868 for HGB environment-only. This contrast motivated the broader grouped transfer analysis in Section 4.8.

The 2024 yield stress test showed the same deployment gap. On the EC-aligned 9,064 test records, MLP G+E was the best learned configuration with RMSE 3.0253 and R2 = -0.0244; on 7,963 records from 959 unseen hybrids, RMSE was 3.0559 and R2 = -0.0451. Using the same 91,135 EC-supported training hybrid-environment records and the same test records, a simple historical-location mean baseline achieved RMSE 2.8738 and R2 = 0.0761 for all 2024 records and RMSE 2.8831 and R2 = 0.0700 for unseen hybrids. Thus, the simple location baseline outperformed the selected learned configurations in this temporal setting. Reaction-norm research supports the value of environmental covariates [24], but reliability remains dependent on trial design and target-population overlap [25, 26].

4.8 Grouped distribution-shift validation

Targeted grouped validation tested whether the high within-distribution G+E result transferred when broader environmental structure was excluded. HGB G+E achieved positive R2 in 6 of 10 held-out years, 17 of 41 held-out locations, 64 of 217 held-out location-years, and 2 of 5 EC-derived climatic regimes. Mean R2 was 0.0579 for leave-one-year-out but fell to -1.3511 for leave-one-location-out, -5.4219 for leave-one-location-year-out, and -0.2069 for leave-one-climatic-regime-out. The corresponding environment-only means were -0.1107, -1.6286, -5.8276, and -0.4690. Genomic information therefore improved the HGB mean relative to environment-only in each grouped analysis, but transfer remained inconsistent and often failed to beat a mean-prediction baseline.

Fig. 4 reports the fraction of held-out groups with positive R2, while Table 5 summarizes group counts and mean R2 intervals; sample-size ranges are reported in Table 3. The strong Pearson correlation observed in some location-year holdouts despite highly negative R2 further indicates that ranking signal can coexist with poor calibration of absolute values.

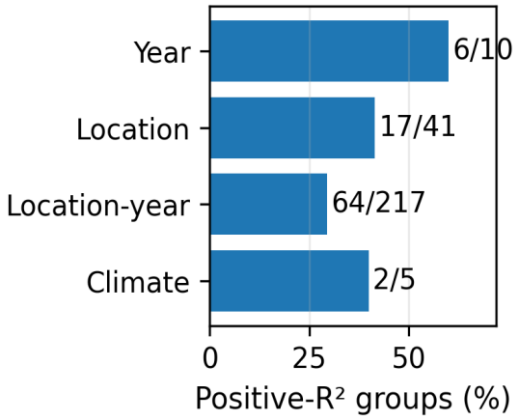


Figure. 4 Proportion of held-out groups with positive R2 for the HGB genomic-plus-environmental configuration under grouped distribution shift

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

4.9 Phenotype sensitivity and corrected statistical inference

Phenotype construction materially changed the configuration ranking. The environment-equal primary target had SD 3.999, whereas the adjusted BLUE target had SD 2.698, so absolute RMSE values across these two targets are not directly comparable. For the original target, the canonical primary-benchmark estimates are reported for configurations already evaluated under the same 4,767 hybrids and ten common 80/20 splits. RF-PCA100 and RF-additive therefore have RMSE values of 2.3982 and 2.3992, respectively. On adjusted BLUEs, GBLUP-additive ranked first (RMSE 1.0282, R2 0.8516, Pearson 0.9233), followed by HGB-additive (RMSE 1.0499, R2 0.8454). RF-additive and RF-PCA100 were lower on the adjusted target. The environment-equal target and BLUEs correlated only r = 0.4455, demonstrating that phenotype adjustment changes the prediction target and can change the leading configuration.

The corrected resampled tests also changed the strength of statistical claims. Relative to RF-PCA100, Holm-adjusted p-values were 1.000 for RF-additive, RF-MAF-weighted, and RF-recessive. GBLUP-additive and MLP-Medium-PCA300 had unadjusted corrected p-values of 0.0393 and 0.0452, but both became 0.1964 after Holm adjustment. Accordingly, the paper emphasizes effect sizes, corrected inference, and groups of practically comparable configurations rather than the numerically best rank alone.

4.10 Limitations and implications

This benchmark uses one maize resource and a predefined, non-factorial configuration panel; its rankings should not be generalized to all breeding populations or interpreted as independent effects of representation and model choice. Cropformer-inspired and DPCformer-inspired systems are controlled adaptations rather than exact reproductions of every original feature-selection, preprocessing, optimization, and interpretation step [14, 17]. The primary target is an environment-equal hybrid mean, and the BLUE sensitivity analysis shows that phenotype construction materially changes both the target distribution and configuration ranking.

