

Beyond Predictive Ability: Multidimensional Evaluation of Single and Multi-Trait Genomic Selection Models in Oat for Enhanced Breeding Efficiency

Muhammad Massub Tehseen

North Dakota State University

Charlotte Brault

University of Minnesota

Naa Korkoi Ardayfio

North Dakota State University

Sepehr Mohajeri Naraghi

North Dakota State University

Michael S McMullen

North Dakota State University

Jason D. Fiedler

Jason.fiedler@usda.gov

Edward T. Schafer Agricultural Research Center

Research Article

Keywords:

Posted Date: June 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8712671/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Genomic selection (GS) can accelerate genetic gain through early marker-based selection, yet model evaluation in many studies remains centered on predictive ability (PA), which may inadequately reflect operational breeding value. Here, we evaluated single-trait (STM) and multi-trait (MTM) genomic prediction in a public oat breeding program using a multidimensional framework integrating predictive performance, selection efficiency, stability, and feasibility. A total of 1,941 oat breeding lines were phenotyped for five key traits (plant height, test weight, grain yield, plump kernels, and thin kernels) over five years (2018–2022) and genotyped using 5,126 genome-wide SNPs. Spatially adjusted phenotypes were used to derive BLUPs for genomic prediction with GBLUP (STM) and Bayesian MTM models. Model performance was assessed across four validation strategies: forward prediction, backward temporal prediction, historical-to-future prediction, and within-year cross-validation (RandomCV and CDmean). PA was highly trait-dependent, with consistently higher performance for plant height and thin kernels than for yield and plump kernels. In cross-validation, CDmean provided marginal PA improvements but required substantially greater computation and showed significantly lower rank correlation with phenotypic rankings, establishing RandomCV as superior for operational implementation. STM and MTM models showed comparable, trait-dependent performance. Importantly, multi-trait selection frameworks achieved hit rates exceeding 50%, demonstrating that moderate PA can deliver high breeding value when aligned with operational selection criteria. These results demonstrate that GS evaluation should complement PA with selection-aligned metrics, and we provide an implementation-focused framework to guide routine GS deployment in oat and similarly structured breeding programs.

Key message

Moderate predictive ability models achieve high breeding value when aligned with multi-trait selection criteria, demonstrating that genomic selection evaluation should complement prediction accuracy with selection outcome metrics.

Introduction

Genomic selection (GS) has emerged as a transformative tool in modern plant breeding, offering the potential to accelerate genetic gain through early and accurate prediction of breeding values using genome-wide molecular markers (Meuwissen et al., 2001; Heffner et al., 2009; Alemu et al., 2024). Unlike marker-assisted selection, which depends on a limited number of significant quantitative trait loci (QTL), GS leverages the effects of thousands of markers, enabling whole-genome modeling for complex traits governed by polygenic genetic architecture. Over the past two decades, GS has demonstrated encouraging results in crops like wheat, maize, barley, oat, and rice, supporting its promise to increase selection intensity, shorten breeding cycles, and reduce phenotyping costs (Crossa et al., 2017; Bernardo, 2020; Desta and Ortiz, 2014; Voss-Fels et al., 2019; Mellers et al., 2020). However, despite substantial theoretical advancements, the path from statistical prediction to actionable selection differential in field breeding remains underdeveloped.

Many GS studies continue to prioritize predictive ability (PA), typically measured as the correlation between genomic estimated breeding values (GEBVs) and best linear unbiased predictions (BLUPs), as the principal metric for evaluating model performance (Heslot et al., 2012; Zhao et al., 2012). While PA provides an important baseline, it is increasingly recognized as a limited proxy for breeding success (Michel et al., 2019). The predictive ability is only one parameter in the breeder's equation (Lush et al., 1937), defined as $\Delta G = h \cdot i \cdot \sigma_a / L$, with i the intensity of selection, h the selection precision (predictive ability in genomic selection), σ_a the additive genetic variance and L the generation interval. Models with high PA do not necessarily lead to superior field performance when other components vary significantly across operational contexts.

Breeding decisions are rarely made on statistical grounds alone; they depend on selection intensity, resource constraints, breeder intuition, and trait-specific priorities (Michel et al., 2019; Heffner et al., 2009). Indeed, empirical reports have shown that genomic models with modest PA values can still achieve high selection overlap and realized gain in real breeding programs (Zhang et al., 2017; Howard et al., 2014).

To address these shortcomings, researchers have proposed augmenting PA with additional evaluation metrics such as selection differential, hit rate, and selection overlap (Akdemir & Isidro-Sánchez, 2019). Hit rate measures the proportion of breeder-selected genotypes correctly identified by the model (analogous to recall in classification), while selection overlap measures the proportion of model-selected genotypes that align with breeder selections (analogous to precision). Both offer practical lenses through which to interpret model utility (Akdemir & Isidro-Sánchez, 2019). Similarly, realized selection differential, when calculated by comparing the performance of model-selected genotypes against both the population mean and the performance of actual breeder selections, can better reflect the direct impact of a model on breeding progress (Desta & Ortiz, 2014; Sallam et al., 2015). However, few studies have systematically integrated these complementary metrics into a single, breeder-relevant evaluation pipeline (Dekkers et al., 2021). Even fewer have done so using real multi-year, multi-environment data under operational constraints typical of public-sector breeding programs, including limited genotyping budgets, smaller breeding populations, irregular phenotyping schedules due to weather/funding variability, fewer testing environments compared to commercial programs, and the need to balance selection differential with cultivar stability across diverse production regions. This gap has limited the broader implementation of GS, especially in resource-constrained settings where genotyping is still expensive and where poor prediction decisions can waste limited field trial resources and delay cultivar release.

Furthermore, the diversity of modeling options, including single-trait (STM) vs. multi-trait (MTM), training population strategies, and cross-validation designs, further complicates genomic prediction deployment and affects predictive ability. Operational realities require breeders to make decisions not only about which model to use, but also how to design training sets, balance traits, and deploy predictions in multi-environment contexts (Montesinos-López et al., 2018; Fernandes et al., 2018; Cooper et al., 2014).

In this study, we present a comprehensive evaluation of genomic prediction strategies grounded in the operational context of a public oat breeding program. We propose a four-dimensional evaluation framework designed to reflect the decision space breeders routinely navigate. Specifically, we assess: (1) trait-wise model performance across validation designs, (2) selection efficiency via hit rate and selection differential, (3) temporal and environmental model robustness, and (4) MTM configuration performance.

Our analysis spans five agronomically important traits plant height, test weight, grain yield, plump kernels, and thin kernels captured from five years of field environments across multiple locations. The dataset reflects the complex structure of real breeding programs, including unbalanced environments, substantial genotype-by-environment interactions that cause differential trait expression across locations and years, and selection bottlenecks. We compare the performance of STM and MTM models under both with and without training set optimization (with RandomCV and CDmean-based validation). Metrics such as predictive ability, selection differential, hit rate, and computational time were assessed to provide a multidimensional view of model performance.

Our goal is to support breeders with an evaluation framework that not only benchmarks models statistically but also reflects real-world decision-making constraints. This study offers a comprehensive evaluation framework that enables breeders to make informed trait-specific decisions about model selection, resource allocation, and performance expectations.

Materials and Methods

Plant Materials and Experimental Design

A total of 1,941 unique NDSU oat breeding lines were evaluated in North Dakota from 2018 through 2022 across preliminary yield trials (PYTs) and advanced yield trials (AYTs). Several check genotypes were included repeatedly across years and trial types to enable comparisons. The experimental design structure, including number of experiments, lines, locations, and replications per year and trial type, is summarized in Table 1.

Table 1
Experimental design summary for oat breeding trials across years (2018–2022), trial types, and locations in North Dakota

Year	Trial Type	No. Experiments	Unique Lines	Locations	Replications per Location
2018	PYT	3	293	Casselton, Fargo	2
2018	AYT	2	95	Carrington, Casselton, Edgeley, Fargo	3
2019	PYT	4	320	Casselton, Fargo	2
2019	AYT	2	96	Carrington, Casselton, Fargo, Minot, Verona	3
2020	PYT	4	325	Casselton, Fargo	2
2020	AYT	2	95	Carrington, Casselton, Fargo, Minot	3
2021	PYT	3	292	Casselton, Fargo	2
2021	AYT	2	125	Carrington, Casselton, Fargo, Minot	3
2022	PYT	5	530	Casselton, Fargo	2
2022	AYT	2	69	Carrington, Casselton, Fargo, Minot	3

The breeding program followed a standard two-stage selection process, with elite lines advancing from PYTs to AYT's based on performance. Advancement of breeding lines from PYTs to AYT's ranged from 23% to 38% across years, reflecting consistent selection intensity. Field evaluation sites were located near Fargo, Casselton, Carrington, Edgeley, Verona, and Minot in North Dakota, representing diverse agro-climatic conditions across the state.

Phenotypic Traits

Experiments were conducted under field conditions at multiple North Dakota experiment stations. Soil amendments were applied when necessary to maintain optimum plant growth according to standard local management practices. Phenotypic evaluation focused on five agronomic traits: plant height, grain yield, test weight, thin kernels, and plump kernels.

Plant height was measured at physiological maturity when plants had begun senescence and was recorded in centimeters from the ground level to the tip of the longest panicle using a measuring pole guided by hand. Grain yield was determined as the total harvested grain weight per plot, adjusted to 14% moisture content, and expressed on an area basis as kilograms per hectare (kg ha^{-1}). Test weight, which

represents grain density, was measured by weighing grain contained in a 500-mL chondrometer and reported in kilograms per hectolitre, with higher test weight indicating higher quality grain.

Plumps and thins are subsets of yield in which thins represent the smallest seed sizes, typically from the tertiary florets, while plumps are the larger seeds, typically from the primary and secondary florets, that meet a desirable milling threshold value. These subsets were determined after threshing by sifting grains with standard oat evaluation sieves. Grains passing through a 5/64-inch slotted sieve were considered thins and presented as a percentage of the total grain sample, while grains retained on a 6/64-inch slotted sieve were considered plumps and expressed on a weight basis in the same units as yield (kg ha⁻¹).

Phenotypic Data Processing

Phenotypic data were processed using a two-stage approach. In Stage 1, spatial analysis was conducted for each trait within each environment using SpATS in R (Rodríguez-Álvarez et al., 2018) to account for field variation through P-spline smoothing and discrete row-column effects. Spatially adjusted genotype means were extracted as empirical best linear unbiased predictors (E-BLUPs). In Stage 2, E-BLUPs from Stage 1 were combined across environments using weighted mixed models in lme4, with genotype-by-environment interactions modeled when data structure permitted. Broad-sense heritability was calculated as $h^2 = \sigma^2_g / (\sigma^2_g + \sigma^2_{ge}/n_e + \sigma^2_\epsilon/n_e)$, where variance components were estimated using REML. The resulting BLUPs served as response variables for genomic prediction analyses. Detailed model specifications are provided in **Supplementary Methods S1**.

Genotypic Data and G-Matrix Construction

Breeding lines (n = 1,941) were genotyped using a dual-enzyme genotyping-by-sequencing (GBS) approach as described by Ardayfio et al. (2025). Following quality control, imputation, and LD-based marker thinning, the final dataset consisted of 5,126 SNPs distributed across the oat genome. The additive genomic relationship matrix (G) was constructed using the VanRaden method (VanRaden, 2008) implemented in rrBLUP (Endelman, 2011). Principal component analysis was performed on the G-matrix to assess population structure (**Supplementary Fig. S1**), with the first five principal components retained as fixed effects in genomic prediction models to account for population stratification. Detailed genotyping and quality control procedures are provided in **Supplementary Methods S2**.

Genomic Prediction Models

Genomic prediction was implemented using both single-trait (STM) and multi-trait (MTM) models to evaluate the potential benefits of leveraging genetic correlations among traits. For single-trait models, each trait was analyzed independently using genomic best linear unbiased prediction (GBLUP)

implemented through the rrBLUP package (Endelman, 2011). For multi-trait models, the MTM package (de los Campos and Grüneberg, 2016) was used to jointly analyze multiple traits.

Three multi-trait model configurations were evaluated:

1. Breeder-defined combinations: Trait combinations based on breeding program priorities:

- • Yield + Height (key agronomic traits)
- • Yield + Height + Test weight (comprehensive agronomic performance)
- • Thins + Plumps (grain quality traits)

2. Correlation-based combinations: Trait pairs were selected using the genetic correlation matrix estimated from Stage-2 genotype BLUPs to represent contrasting levels and directions of genetic association. Specifically, we included one strong positive pair, one strong negative pair, and one weakly correlated comparator combination:

- • Thins + Plumps
- • Yield + Plumps
- • Height + Test weight

3. All-traits model: Joint analysis of all five traits simultaneously to maximize information utilization.

For single-trait GBLUP, the model was:

$$y = X\beta + Zu + \varepsilon$$

where y is the vector of BLUPs, X is the design matrix for fixed effects (intercept and first 5 PCs), β is the vector of fixed effects, Z is the identity matrix linking genotypes to observations, $u \sim N(0, G\sigma^2_u)$ is the vector of genomic estimated breeding values, and $\varepsilon \sim N(0, I\sigma^2_e)$ is the residual error.

For multi-trait models, the formulation was extended to include the same fixed effects as the STM (population structure PCs):

$$Y = XB + ZU + E$$

where Y is the $n \times t$ matrix of BLUPs for t traits, B is the matrix of fixed effects, $U \sim MN(0, G, \Sigma_u)$ follows a matrix-normal distribution with genetic covariance matrix Σ_u among traits, and $E \sim MN(0, I, \Sigma_e)$ represents residual errors with covariance matrix Σ_e .

Single-trait models were fitted with the genomic relationship matrix as the kernel. Variance components were estimated using restricted maximum likelihood (REML). Multi-trait models were implemented using Bayesian methods via Markov Chain Monte Carlo (MCMC) sampling in the MTM package. Posterior

means were used as point estimates for genomic estimated breeding values (GEBVs) and variance-covariance parameters.

Validation Strategy

A four-strategy validation framework addressed distinct practical questions in genomic selection. Strategy 1 (Forward prediction) evaluated the use of historical data (2018–2021) to predict new genotypes in subsequent years, mimicking operational breeding scenarios. Strategy 2 (Backward prediction) assessed training set size effects by varying the temporal window used to predict 2021 performance (e.g., 2020 alone vs. 2018–2020 combined). Strategy 3 (Historical prediction) tested the value of maintaining extensive datasets by comparing predictions using PYT-only versus combined PYT + AYT training populations. Strategy 4 (Within-year cross-validation) compared two validation methodologies, random 5-fold cross-validation (repeated 10 times) versus CDmean optimization (Rincant et al., 2012), across all years (2018–2022). To ensure temporal separation and avoid data leakage, trait-year-type-specific BLUPs were computed separately for each validation strategy using only the appropriate training years' data in Stage 2 BLUP calculation. Detailed validation scenarios and implementation are provided in **Supplementary Methods S3**.

Selection-Efficiency Analysis

Selection efficiency was evaluated to determine how well genomic prediction models captured breeder advancement decisions and to quantify the genetic gain obtainable from model-based selection. For each prediction scenario and trait, genotypes were ranked by direction-adjusted GEBVs, where a coefficient $d = +1$ (plumps, test weight, yield) or -1 (height, thins) ensured that larger values always reflected improvement. Three population-based selection intensities (10%, 20%, 30% of the preliminary-yield-trial population) were examined; results reported here focus on the breeding-realistic 30% threshold.

Model selections, defined as the top 30% of lines based on GEBVs, were compared with breeder-advanced lines recorded in the NDSU oat program archives. Advancement rates from PYTs to AYT_s varied across years, ranging from 22.9% to 38.0%. Overlap between model-based selections and breeder-advanced lines was evaluated using hit rate, calculated as the percentage of breeder-selected lines successfully captured by the genomic prediction model.

Selection differential was quantified using two complementary metrics: (1) percentage change relative to the population mean, calculated as $[(\text{mean of selected} - \text{population mean}) / \text{population mean}] \times 100$; and (2) standardized selection differential, calculated as $(\text{mean of selected} - \text{population mean}) / \text{phenotypic standard deviation}$. Both metrics were based on phenotypic BLUP values and express the expected selection differential from genomic selection in absolute percentage terms and in standard deviation units, respectively.

Advanced Analytical Framework

To provide a comprehensive evaluation of genomic selection performance, four complementary analyses were conducted. Analysis 1 summarized mean predictive ability, hit rate, and standardized selection differential across all validation strategies using multi-trait model outputs. Analysis 2 compared three multi-trait modeling strategies (All_Traits, Breeder_Selected, Correlation_Based) to identify optimal trait-combination approaches. Analysis 3 evaluated prediction stability using temporal stability (consistency across validation scenarios) and environmental stability (consistency across crop years). Analysis 4 assessed the association between predictive ability and realized selection differential to determine whether higher accuracy translates to superior parent selection. Detailed analytical specifications are provided in **Supplementary Methods S4**.

Results

Overview of Experimental Dataset

A comprehensive six-year, multi-environment dataset comprising 15,764 plot observations provided the foundation for all genomic selection analyses conducted in this study. The dataset encompassed experimental entries collected over five consecutive years (2018–2022) from multiple field environment locations, reflecting the typical operational rhythm of a public oat breeding program where candidate lines are evaluated under diverse environmental conditions to enable robust selection decisions. This temporal structure allowed for realistic assessment of genomic prediction performance under operational breeding scenarios.

Five key agronomic traits were consistently recorded across five years (2018–2022) at six North Dakota locations, yielding 21 unique environments (location × year combinations): plant height, test weight (TWT), yield, plump kernels, and thin kernels. Test weight and yield were the most frequently evaluated traits, being measured in 98 environments each, followed closely by plump and thin kernels (97 environments each). Plant height had slightly fewer observations (78 environments).

Phenotypic variation, based on raw field data, was substantial across all traits, indicating sufficient genetic variance for genomic selection implementation. Yield showed the broadest range of variation, averaging approximately 4,316 kg/ha with observed values spanning from ~ 2,127 to ~ 5,634 kg/ha, representing a 2.6-fold range that reflects both genetic differences and environmental effects. Test weight values averaged ~ 53.9 kg/hl, ranging from ~ 44.9 to ~ 60.0, while plant height varied from ~ 84 to ~ 127 cm with a mean of ~ 108 cm, reflecting genetic differences in plant architecture. Kernel quality traits also exhibited broad variation: plump kernels averaged ~ 2,044 kg/ha and thin kernels averaged ~ 53.3%, indicating substantial genetic control over grain filling characteristics.

Heritability and Genetic Correlations

Heritability estimates revealed strong genetic control for multiple traits, with important implications for genomic selection implementation. Thin kernels and plant height displayed generally high heritability values in most years (often > 0.85), although notable exceptions were observed (e.g., thins in 2020 and height in 2019). Yield exhibited moderate heritability (mean = 0.61, SD = 0.11), ranging from 0.43 to 0.75 across years, indicating substantial temporal variation in genetic signal strength. Temporal analysis highlighted considerable year-to-year variability in heritability across traits, with test weight and plump kernels exhibiting the highest fluctuations. Test weight had a mean heritability of 0.70 (SD = 0.34), ranging from 0.03 (2019) to 0.92 (2016). Similarly, plump kernels had a mean heritability of 0.60 (SD = 0.26), spanning from 0.08 (2016) to 0.81 (2020) (**Supplementary Table 1**).

The genetic correlation matrix provided insights into trait interdependencies relevant to multi-trait prediction strategies (Fig. 1). A strong positive genetic correlation was observed between plump kernels and yield ($r_g = 0.85$), indicating substantial shared genetic control and suggesting that leveraging one trait may improve prediction of the other in multi-trait models. Thin kernels were strongly negatively correlated with plumps ($r_g = -0.52$), consistent with the expected trade-off between kernel filling and the thin fraction. Height and test weight exhibited a weak positive genetic correlation ($r_g = 0.15$) and were retained as a weak-association comparator trait pair in correlation-based MTM evaluations. Correlations between thin kernels and height (-0.06), and between other trait pairs, were generally weak.

Strategy 1 - Forward Prediction

Three training schemes were implemented: a single-year training scenario (2018→2019), a two-year cumulative training (2018–2019→2020), and a three-year cumulative training (2018-2019-2020→2021). Both single-trait (STM) and multi-trait (MTM) models were tested across five traits, capturing variation in heritability and genetic complexity.

PA, measured as the Pearson correlation between genomic predictions and observed phenotypes, showed clear dependence on training data accumulation. Mean accuracy across all traits increased from 0.151 to 0.219 for single-trait models (STM) as training expanded from single-year (2018→2019) to three-year cumulative (2018-2019-2020→2021). Multi-trait models (MTM) showed a similar overall pattern (0.145 to 0.150) though with reduced performance in the two-year scheme (0.108). Individual predictions ranged from – 0.03 to 0.44, with the highest accuracies achieved for height (STM: 0.443; MTM: 0.432 for thin kernels) (Fig. 2), demonstrating substantial variation across traits and scenarios (**Supplementary Table 1**).

At the trait level, mean PA across all training schemes was highest for height (0.199) and thin kernels (0.188), while plump kernels and yield exhibited lower mean accuracies (0.082 and 0.081, respectively). Among MTM model configurations, the breeder-selected trait combination provided the highest mean accuracy (0.204), while correlation-based selection showed the lowest (0.175) (Fig. 2).

Comparing modeling approaches across all scenario-trait combinations, MTM achieved higher PA than STM in 73% of cases (11 of 15 direct comparisons), while STM outperformed MTM in 20% of cases (3 of

15), demonstrating the general utility of multi-trait approaches despite comparable overall mean accuracy across schemes (STM: 0.174; MTM: 0.132).