The strong within-distribution results did not transfer reliably to unseen parents, sites, location-years, climatic regimes, or future-year yield. Grouped environmental analyses also show that environmental covariates can support interpolation while failing under broader exclusion. The 2024 stress test is therefore treated as a deployment-oriented failure analysis rather than as a model-selection benchmark, and no configuration is described as breeding-ready. Simple baselines, including historical location means, must be retained when evaluating future deployment.

Table 5. Grouped environmental transfer summary; sample-size ranges are reported in Table 3

Group	HGB G+E mean R2 [95% CI]	Pos. R2	E-only mean R2
Year	0.0579 [-0.2053, 0.3211]	6/10	-0.1107
Location	-1.3511 [-2.5514, -0.1507]	17/41	-1.6286
Location-year	-5.4219 [-8.0304, -2.8135]	64/217	-5.8276
Climate regime	-0.2069 [-0.6946, 0.2809]	2/5	-0.4690

Table 6. Phenotype-construction sensitivity on ten common holdout splits. Original-target entries use the canonical primary-benchmark estimates for the same configurations and splits

Target / config.	RMSE [95% descriptive t CI]	R2	Pearson r
Original / RF-PCA100	2.3982 [2.3394, 2.4570]	0.6411	0.8019
Original / RF-additive	2.3992 [2.3502, 2.4482]	0.6409	0.8021
Original / HGB-additive	2.4231 [2.3709, 2.4753]	0.6337	0.7965
Original / GBLUP-additive	2.5114 [2.4441, 2.5788]	0.6063	0.7799
BLUE / GBLUP-additive	1.0282 [1.0068, 1.0496]	0.8516	0.9233
BLUE / HGB-additive	1.0499 [1.0235, 1.0762]	0.8454	0.9201
BLUE / RF-additive	1.1632 [1.1321, 1.1943]	0.8103	0.9030
BLUE / RF-PCA100	1.1879 [1.1518, 1.2240]	0.8021	0.8974

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Future work should evaluate independent model and representation effects using a complete factorial design, use mixed-model phenotype adjustment and formal reliability estimates, investigate calibration at phenotype extremes, and develop domain-adaptation and uncertainty-aware methods across families, environments, and years. Training-population and multi-environment trial design should also be optimized because relationship structure and environmental coverage strongly affect transfer [25, 47].

5. Conclusions

This study compared selected SNP representation-model configurations under random, trait-specific, parent-disjoint, environmental, phenotype-sensitivity, and temporal validation. Because the experiment was not a complete factorial design, it does not establish an independently superior SNP representation or learning algorithm. For the primary pollen-DAP target, RF-PCA100 had the lowest mean RMSE, while RF-additive and RF-MAF-weighted produced nearly identical accuracy and slightly higher Pearson correlation. Corrected resampled tests with Holm adjustment found no significant RMSE difference among the leading RF configurations. PCA100 therefore offers a favorable accuracy-dimensionality trade-off rather than a universal accuracy advantage.

Generalization failure was a primary result. All evaluated parent-disjoint configurations had negative mean R2. HGB G+E achieved positive R2 in only 6/10 held-out years, 17/41 locations, 64/217 location-years, and 2/5 climatic regimes. In the aligned 2024 yield test, the historical-location mean baseline outperformed the best learned configuration for both all records and unseen hybrids. These findings show that strong within-distribution accuracy is insufficient evidence for deployment to new families, sites, or years.

Across nine traits, leading evaluated configurations varied among RF, HGB, and GBLUP. Phenotype adjustment also changed the ranking: GBLUP-additive led the adjusted-BLUE sensitivity panel, whereas RF configurations led the original environment-equal target. The overall evidence therefore supports configuration-level benchmarking, explicit phenotype-definition sensitivity, simple deployment baselines, and demanding distribution-shift validation before genomic prediction outputs are used to guide breeding decisions.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Rizki Darmawan: conceptualization, methodology, software, validation, formal analysis, data curation, visualization, and writing—original draft preparation. Heru Agus Santoso: supervision, validation, and writing—review and editing. Abdul Syukur: methodology review, supervision, and writing—review and editing. Affandy: supervision, validation, and project administration. All authors have read and agreed to the submitted version of the manuscript.