Hit rates (percentage of breeder-selected lines that were also picked by model) increased with training data accumulation and were highest under the three-year training scenario (2018–2019–2020 → 2021). Under this scheme, single-trait models (STM) achieved a maximum hit rate of 42.4% for plump kernels, while multi-trait models (MTM) reached comparable peak performance for plump kernels and test weight (40.9–41.0%). Across all trait–scenario–model comparisons, MTM outperformed STM in 20 of 54 cases for hit rate and in 30 of 54 cases for standardized selection differential, with the largest standardized differential observed for height under the two-year training scheme (MTM = 0.478).

To evaluate whether observed hit rates exceeded expectations under random selection, genomic selection outcomes were compared to random baselines generated using 1,000 Monte Carlo simulations at 30% selection intensity. The expected random hit rate was 29.96%. While mean hit rates averaged across all traits and scenarios approximated this baseline (MTM: 30.55%, STM: 30.56%), stratification by training scheme revealed a systematic dependence on training data accumulation. Single-year training schemes generally performed below random expectation (26–27%), two-year schemes were comparable to random (29–31%), whereas three-year cumulative training schemes consistently exceeded random performance for both STM (34.54%, + 4.6%) and MTM (35.52%, + 5.6%), with individual trait–scenario combinations exceeding random overlap by up to 12.7% (plump kernels).

To more directly reflect the multi-trait nature of breeder advancement decisions, combined trait-based selection strategies were evaluated by identifying genotypes ranked within the top 30% for multiple traits simultaneously (Fig. 3). Among all strategies tested, the strongest alignment with breeder selections was achieved when genotypes were prioritized based on membership in the top 30% for at least two traits, specifically height and test weight with equal weighting. Under this constrained multi-trait framework, MTM achieved the highest overall hit rate of 56.1%, while the corresponding STM strategy achieved a similarly elevated hit rate of 54.5%. These values substantially exceeded single-trait performance and reflected closer concordance with breeder decision-making. In contrast, increasingly restrictive strategies requiring genotypes to simultaneously rank within the top 30% for three or more traits resulted in sharp declines in hit rates, reflecting biological trade-offs among traits.

Strategy 2 - Backward Prediction

Backward temporal validation using three training schemes assessed model generalizability across increasing temporal distances between training and prediction datasets: 1-year (2020→2021), 2-year (2019–2020→2021, where "2-year" indicates training data from both 2019 and 2020), and 3-year (2018–2019–2020→2021). This approach evaluated whether models trained on more recent data could accurately predict performance of older genotypes, simulating scenarios where predictions must be made retrospectively.

Mean PA varied substantially by trait, training scheme, and model type (**Supplementary Table 1**). Despite this variation, averaging across all five traits revealed that PA generally improved with expanded training windows for STM (0.163 for 1-year, 0.205 for 2-year, 0.219 for 3-year), while MTM showed initial improvement (0.129 to 0.208) followed by decline in the 3-year scheme (0.188). This overall pattern, however, masked important trait-specific differences: highly heritable traits such as height maintained consistently strong PA across all schemes (STM: 0.386–0.424; MTM: 0.352–0.403), while yield and plump kernels showed substantially lower performance (0.021–0.122). Test weight exhibited notable model-dependent behavior, with MTM achieving particularly strong accuracy in the 2-year scheme (0.243) compared to STM (0.156), while thin kernels showed consistent STM advantage across schemes (0.151–0.254 vs. 0.120–0.217).

At the individual trait level, the highest prediction accuracies achieved (2018-2019-2020→2021) demonstrated substantial variation by trait and model type (Fig. 4). Height achieved the strongest and equivalent performance for both approaches (0.42), followed by test weight (MTM: 0.27; STM: 0.20) and thin kernels (MTM: 0.22; STM: 0.25). Yield and plump kernels exhibited lower maximum accuracies (MTM: 0.14 and 0.14; STM: 0.12 and 0.10, respectively). MTM achieved higher peak accuracy than STM for test weight (+ 0.07), yield (+ 0.02), and plump kernels (+ 0.04), while STM showed advantage for thin kernels (+ 0.03), with both approaches performing equally for height. Across the five traits, MTM achieved superior or equivalent maximum accuracy in four cases (height, test weight, yield, plump kernels), indicating that optimal model selection depends on breeding objectives and target trait profiles (Fig. 4).

Strategy 3 - Historical Prediction Stability

Historical prediction analysis evaluated model performance when trained on extended multi-year archives (2018-2019-2020) to predict 2021 performance, assessing the utility of accumulated historical data for genomic selection. This analysis addresses the practical question of whether maintaining extensive historical datasets improves prediction performance for operational breeding decisions.

PA varied substantially by trait and modeling approach (**Supplementary Table 1**). Comparing modeling approaches, multi-trait models outperformed single-trait models for three traits: test weight (best MTM: 0.293 vs. STM: 0.253, + 0.040), thin kernels (best MTM: 0.244 vs. STM: 0.232, + 0.012), and height (best MTM: 0.444 vs. STM: 0.435, + 0.009). Conversely, single-trait models performed better for yield (STM: 0.139 vs. best MTM: 0.119, + 0.020) and plump kernels (STM: 0.109 vs. best MTM: 0.080, + 0.029).

Among multi-trait model configurations, performance differences were generally modest. The All_Traits approach achieved the highest accuracy for height (0.444), test weight (0.293), and thin kernels (0.244), though differences between configurations were typically small (< 0.01) for most traits. For yield, all three MTM configurations performed similarly poorly (0.093–0.119), substantially below the STM benchmark (0.139). These patterns suggest that for traits where MTM approaches provide advantages (height, test weight, thin kernels), the specific trait selection strategy has limited impact on performance,

while for traits where MTM underperforms STM (yield, plump kernels), no MTM configuration provides substantial improvement.

The inclusion of Advanced Yield Trial (AYT) data in MTM training sets showed minimal to negative effects on PA (**Supplementary Fig. S2**). Comparing models trained with PYT-only data versus PYT + AYT data revealed that height and thin kernels showed no change (0.44 and 0.24, respectively), while test weight, yield, and plump kernels showed reduced accuracy when AYT data was incorporated (reductions of 0.02, 0.05, and 0.02, respectively). These results suggest that expanded training data from advanced trials does not enhance model performance when predicting preliminary yield trial outcomes and may introduce noise, underscoring the importance of training data relevance and trial stage alignment in genomic selection applications.

Strategy 4 - Cross-Validation Strategy Optimization

Cross-validation strategy comparison evaluated Random CV versus CDmean optimization approaches, addressing computational and performance trade-offs between simple random sampling and relationship matrix-based optimization. CDmean optimization and Random CV partitioning were conducted separately for each trait-year dataset using only genotypes with complete phenotypes. For single-trait models, this resulted in independent optimizations for 25 trait-year combinations (5 traits × 5 years); multi-trait models were optimized on genotypes phenotyped for all traits in each configuration.

For single-trait models, Random CV and CDmean produced comparable predictive ability (0.352 vs. 0.362) and selection efficiency. Hit rate (41.6% vs. 40.2%; $p = 0.676$), selection overlap (42.4% vs. 41.1%; $p = 0.735$), and standardized selection differential (0.95 vs. 0.885 SD units) did not differ significantly. However, Random CV maintained significantly higher rank correlation (0.925 vs. 0.859; $p < 0.001$), indicating superior genotype ordering despite similar predictive accuracy (Table 2).

Multi-trait model performance was evaluated across three trait-combination approaches: All_Traits (all five traits jointly), Breeder_Selected (domain knowledge-informed), and Correlation_Based (empirically determined). Across approaches combined, Random CV and CDmean showed no significant differences in predictive ability (0.217 vs. 0.168), hit rate (39.8% vs. 40.1%; $p = 0.852$), selection overlap (40.7% vs. 41.3%; $p = 0.800$), or standardized selection differential (0.758 vs. 0.703 SD units). Random CV again exhibited significantly higher rank correlation (0.773 vs. 0.642; $p < 0.001$) (**Supplementary Fig. S4**). Among the three approaches evaluated under Random CV, performance was statistically indistinguishable (ANOVA $p > 0.80$ for all metrics), with Correlation_Based showing numerically highest values across metrics, followed closely by Breeder_Selected and All_Traits.

Computational efficiency strongly favored Random CV, requiring substantially less time than CDmean's relationship matrix-based optimization. While CDmean contributed to some best-performing individual combinations, Random CV demonstrated superior rank correlation across both model types with dramatically reduced computational burden (Table 2). These findings support prioritizing Random CV for

routine implementation, with CDmean applied selectively where relationship-based optimization provides demonstrated benefits.

Table 2

Performance comparison between Random CV and CDmean optimization strategies across single-trait and multi-trait genomic prediction models.

Performance Metric	Single-Trait Models			Multi-Trait Models (All Approaches)		
	RandomCV	CDmean	p-value	RandomCV	CDmean	p-value
Mean Predictive Ability	0.352	0.362	-	0.217	0.168	-
Mean Hit Rate (%)	41.6	40.2	0.676	39.8	40.1	0.852
Mean Selection Overlap (%)	42.4	41.1	0.735	40.7	41.3	0.800
Mean Rank Correlation	0.925	0.859	< 0.001	0.773	0.642	< 0.001
Mean Std Selection Differential (SD)	0.95	0.885	-	0.758	0.703	-

p-values calculated using independent samples t-tests.

Advanced Analytical Framework

Four comprehensive analyses provided additional insights into trait performance, stability, and optimization strategies that inform genomic selection implementation.

Trait performance analysis

Across validation strategies revealed height as the most accurate but variable trait (accuracy = 0.211 ± 0.072 , hit rate = 26.7%, standardized Selection Differential = 0.036), while thins demonstrated consistent balanced performance (Fig. 5). Statistical analysis confirmed significant trait differences across validation strategies ($F = 4.23$, $p = 0.017$), emphasizing the importance of trait-specific optimization strategies.

Multi Trait Model combination analysis: Multi-trait model performance varied by trait combination approach, though differences between strategies were not statistically significant (**Supplementary Fig. S5**). Across all trait-year evaluations, mean prediction accuracies were similar for Breeder_Selected (0.21), All_Traits (0.20), and Correlation_Based (0.19) approaches, indicating that the choice of trait combination strategy had modest impact on overall performance. However, trait-specific patterns revealed more nuanced differences in optimal configurations. Examining best-performing combinations by trait, the All_Traits approach (including all five traits simultaneously) achieved highest accuracy for height (0.43), thin kernels (0.67), and test weight (0.39). In contrast, targeted two-trait combinations

excelled for specific traits: yield_height achieved highest accuracy for yield (0.31), while thins_plumps performed best for plump kernels (0.52). These patterns suggest that for highly heritable traits with broad genetic correlations (height, thin kernels), incorporating all available trait information enhances prediction, while traits with more specific or complex genetic architectures (yield, plump kernels) benefit from selective trait pairing. Performance matrices for specific trait combinations revealed consistent patterns across approaches. Among Breeder_Selected combinations, yield_height_twt provided balanced performance across multiple traits (height: 0.279, test weight: 0.180, yield: 0.127), while yield_height showed comparable height prediction (0.271) with slightly reduced yield accuracy (0.125). The thins_plumps combination achieved moderate accuracies for both traits (0.258 and 0.148, respectively). Correlation_Based combinations showed similar performance profiles: height_twt achieved accuracies of 0.271 and 0.173 for its component traits, while plumps_yield reached 0.142 and 0.119. The All_Traits configuration demonstrated moderate but stable performance across all five traits (0.142–0.265), though generally underperforming trait-specific combinations for individual trait predictions. These results indicate that while multi-trait approaches provide advantages over single-trait models for certain traits, the specific trait combination selected influences PA less than whether a multi-trait approach is appropriate for the trait's genetic architecture.

Environmental and temporal stability analysis revealed trait-dependent variation in model robustness (**Supplementary Fig. S6**). Temporal stability, quantified as CV across the four validation strategies, showed height exhibited the highest consistency (mean CV = 0.268), followed by thin kernels (CV = 0.478) and test weight (CV = 0.612), while yield (CV = 0.735) and plump kernels (CV = 0.797) showed greater sensitivity to validation design. Environmental stability, assessed using Strategy 4 predictions across crop years (2018–2022), showed similar rankings, with height achieving the highest score (0.779), followed by test weight (0.654) and thin kernels (0.609), while yield (0.606) and plump kernels (0.550) demonstrated greater sensitivity to inter-annual environmental variation. Temporal patterns showed variation in prediction accuracies across crop years (2018–2022), with performance differences reflecting annual changes in environmental conditions.

Accuracy-Selection differential relationship analysis: To directly evaluate whether predictive ability reflects breeding utility, we quantified the relationship between PA and realized selection differential across all prediction scenarios. The relationship between PA and realized selection differential revealed a moderate but statistically significant positive correlation (Fig. 6A), suggesting an overall tendency for higher PA to improve selection outcomes among top-ranked candidates. Trait-specific analysis showed substantial heterogeneity in this relationship (Fig. 6B), with kernel quality traits exhibiting statistically significant correlations, Plumps ($r = 0.419$, $p = 0.005$) and Thins ($r = 0.376$, $p = 0.031$), whereas agronomic traits showed weak and non-significant associations: TWT ($r = 0.251$, $p = 0.159$), Height ($r = 0.104$, $p = 0.503$), and Yield ($r = -0.121$, $p = 0.435$).

Discussion

This comprehensive evaluation of genomic prediction strategies in oat breeding provides critical insights that challenge several conventional assumptions in the field while offering practical guidance for implementation. The integration of four validation strategies with five analytical frameworks reveals the complexity of optimizing genomic selection beyond simple accuracy maximization, highlighting the importance of considering selection efficiency, temporal stability, and resource allocation in breeding program design. These findings have broad implications for how the genomic selection community approaches model development, evaluation, and implementation across diverse crop species and breeding contexts.

Rethinking Trait Prioritization: Beyond PA

The trait-specific performance patterns observed in this study underscore the need to complement predictive ability with selection outcome metrics when evaluating genomic selection value for operational breeding programs. While predictive ability, quantified as rank correlation between genomic predictions and observed phenotypes, remains fundamental for model evaluation (Meuwissen et al., 2001; Goddard & Hayes, 2007), our results demonstrate that this single metric provides an incomplete picture of breeding value, particularly when programs simultaneously optimize multiple traits with differential priorities. Our aim is not to suggest that selection outcome metrics supersede predictive ability, but rather to demonstrate that genomic selection can add value to breeding programs even with modest predictive ability by enriching the top candidates that breeders actually advance.

The most significant finding is that strategic multi-trait selection frameworks substantially improved alignment with breeder decisions beyond single-trait approaches. When genomic selection prioritized genotypes meeting top-30% criteria for two breeder-selected traits (height and test weight), hit rates in forward prediction reached 56.1% (MTM) and 54.5% (STM), exceeding the best single-trait performance (42.4% for plump kernels) by 13–14 percentage points and surpassing random baseline expectations (29.96%) by 26 percentage points. Remarkably, these forward prediction values approached or exceeded within-year cross-validation performance (Strategy 4: mean $\sim 40\%$, individual traits 45–59%), despite the substantially more challenging task of predicting across crop years where temporal effects, environmental variation, and germplasm turnover reduce predictive ability (Heslot et al., 2012). This demonstrates that multi-trait selection frameworks can compensate for the genotype-by-environment interactions and temporal drift inherent in operational breeding contexts, making them particularly valuable for realistic implementation. Single-trait performance patterns revealed substantial variation reflecting both genetic architecture and breeder prioritization. Thin kernels demonstrated consistent predictive ability ($\rho = 0.204 \pm 0.0388$) and the highest selection differential ($+ 0.697$ SD), likely reflecting simpler genetic architecture controlled by fewer major genes with stable effects characteristic of kernel morphology traits (Tinker et al., 2009; Huang et al., 2014; Yan et al., 2017; Bekele et al., 2018). In contrast, plump kernels achieved the highest single-trait hit rate (42.4%) despite modest predictive ability ($\rho = 0.099$), suggesting that this trait, as a critical market quality determinant, receives elevated priority in breeder advancement decisions regardless of genomic prediction precision. This distinction has important implications: hit rates reflect both model predictive performance and trait importance in

operational decision-making, indicating that different traits may require different evaluation frameworks even when using identical genomic prediction methodologies (Heslot et al., 2012; Lozada et al., 2019). This confounding represents a key limitation when interpreting single-trait hit rates: comparing individual trait genomic predictions against multi-trait breeder selections conflates model quality with trait weighting, such that even perfect predictive ability for a deprioritized trait may yield modest hit rates. Calculating realized selection indexes from actual breeder decisions could quantify trait importance explicitly, though such analysis was not feasible with the current dataset structure. This limitation underscores the value of the two-trait selection framework evaluated here, which directly addresses the multi-trait selection reality rather than attempting to infer trait weights from single-trait outcomes.

Critically, comparison to random baseline expectations (29.96% under 30% selection intensity) demonstrates that observed outcomes reflect genuine predictive signal beyond chance overlap or trait prioritization alone. Training data accumulation showed systematic progression: single-year training performed below random (26–27%), two-year approximated random (29–31%), while three-year cumulative training exceeded random baselines by 4.6–5.6%, with individual traits improving up to 12.7% above chance. This progression provides key evidence: if hit rates solely reflected breeder trait priorities independent of model quality, all training scenarios would show similar overlap regardless of training population size. The observation that inadequate training produces below-random performance demonstrates that predictive ability meaningfully translates to breeding value when models are supported by sufficient data, and that genomic selection requires strategic implementation to realize value. Yield exhibited variable and modest predictive ability ($p = 0.024$ – 0.122) yet achieved meaningful hit rates (maximum 36.4%), reinforcing long-standing observations about the complexity of this trait involving numerous small-effect loci, epistatic interactions, and substantial genotype-by-environment components (Bernardo & Yu, 2007; Lorenz et al., 2013). However, these selection outcomes, particularly when evaluated using multi-trait frameworks, suggest that breeding programs may have been overly conservative in dismissing genomic selection for complex agronomic traits based solely on correlation-based predictive ability (Crossa et al., 2017; Hickey et al., 2019). Alternative evaluation metrics focused on selection outcomes rather than prediction precision may reveal broader utility across trait types when implementation incorporates strategic multi-trait frameworks, systematic training data accumulation, and validation approaches that realistically assess operational performance.

The PA-Selection Differential Relationship: Implications for Model Evaluation

Predictive ability (PA) and selection differential showed only moderate alignment in our study ($r = 0.336$, $p = 1.3 \times 10^{-6}$), illustrating why PA, while fundamental for model evaluation, provides an incomplete picture of genomic selection value for operational breeding programs. This decoupling arises because PA reflects population-wide prediction performance, whereas selection differential depends on the genetic superiority of the top-ranked fraction that breeders actually advance. Population structure and the distribution of genotypic values can therefore cause these metrics to diverge, particularly when superior

genotypes cluster non-randomly or when prediction errors disproportionately affect distributional tails where selection occurs (Daetwyler et al., 2010; VanRaden, 2008; de los Campos et al., 2013).

Trait-specific correlations reveal fundamental differences in how genetic architecture affects PA-selection differential relationships. Only kernel quality traits, plump ($r = 0.419$, $p = 0.005$) and thin kernels ($r = 0.376$, $p = 0.031$), demonstrated statistically significant correlations, consistent with their higher heritability and simpler genetic architecture controlled by fewer major genes (Rutkoski et al., 2012; Huang et al., 2016). In contrast, all agronomic traits showed weak or non-significant relationships, with yield exhibiting a slightly negative correlation ($r = -0.121$, $p = 0.435$). For these complex traits, factors beyond predictive ability, including genotype-by-environment interactions, epistatic effects, or population substructure, substantially influence realized selection differential (Heffner et al., 2011; Poland et al., 2012; Araus et al., 2008; Reynolds et al., 2009). This trait-dependent variation underscores that evaluation frameworks must account for biological complexity when assessing genomic selection value across diverse breeding targets.

These findings reinforce the need to complement predictive ability with selection outcome metrics when evaluating genomic prediction methods or training designs. While PA remains essential for assessing model quality, incorporating hit rate and standardized selection differential provides direct evidence of whether predictions translate to genetic gain in operational contexts where candidates are advanced based on rank. This combined evaluation framework ensures that model comparisons reflect performance in the decision context where breeding value is ultimately realized.

Multi-Trait Modeling: Context-Dependent Benefits

The comparative evaluation of single-trait and multi-trait models reveals that MTM benefits are highly context-dependent. While some studies demonstrate clear advantages of multi-trait approaches (Jia & Jannink, 2012; Calus & Veerkamp, 2011), others report superior single-trait performance under certain conditions (Brault et al., 2021). In our study, single-trait models outperformed multi-trait models in 18 of 25 comparisons, suggesting that theoretical benefits of information borrowing may not translate to practical improvements when trait correlations are weak or training sets are limited.

Multi-trait models showed advantages for specific traits with strong biological relationships. Test weight benefited from joint modeling (MTM PA = 0.289 vs STM PA = 0.228), likely reflecting shared grain filling processes (Haile et al., 2018). Similarly, the strong negative correlation between thin and plump kernels (-0.65) enabled effective joint modeling (Peltonen-Sainio et al., 2007). The lack of significant differences among multi-trait strategies ($p = 0.7121$) suggests that simpler trait combination approaches may be as effective as sophisticated correlation-based methods (Guo et al., 2014), potentially reducing implementation complexity.