Acknowledgments

The authors thank Dian Nuswantoro University and the Genomes to Fields community for academic and data support.

Data and Code Availability

The raw genotype, phenotype, and environmental data are publicly available from the Genomes to Fields (G2F) Initiative [34]. The supporting archive contains the versioned scripts reviewer_all_requests_experiment_g2f_final_v4_1.py and ijies_targeted_revision_experiments_v5_1.py preprocessing logic, split identifiers, run configurations, integrity reports, sample-level predictions, per-run metrics, corrected resampled tests, grouped-transfer outputs, phenotype-sensitivity results, and baseline tables. These materials are publicly accessible through: https://drive.google.com/drive/folders/1pDBk6sPc7sgUFs096ETa-kgFIj0JvGOA?usp=sharing.

The archived files are intended to permit reconstruction of the reported preprocessing, training, validation, statistical comparisons, and targeted revision analyses. The manuscript reports the software versions and principal settings needed to interpret those files consistently.

References

[1] D. K. Ray, N. D. Mueller, P. C. West, and J. A. Foley, “Yield trends are insufficient to double global crop production by 2050”, *PLoS One*, Vol. 8, No. 6, p. e66428, 2013, doi: 10.1371/journal.pone.0066428.

[2] T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard, “Prediction of total genetic value using genome-wide dense marker maps”, *Genetics*, Vol. 157, No. 4, pp. 1819–1829, 2001, doi: 10.1093/genetics/157.4.1819.

[3] J. Crossa, *et al.*, “Genomic selection in plant breeding: methods, models, and perspectives”, *Trends Plant Sci.*, Vol. 22, No. 11, pp. 961–975, 2017, doi: 10.1016/j.tplants.2017.08.011.

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

[4] T. He and C. Li, "Harness the power of genomic selection and the potential of germplasm in crop breeding for global food security in the era with rapid climate change", *Crop J.*, Vol. 8, No. 5, pp. 688–700, 2020, doi: 10.1016/j.cj.2020.04.005.

[5] R. Bernardo and J. Yu, "Prospects for genomewide selection for quantitative traits in maize", *Crop Sci.*, Vol. 47, No. 3, pp. 1082–1090, 2007, doi: 10.2135/cropsci2006.11.0690.

[6] D. C. Lima, *et al.*, "Genomes to Fields 2022 maize genotype by environment prediction competition", *BMC Res. Notes*, Vol. 16, No. 1, p. 148, 2023, doi: 10.1186/s13104-023-06421-z.

[7] M. Lopez-Cruz, *et al.*, "Leveraging data from the Genomes-to-Fields initiative to investigate genotype-by-environment interactions in maize in North America", *Nat. Commun.*, Vol. 14, No. 1, p. 6904, 2023, doi: 10.1038/s41467-023-42687-4.

[8] M. Pérez-Enciso and L. M. Zingaretti, "A guide on deep learning for complex trait genomic prediction", *Genes*, Vol. 10, No. 7, p. 553, 2019, doi: 10.3390/genes10070553.

[9] O. A. Montesinos-López, *et al.*, "A review of deep learning applications for genomic selection", *BMC Genomics*, Vol. 22, No. 1, p. 19, 2021, doi: 10.1186/s12864-020-07319-x.

[10] R. Abdollahi-Arpanahi, D. Gianola, and F. Peñagaricano, "Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes", *Genet. Sel. Evol.*, Vol. 52, No. 1, p. 12, 2020, doi: 10.1186/s12711-020-00531-z.

[11] Z. Wang, W. Li, and Z. Tang, "Enhancing the genomic prediction accuracy of swine agricultural economic traits using an expanded one-hot encoding in CNN models", *J. Integr. Agric.*, Vol. 24, No. 9, pp. 3574–3582, 2025, doi: 10.1016/j.jia.2024.03.071.

[12] L. Krusenbaum and M. Wissuwa, "Low-call-rate SNPs and presence-absence variation identified in the rice pan-genome can improve genomic prediction of rice gene bank accessions", *Theor. Appl. Genet.*, Vol. 138, No. 12, p. 295, 2025, doi: 10.1007/s00122-025-05080-x.

[13] V. Manthena, *et al.*, "Evaluating dimensionality reduction for genomic prediction", *Front. Genet.*, Vol. 13, p. 958780, 2022, doi: 10.3389/fgene.2022.958780.

[14] H. Wang, *et al.*, "Cropformer: an interpretable deep learning framework for crop genomic prediction", *Plant Commun.*, Vol. 6, No. 3, p. 101223, 2025, doi: 10.1016/j.xplc.2024.101223.