Cross-Validation Strategy: Efficiency Versus Sophistication

Random cross-validation and CDmean optimization produced comparable selection efficiency across single-trait and multi-trait models, with no significant differences in hit rate (STM: 41.6% vs 40.2%, $p = 0.676$; MTM: 39.8% vs 40.1%, $p = 0.852$) or selection overlap (STM: 42.4% vs 41.1%, $p = 0.735$; MTM: 40.7% vs 41.3%, $p = 0.800$). However, Random CV maintained significantly higher rank correlation with phenotypic rankings in both contexts (STM: 0.925 vs 0.859, $p < 0.001$; MTM: 0.773 vs 0.642, $p < 0.001$), while requiring substantially less computational time. This performance pattern suggests that while both methods achieve similar selection outcomes, Random CV provides more reliable genotype ordering, a critical advantage for breeding decisions, without the computational overhead of relationship matrix-based optimization. CDmean optimization is designed to reduce phenotyping costs by selecting informative subsets that maintain predictive ability with fewer genotypes (Saatchi et al., 2011; Rincent et al., 2012), not to maximize performance with unrestricted training populations. When training sets are large and phenotyping resources are not limiting, random sampling may be sufficient and computationally efficient (Endelman, 2011). These results suggest that training set optimization methods like CDmean are most valuable when phenotyping budgets constrain training population size, rather than as a universal strategy for improving genomic selection performance.

Implications for Breeding Program Implementation

The collective findings of this study support a strategic, context-dependent approach to genomic selection implementation that challenges the pursuit of universal optimization solutions. The superior performance of multi-trait selection frameworks, particularly when aligned with breeder priorities (56.1% hit rate for height and test weight combined), demonstrates that genomic selection delivers greatest value when implementation mirrors operational decision-making rather than maximizing single-trait predictive ability. This finding has immediate practical relevance: breeding programs should design genomic selection strategies around their actual selection criteria, which typically involve simultaneous evaluation of multiple traits with differential priorities, rather than optimizing individual trait predictions in isolation (Heslot et al., 2012; Crossa et al., 2017).

The moderate correlation between predictive ability and selection differential ($r = 0.336$) reinforces that evaluation frameworks must extend beyond correlation-based metrics to assess practical breeding value. Yield exemplifies this principle: despite variable and modest predictive ability ($\rho = 0.024$ – 0.122), this trait achieved meaningful hit rates (maximum 36.4%) and substantial selection differentials ($+ 0.9$ SD), demonstrating that moderate-accuracy models can deliver genetic gain when evaluated through selection outcome metrics. This suggests that breeding programs may have been overly conservative in dismissing genomic selection for complex agronomic traits based solely on predictive ability (Bernardo & Yu, 2007; Lorenz et al., 2011; Hickey et al., 2019). Complementing prediction precision with selection efficiency metrics better aligns genomic selection research with breeding program objectives and may reveal broader utility across trait types, particularly for integrative traits influenced by genotype-by-environment interactions.

Implementation accessibility represents a critical practical consideration. The comparable performance of single-trait and multi-trait models, combined with the superior efficiency of random cross-validation (achieving equivalent selection outcomes while requiring substantially less computational time than CDmean optimization), demonstrates that effective genomic selection can be implemented without sophisticated modeling approaches or extensive computational infrastructure (Thorwarth et al., 2017; Tiede & Smith, 2018). This finding is particularly relevant for public breeding programs, smaller commercial operations, and crops in resource-limited regions where genotyping and computational resources constrain adoption (Cossa et al., 2017). However, training set optimization methods like CDmean remain valuable when phenotyping budgets limit training population size, as their design purpose is reducing phenotyping costs while maintaining predictive ability rather than maximizing performance with unrestricted populations (Rincent et al., 2012; Endelman, 2011).

The systematic progression in training data accumulation, from below-random performance with single-year training (26–27%) to exceeding random baselines by up to 12.7% with three-year cumulative training, underscores that genomic selection requires strategic implementation to realize value. This evidence demonstrates that predictive ability meaningfully translates to breeding value when models are supported by sufficient training data, and that implementation success depends on thoughtful integration of multi-trait frameworks, realistic validation approaches, and evaluation metrics aligned with operational breeding contexts.

Conclusion

This comprehensive evaluation of genomic selection strategies in oat demonstrates that optimizing breeding value requires moving beyond single-metric evaluation frameworks to incorporate selection outcome measures that directly assess genetic gains. The moderate correlation between predictive ability and selection differential, combined with the superior performance of multi-trait selection frameworks aligned with breeder priorities, establishes that predictive ability, while fundamental, provides an incomplete picture of genomic selection value for operational breeding programs. These findings challenge the conventional emphasis on maximizing correlation-based PA and support evaluation frameworks that complement predictive ability with hit rate and selection differential metrics.

The context-dependent nature of optimization strategies, including the comparable performance of single-trait and multi-trait models, the substantial efficiency advantages of random cross-validation, and the trait-specific benefits of multi-trait modeling, demonstrates that effective implementation requires strategic decision-making rather than universal solutions. Within-year cross-validation revealed that multiple trait- combination approaches (breeder-defined, correlation-based, all-traits) achieved statistically indistinguishable performance, providing breeding programs with implementation flexibility rather than mandating specific optimization strategies. The four-strategy validation framework developed here provides breeding programs with comprehensive assessment tools spanning forward prediction, temporal effects, historical data integration, and within-year optimization. The systematic progression from below-random performance with inadequate training to substantial improvements with

three-year cumulative data underscores that genomic selection success depends on thoughtful integration of sufficient training populations, multi-trait frameworks that mirror operational decision-making, and validation approaches that realistically assess breeding outcomes.

Declarations

Supplementary Information

The online version contains supplementary material available at

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

Funding

This research used resources provided by the USDA-ARS SCINet project 0201-88888-003-000D and USDA-ARS CRIS project number 3060-21000-046-000-D.

Author Contribution

Muhammad Massub Tehseen contributed to conceptualization, formal analysis, data curation, validation, visualization, drafting the manuscript. Charlotte Brault was involved in methodology, validation, and writing, review and editing. Naa Korkoi Ardayfio contributed to investigation, data curation, methodology, and writing, review and editing. Sepehr Mohajeri Naraghi was involved in methodology, resources, data curation, review and editing. Michael S. McMullen contributed to methodology and resources. Jason D. Fiedler was involved in conceptualization, methodology, validation, resources, supervision, funding acquisition, project administration, and writing, review and editing. All authors reviewed and edited the manuscript and approved the final version.

Data Availability

The phenotypic and genotypic datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request. The R scripts and code used for genomic

prediction models, data processing, and figure generation are also available from the corresponding author upon reasonable request.

References

1. Akdemir, D., & Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Scientific Reports*, 9(1), 1446. <https://doi.org/10.1038/s41598-018-38081-6>
2. Alemu, A., Åstrand, J., Montesinos-López, O. A., Isidro Y Sánchez, J., Fernández-González, J., Tadesse, W., Vetukuri, R. R., Carlsson, A. S., Ceplitis, A., Crossa, J., Ortiz, R., & Chawade, A. (2024). Genomic selection in plant breeding: Key factors shaping two decades of progress. *Molecular Plant*, 17(4), 552–578. <https://doi.org/10.1016/j.molp.2024.03.007>
3. Araus, J. L., Slafer, G. A., Royo, C., & Serret, M. D. (2008). Breeding for yield potential and stress adaptation in cereals. *Critical Reviews in Plant Sciences*, 27(6), 377–412. <https://doi.org/10.1080/07352680802467736>
4. Ardayfio, N. K., Kinney, C., Li, D., Naraghi, S. M., McMullen, M. S., & Fiedler, J. D. (2025). Genome-wide association studies reveal genetic control of nutritional quality, milling traits, and agronomic characteristics in oat (*Avena sativa* L.). *The Plant Genome*, 18(3), e70060.
5. Bekele, W. A., Wight, C. P., Chao, S., Howarth, C. J., & Tinker, N. A. (2018). Haplotype-based genotyping-by-sequencing in oat genome research. *Plant Biotechnology Journal*, 16(8), 1452–1463. <https://doi.org/10.1111/pbi.12888>
6. Bernardo, R. (2020). Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. *Heredity*, 125(6), 375–385. <https://doi.org/10.1038/s41437-020-0312-1>
7. Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, 47(3), 1082–1090. <https://doi.org/10.2135/cropsci2006.11.0690>
8. Brault, C., Labrousse, Y., Rincet, R., Remoué, C., Goudemand, E., Girousse, C., Bogard, M., Heumez, E., Orford, S., Griffiths, S., Foulkes, J., & Le Gouis, J. (2021). Genome-wide association mapping and genomic prediction for grain yield in a winter wheat population. *Plant Genome*, 14(3), e20148. <https://doi.org/10.1002/tpg2.20148>
9. Calus, M. P., & Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution*, 43(1), 26. <https://doi.org/10.1186/1297-9686-43-26>
10. Cooper, M., Messina, C. D., Podlich, D., Totir, L. R., Baumgarten, A., Hausmann, N. J., Wright, M. H., & Hammer, G. L. (2014). Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. *Crop and Pasture Science*, 65(4), 311–336. <https://doi.org/10.1071/CP14007>
11. Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic selection in

- plant breeding: methods, models, and perspectives. *Trends in Plant Science*, 22(11), 961–975. <https://doi.org/10.1016/j.tplants.2017.08.011>
12. Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, 185(3), 1021–1031. <https://doi.org/10.1534/genetics.110.116855>
 13. de los Campos, G., & Grüneberg, A. (2016). MTM (Multi-Trait Model) package. R package version 1.0. <https://CRAN.R-project.org/package=MTM>
 14. de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193(2), 327–345. <https://doi.org/10.1534/genetics.112.143313>
 15. Dekkers, J. C., Su, H., & Cheng, J. (2021). Predicting the accuracy of genomic predictions. *Genetics Selection Evolution*, 53, 47. <https://doi.org/10.1186/s12711-021-00647-w>
 16. Desta, Z. A., & Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science*, 19(9), 592–601. <https://doi.org/10.1016/j.tplants.2014.05.006>
 17. Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*, 4(3), 250–255. <https://doi.org/10.3835/plantgenome2011.08.0024>
 18. Fernandes, S. B., Dias, K. O., Ferreira, D. F., & Brown, P. J. (2018). Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theoretical and Applied Genetics*, 131(3), 747–755. <https://doi.org/10.1007/s00122-017-3033-y>
 19. Goddard, M. E., & Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*, 124(6), 323–330. <https://doi.org/10.1111/j.1439-0388.2007.00702.x>
 20. Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., & Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genetics*, 15(1), 30. <https://doi.org/10.1186/1471-2156-15-30>
 21. Haile, J. K., N'Diaye, A., Clarke, F., Clarke, J., Knox, R., Rutkoski, J., Bassi, F. M., & Pozniak, C. J. (2018). Genomic selection for grain yield and quality traits in durum wheat. *Molecular Breeding*, 38(6), 75. <https://doi.org/10.1007/s11032-018-0818-x>
 22. Heffner, E. L., Jannink, J. L., & Sorrells, M. E. (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Genome*, 4(1), 65–75. <https://doi.org/10.3835/plantgenome2010.12.0029>
 23. Heffner, E. L., Sorrells, M. E., & Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Science*, 49(1), 1–12. <https://doi.org/10.2135/cropsci2008.08.0512>
 24. Heslot, N., Yang, H. P., Sorrells, M. E., & Brown-Guedira, G. L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Science*, 52(1), 146–160. <https://doi.org/10.2135/cropsci2011.06.0297>
 25. Hickey, L. T., Hafeez, A. N., Robinson, H., Jackson, S. A., Leal-Bertioli, S. C., Tester, M., Gao, C., Godwin, I. D., Hayes, B. J., & Wulff, B. B. (2019). Breeding crops to feed 10 billion. *Nature Biotechnology*, 37(7), 744–754. <https://doi.org/10.1038/s41587-019-0152-9>

26. Howard, R., Carriquiry, A. L., & Beavis, W. D. (2014). Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3: Genes, Genomes, Genetics*, 4(6), 1027–1046. <https://doi.org/10.1534/g3.114.010298>
27. Huang, M., Cabrera, A., Hoffstetter, A., Griffey, C., Van Sanford, D., Costa, J., McKendry, A., Chao, S., & Sneller, C. (2016). Genomic selection for wheat traits and trait stability. *Theoretical and Applied Genetics*, 129(9), 1697–1710. <https://doi.org/10.1007/s00122-016-2733-z>
28. Huang, Y. F., Poland, J. A., Wight, C. P., Jackson, E. W., & Tinker, N. A. (2014). Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. *PLoS One*, 9(7), e102448. <https://doi.org/10.1371/journal.pone.0102448>
29. Jia, Y., & Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value predictive ability. *Genetics*, 192(4), 1513–1522. <https://doi.org/10.1534/genetics.112.144246>
30. Lorenz, A. J. (2013). Resource allocation for maximizing prediction accuracy and selection differential of genomic selection in plant breeding: a simulation experiment. *G3: Genes, Genomes, Genetics*, 3(3), 481–491. <https://doi.org/10.1534/g3.112.004911>
31. Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., Smith, K. P., Sorrells, M. E., & Jannink, J. L. (2011). Genomic selection in rice breeding. *Plant Genome*, 4(1), 22–27. <https://doi.org/10.3835/plantgenome2010.04.0022>
32. Lozada, D. N., Mason, R. E., Sarinelli, J. M., & Brown-Guedira, G. (2019). Accuracy of genomic selection for grain yield and agronomic traits in soft red winter wheat. *BMC Genetics*, 20(1), 82. <https://doi.org/10.1186/s12863-019-0785-1>
33. Lush, J. L. (1937). *Animal breeding plans*. Iowa State College Press.
34. Mellers, G., Mackay, I., Cowan, A., Griffiths, I., Martinez-Martin, P., Spannagl, M., Pozniak, C. J., Foulkes, M. J., & Bentley, A. R. (2020). Implementing genomic selection in wheat breeding programs. *Plant Genome*, 13(3), e20064. <https://doi.org/10.1002/tpg2.20064>
35. Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4), 1819–1829. <https://doi.org/10.1093/genetics/157.4.1819>
36. Michel, S., Löschenberger, F., Ametz, C., Pachler, B., Sperry, E., & Bürstmayr, H. (2019). Combining grain yield, protein content and protein quality by multi-trait genomic selection in bread wheat. *Theoretical and Applied Genetics*, 132(8), 2767–2784. <https://doi.org/10.1007/s00122-019-03386-1>
37. Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Gianola, D., Hernández-Suárez, C. M., & Martín-Vallejo, J. (2018). Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. *G3: Genes, Genomes, Genetics*, 8(12), 3829–3840. <https://doi.org/10.1534/g3.118.200728>
38. Peltonen-Sainio, P., Kangas, A., Salo, Y., & Jauhiainen, L. (2007). Grain number dominates grain weight in temperate cereal yield determination: evidence based on 30 years of multi-location environments. *Field Crops Research*, 100(2–3), 179–188. <https://doi.org/10.1016/j.fcr.2006.07.002>

39. Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M., & Jannink, J. L. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome*, 5(3), 103–113.
<https://doi.org/10.3835/plantgenome2012.06.0006>
40. Reynolds, M., Foulkes, M. J., Slafer, G. A., Berry, P., Parry, M. A., Snape, J. W., & Angus, W. J. (2009). Raising yield potential in wheat. *Journal of Experimental Botany*, 60(7), 1899–1918.
<https://doi.org/10.1093/jxb/erp016>
41. Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., ... Moreau, L. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics*, 192(2), 715–728.
42. Rodríguez-Álvarez, M. X., Boer, M. P., van Eeuwijk, F. A., & Eilers, P. H. (2018). Correcting for spatial heterogeneity in plant breeding experiments with P-splines. *Spatial Statistics*, 23, 52–71.
<https://doi.org/10.1016/j.spasta.2017.10.003>
43. Rutkoski, J., Benson, J., Jia, Y., Brown-Guedira, G., Jannink, J. L., & Sorrells, M. (2012). Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. *Plant Genome*, 5(2), 51–61. <https://doi.org/10.3835/plantgenome2012.02.0001>
44. Saatchi, M., McClure, M. C., McKay, S. D., Rolf, M. M., Kim, J., Decker, J. E., Taxis, T. M., Chapple, R. H., Ramey, H. R., Northcutt, S. L., Bauck, S., Woodward, B., Dekkers, J. C., Fernando, R. L., Schnabel, R. D., Garrick, D. J., & Taylor, J. F. (2011). Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetics Selection Evolution*, 43(1), 40.
<https://doi.org/10.1186/1297-9686-43-40>
45. Sallam, A. H., Endelman, J. B., Jannink, J. L., & Smith, K. P. (2015). Assessing genomic selection prediction accuracy in a dynamic barley breeding population. *Plant Genome*, 8(1), plantgenome2014.05.0020. <https://doi.org/10.3835/plantgenome2014.05.0020>
46. Thorwarth, P., Ahlemeyer, J., Bochard, A. M., Krumnacker, K., Blümel, H., Laubach, E., Knöchel, N., Cselényi, L., Ordon, F., & Longin, C. F. H. (2017). Genomic prediction ability for yield-related traits in German winter barley elite material. *Theoretical and Applied Genetics*, 130(8), 1669–1683.
<https://doi.org/10.1007/s00122-017-2917-1>
47. Tiede, T., & Smith, K. P. (2018). Evaluation and comparison of genomic selection methods for small grain improvement. *Crop Science*, 58(1), 299–312. <https://doi.org/10.2135/cropsci2017.05.0271>
48. Tinker, N. A., Kilian, A., Wight, C. P., Heller-Uszynska, K., Wenzl, P., Rines, H. W., Bjørnstad, Å., Howarth, C. J., Jannink, J. L., Anderson, J. M., Rossnagel, B. G., Stuthman, D. D., Sorrells, M. E., Jackson, E. W., Tuveson, S., Kolb, F. L., Olsson, O., Federizzi, L. C., Carson, M. L., Ohm, H. W., Molnar, S. J., Scoles, G. J., Eckstein, P. E., Bonman, J. M., Ceplitis, A., & Langdon, T. (2009). New DArT markers for oat provide enhanced map coverage and global germplasm characterization. *BMC Genomics*, 10(1), 39.
<https://doi.org/10.1186/1471-2164-10-39>

49. VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11), 4414–4423. <https://doi.org/10.3168/jds.2007-0980>
50. Voss-Fels, K. P., Cooper, M., & Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics*, 132(3), 669–686. <https://doi.org/10.1007/s00122-018-3270-8>
51. Yan, W., Tinker, N. A., Bekele, W. A., Hodson, D., Selles, F., Navabi, A., Eckstein, P. E., Rosnagel, B. G., Scoles, G. J., & Hucl, P. (2017). Theoretical unification and practical integration of conventional and molecular marker-based plant breeding. *Crop Science*, 57(5), 2372–2382. <https://doi.org/10.2135/cropsci2016.09.0780>
52. Zhang, A., Wang, H., Beyene, Y., Semagn, K., Liu, Y., Cao, S., Cui, Z., Ruan, Y., Burgueño, J., San Vicente, F., Olsen, M., Prasanna, B. M., Crossa, J., Yu, H., & Zhang, X. (2017). Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. *Frontiers in Plant Science*, 8, 1916. <https://doi.org/10.3389/fpls.2017.01916>
53. Zhao, Y., Gowda, M., Liu, W., Würschum, T., Maurer, H. P., Longin, F. H., Ranc, N., & Reif, J. C. (2012). Accuracy of genomic selection in European maize elite breeding populations. *Theoretical and Applied Genetics*, 124(4), 769–776. <https://doi.org/10.1007/s00122-011-1745-y>