[15] H. Wu, *et al.*, "AutoGP: an intelligent breeding platform for enhancing maize genomic selection", *Plant Commun.*, Vol. 6, No. 4, p. 101240, 2025, doi: 10.1016/j.xplc.2025.101240.

[16] C. Wang, D. Zhang, Y. Ma, Y. Zhao, P. Liu, and X. Li, "WheatGP, a genomic prediction method based on CNN and LSTM", *Brief. Bioinform.*, Vol. 26, No. 2, p. bbaf191, 2025, doi: 10.1093/bib/bbaf191.

[17] P. Deng, *et al.*, "DPCformer: an interpretable deep learning model for genomic prediction in crops", *arXiv Preprint*, arXiv:2510.08662, 2025, doi: 10.48550/arXiv.2510.08662.

[18] J. Li, *et al.*, "Leveraging weighted embedding and Transformer architecture to improve phenotype prediction of complex traits for crops", *Nat. Commun.*, Vol. 17, No. 1, p. 4427, 2026, doi: 10.1038/s41467-026-71035-5.

[19] M. Gill, *et al.*, "Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction", *BMC Plant Biol.*, Vol. 22, No. 1, p. 180, 2022, doi: 10.1186/s12870-022-03559-z.

[20] F. Mora-Poblete, C. Maldonado, L. Henrique, R. Uhdre, C. A. Scapim, and C. A. Mangolim, "Multi-trait and multi-environment genomic prediction for flowering traits in maize: a deep learning approach", *Front. Plant Sci.*, Vol. 14, p. 1153040, 2023, doi: 10.3389/fpls.2023.1153040.

[21] T. Pook, J. Freudenthal, A. Korte, and H. Simianer, "Using local convolutional neural networks for genomic prediction", *Front. Genet.*, Vol. 11, p. 561497, 2020, doi: 10.3389/fgene.2020.561497.

[22] I.-T. Vourlaki, S. E. Ramos-Onsins, M. Pérez-Enciso, and R. Castanera, "Evaluation of deep learning for predicting rice traits using structural and single-nucleotide genomic variants", *Plant Methods*, Vol. 20, No. 1, p. 121, 2024, doi: 10.1186/s13007-024-01250-y.

[23] A. Alemu, *et al.*, "Genomic selection in plant breeding: key factors shaping two decades of progress", *Mol. Plant*, Vol. 17, No. 4, pp. 552–578, 2024, doi: 10.1016/j.molp.2024.03.007.

[24] D. Jarquín, *et al.*, "A reaction norm model for genomic selection using high-dimensional genomic and environmental data", *Theor. Appl. Genet.*, Vol. 127, No. 3, pp. 595–607, 2014, doi: 10.1007/s00122-013-2243-1.

[25] S. Rio, D. Akdemir, T. Carvalho, and J. I. Y. Sánchez, "Assessment of genomic prediction reliability and optimization of experimental designs in multi-environment trials", *Theor.*

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

Appl. Genet., Vol. 135, No. 2, pp. 405–419, 2022, doi: 10.1007/s00122-021-03972-2.

[26] C. A. V. Barreto, *et al.*, "Genomic prediction in multi-environment trials in maize using statistical and machine learning methods", *Sci. Rep.*, Vol. 14, No. 1, p. 1062, 2024, doi: 10.1038/s41598-024-51792-3.

[27] G. Costa-Neto, R. Fritsche-Neto, and J. Crossa, "Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials", *Heredity*, Vol. 126, No. 1, pp. 92–106, 2021, doi: 10.1038/s41437-020-00353-1.

[28] M. Sinecen, "Comparison of genomic best linear unbiased prediction and Bayesian regularization neural networks for genomic selection", *IEEE Access*, Vol. 7, pp. 79199–79210, 2019, doi: 10.1109/access.2019.2922006.

[29] A. Montesinos-López, O. A. Montesinos-López, J. C. Montesinos-López, C. A. Flores-Cortes, R. de la Rosa, and J. Crossa, "A guide for kernel generalized regression methods for genomic-enabled prediction", *Heredity*, Vol. 126, No. 4, pp. 577–596, 2021, doi: 10.1038/s41437-021-00412-1.

[30] L. Ornella, *et al.*, "Genomic-enabled prediction with classification algorithms", *Heredity*, Vol. 112, No. 6, pp. 616–626, 2014, doi: 10.1038/hdy.2013.144.