Figures

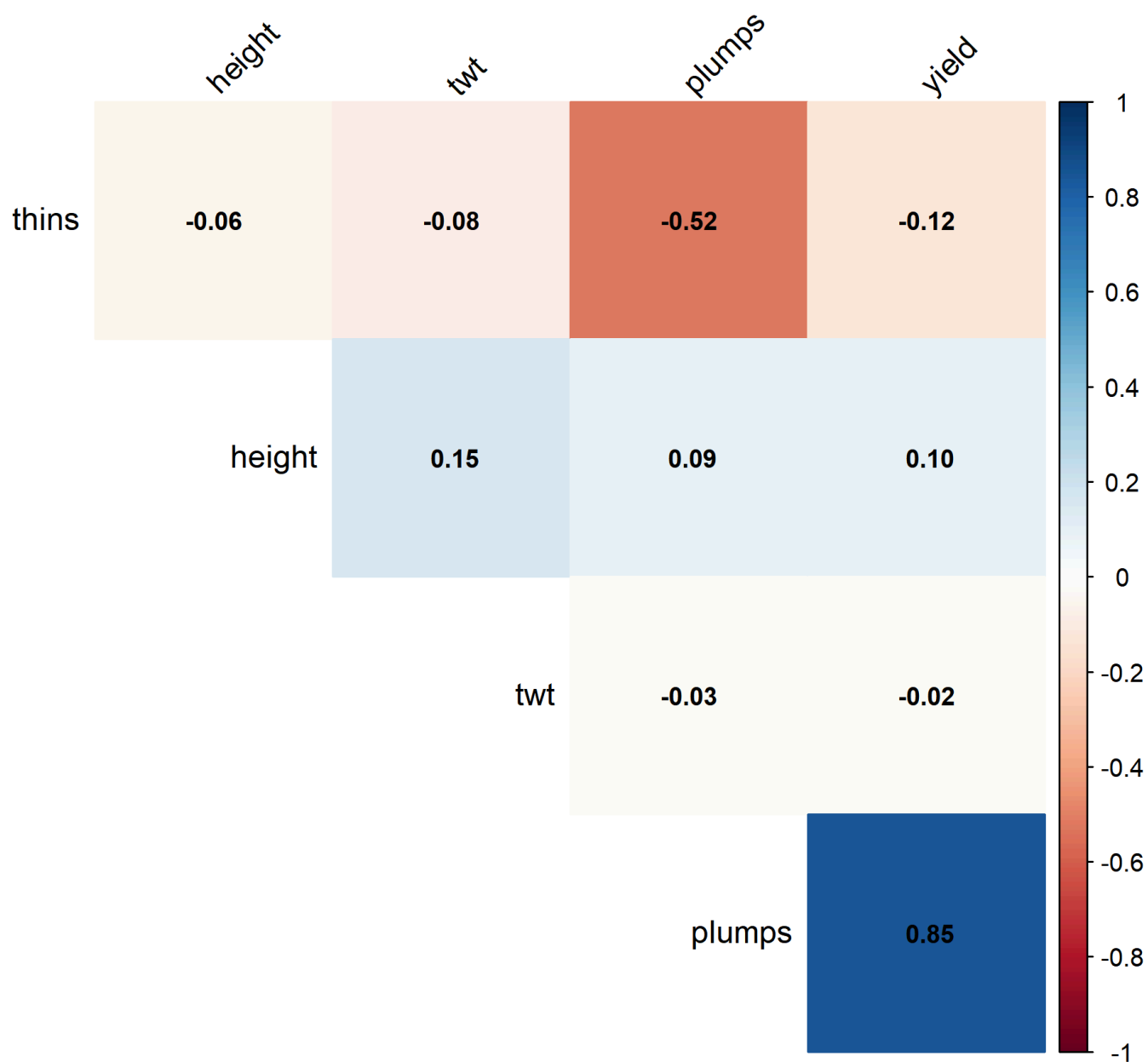


Figure 1

Genetic correlations among five agronomic traits in oat. Twt: test weight.

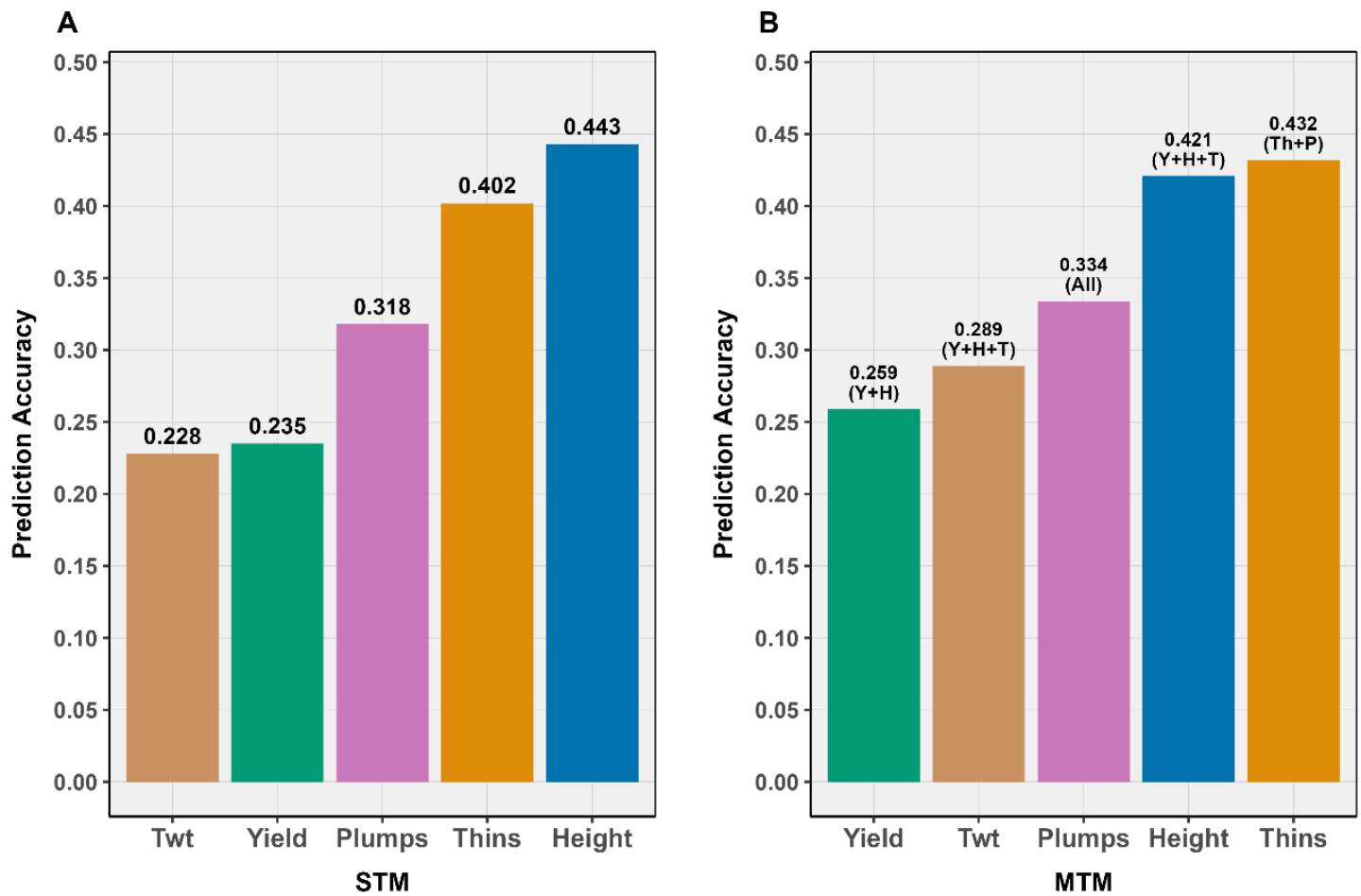


Figure 2

Trait-wise PA's in Strategy 1 forward prediction scenarios. Panel A shows the best single-trait PA for each trait. Panel B presents the highest multi-trait PA achieved per trait based on optimal trait combinations. Abbreviations in panel B: Y=Yield, H=Height, T=TWT, Th=Thins, P=Plumps.

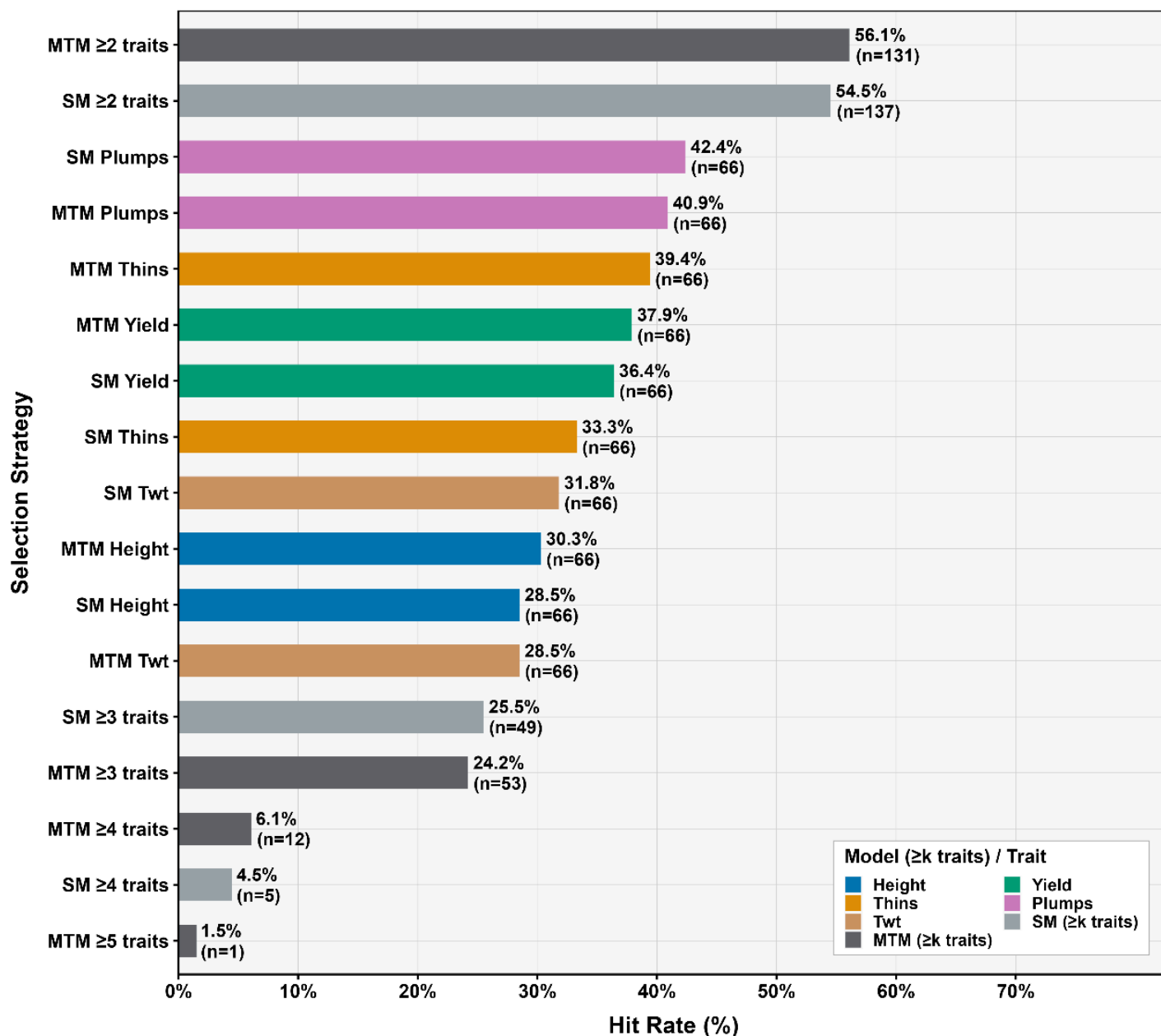


Figure 3

Selection performance based on combined trait prioritization under a 30% selection threshold, three-year cumulative scenario (2018–2019–2020 → 2021). Hit rate = % of breeder-advanced lines recovered; n = number of lines selected.

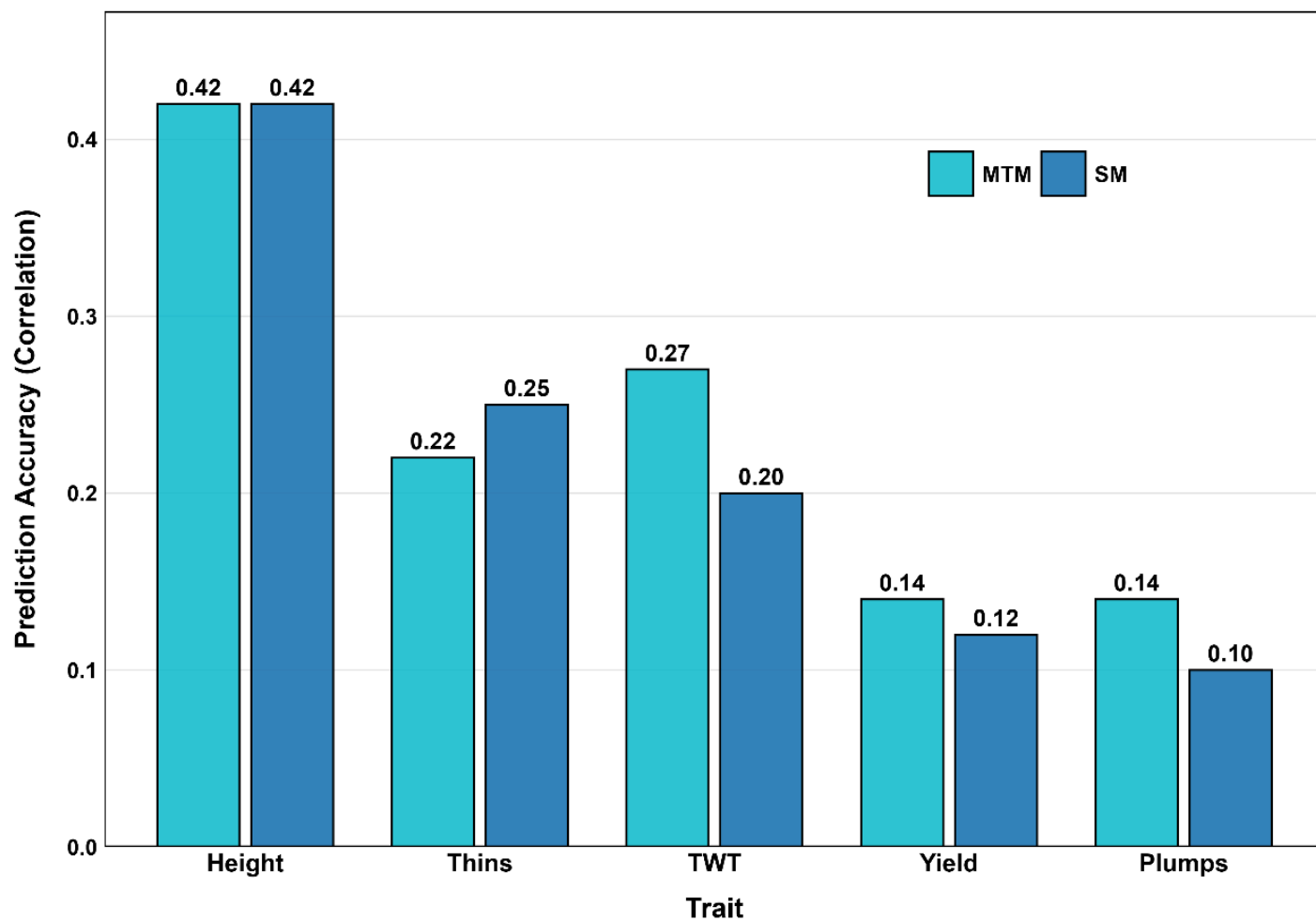


Figure 4

Trait-wise mean PA between single-trait (SM) and multi-trait (MTM) models in backward prediction with varying training set sizes (Strategy 2).

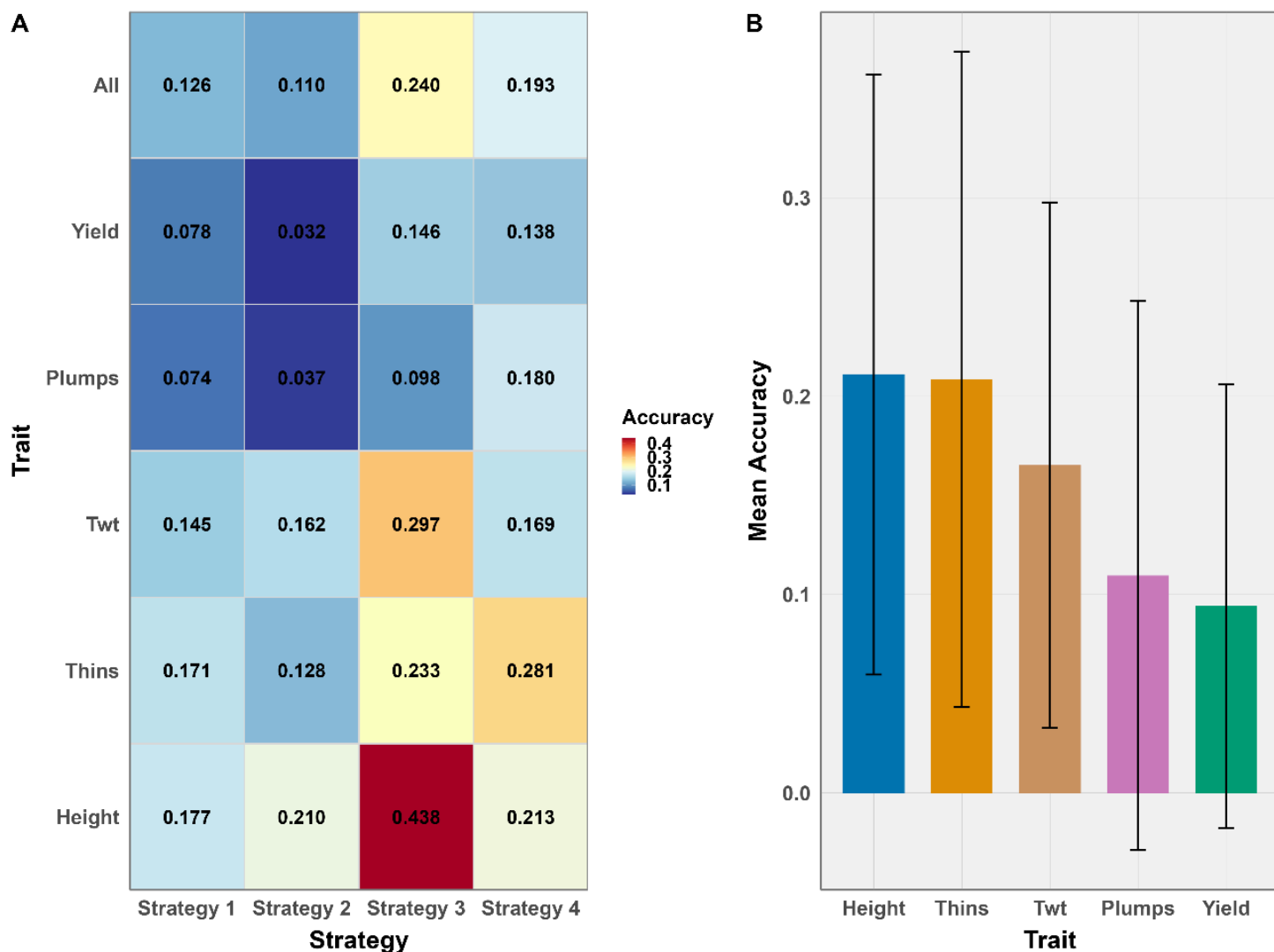


Figure 5

Trait performance across validation strategies. (A) Heatmap of mean PA for each trait × strategy. The “All” row reports the mean across traits within each strategy. (B) Trait-wise mean ± standard deviation accuracy across strategies.

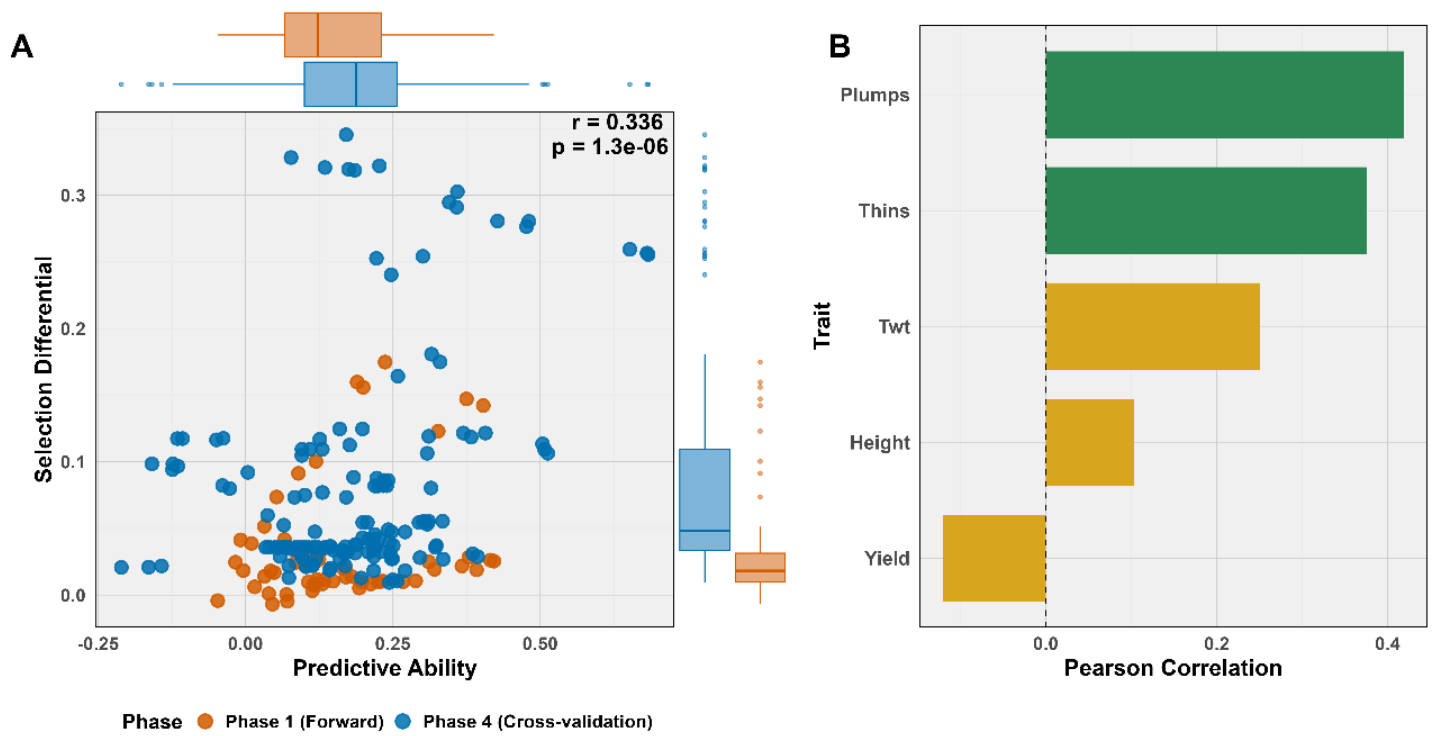


Figure 6

Accuracy-Selection differential Analysis. Relationship Between PA and Selection Differential. (A) Correlation between PA and selection differential across all prediction scenarios from Strategy 1 (orange) and Strategy 4 (blue). Marginal boxplots show distributions by strategy. (B) Trait-specific correlations. Green bars: $p < 0.05$; gold bars: $p \geq 0.05$.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supp.Tables.xlsx](#)
- [SupplementaryFiguresandFile.pdf](#)