[31] J. M. González-Camacho, *et al.*, "Genome-enabled prediction of genetic values using radial basis function neural networks", *Theor. Appl. Genet.*, Vol. 125, No. 4, pp. 759–771, 2012, doi: 10.1007/s00122-012-1868-9.

[32] J. Guo, *et al.*, "Multi-trait genomic prediction of yield-related traits in US soft wheat under variable water regimes", *Genes*, Vol. 11, No. 11, p. 1270, 2020, doi: 10.3390/genes11111270.

[33] C. B. Azodi, J. Pardo, R. VanBuren, G. de los Campos, and S.-H. Shiu, "Transcriptome-based prediction of complex traits in maize", *Plant Cell*, Vol. 32, No. 1, pp. 139–151, 2020, doi: 10.1105/tpc.19.00332.

[34] The Genomes to Fields Initiative, "The Genomes to Fields Initiative", [Online]. Available: https://www.genomes2fields.org/home/, accessed Jul. 19, 2026.

[35] G. Sahana, Z. Cai, M. P. Sanchez, A. C. Bouwman, and D. Boichard, "Invited review: good practices in genome-wide association studies to identify candidate sequence variants in dairy cattle", *J. Dairy Sci.*, Vol. 106, No. 8, pp. 5218–5241, 2023, doi: 10.3168/jds.2022-22694.

[36] F. Mittag, M. Römer, and A. Zell, "Influence of feature encoding and choice of classifier on disease risk prediction in genome-wide association studies", *PLoS One*, Vol. 10, No. 8, p. e0135832, 2015, doi: 10.1371/journal.pone.0135832.

[37] Y. Da, C. Wang, S. Wang, and G. Hu, "Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers", *PLoS One*, Vol. 9, No. 1, p. e87666, 2014, doi: 10.1371/journal.pone.0087666.

[38] P. M. VanRaden, "Efficient methods to compute genomic predictions", *J. Dairy Sci.*, Vol. 91, No. 11, pp. 4414–4423, 2008, doi: 10.3168/jds.2007-0980.

[39] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? - arguments against avoiding RMSE in the literature", *Geosci. Model Dev.*, Vol. 7, No. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.

[40] K. Wang, M. A. Abid, A. Rasheed, J. Crossa, S. Hearne, and H. Li, "DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants", *Mol. Plant*, Vol. 16, No. 1, pp. 279–293, 2023, doi: 10.1016/j.molp.2022.11.004.

[41] C. B. Azodi, E. Bolger, A. McCarren, M. Roantree, G. de los Campos, and S.-H. Shiu, "Benchmarking parametric and machine learning models for genomic prediction of complex traits", *G3 Genes Genomes Genet.*, Vol. 9, No. 11, pp. 3691–3702, 2019, doi: 10.1534/g3.119.400498.

[42] P. Bellot, G. de los Campos, and M. Pérez-Enciso, "Can deep learning improve genomic prediction of complex human traits?", *Genetics*, Vol. 210, No. 3, pp. 809–819, 2018, doi: 10.1534/genetics.118.301298.

[43] P. Waldmann, "Approximate Bayesian neural networks in genomic prediction", *Genet. Sel. Evol.*, Vol. 50, No. 1, p. 70, 2018, doi: 10.1186/s12711-018-0439-1.

[44] O. A. Montesinos-López, A. Montesinos-López, J. Crossa, D. Gianola, C. M. Hernández-Suárez, and J. Martín-Vallejo, "Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits", *G3 Genes Genomes Genet.*, Vol. 8, No. 12, pp. 3829–3840, 2018, doi: 10.1534/g3.118.200728.

[45] A. Montesinos-López, *et al.*, "Multimodal deep learning methods enhance genomic prediction

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/

of wheat breeding", *G3: Genes*, Genomes, Genetics, Vol. 13, No. 5, p. jkad045, 2023, doi: 10.1093/g3journal/jkad045.

[46] A. Montesinos-López, *et al.*, "Deep learning methods improve genomic prediction of wheat breeding", *Front. Plant Sci.*, Vol. 15, p. 1324090, 2024, doi: 10.3389/fpls.2024.1324090.

[47] G. de los Campos, J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, "Whole-genome regression and prediction methods applied to plant and animal breeding", *Genetics*, Vol. 193, No. 2, pp. 327–345, 2013, doi: 10.1534/genetics.112.143313.

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
License details: https://creativecommons.org/licenses/by-sa/4.0/