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

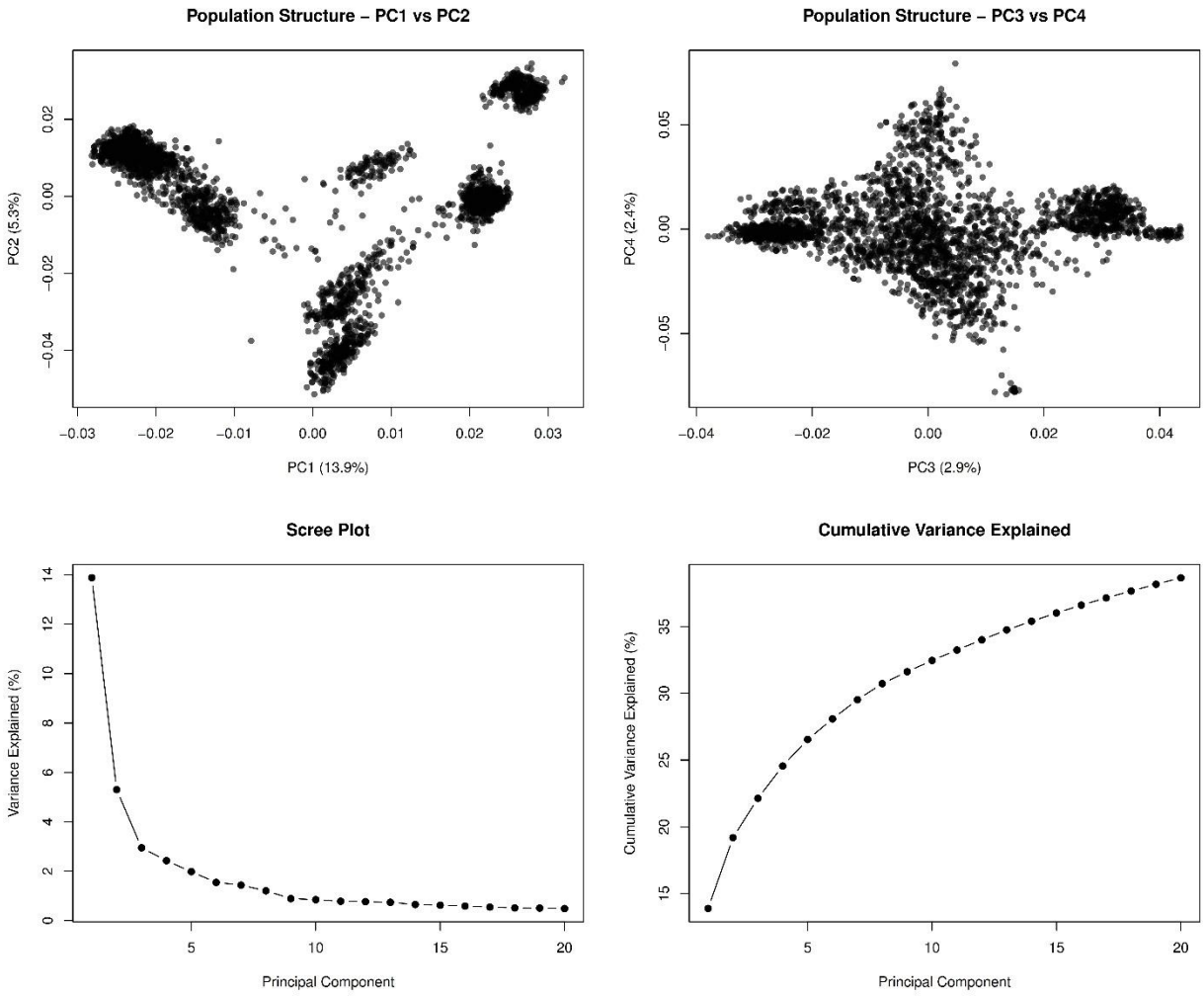
Muhammad Massub Tehseen[1], Charlotte Brault[2], Naa Korkoi Ardayfio[1], Sepehr Mohajeri Naraghi[1], Michael S. McMullen[1] and Jason D. Fiedler[3]*

[1]Department of Plant Sciences, North Dakota State University, Fargo, ND 58102, USA

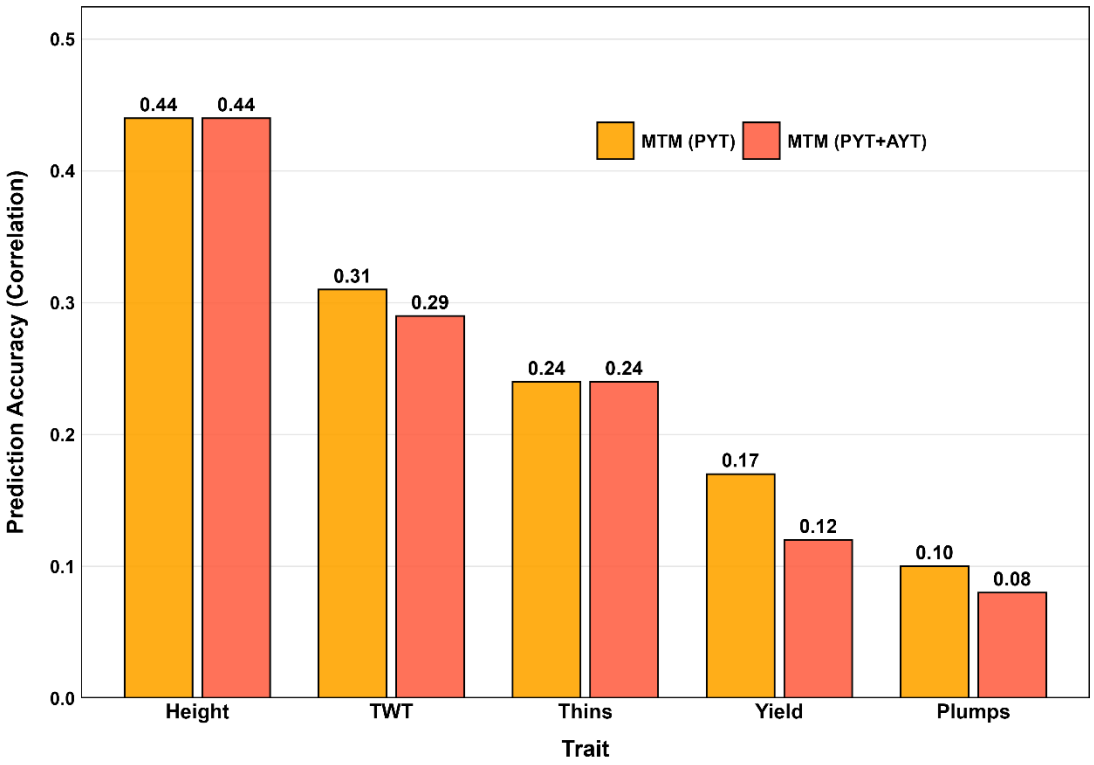
[2]Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108, USA

[3]USDA-ARS Cereals Crops Improvement Research Unit, Edward T. Schafer Agricultural Research Center, Fargo, ND 58102, USA

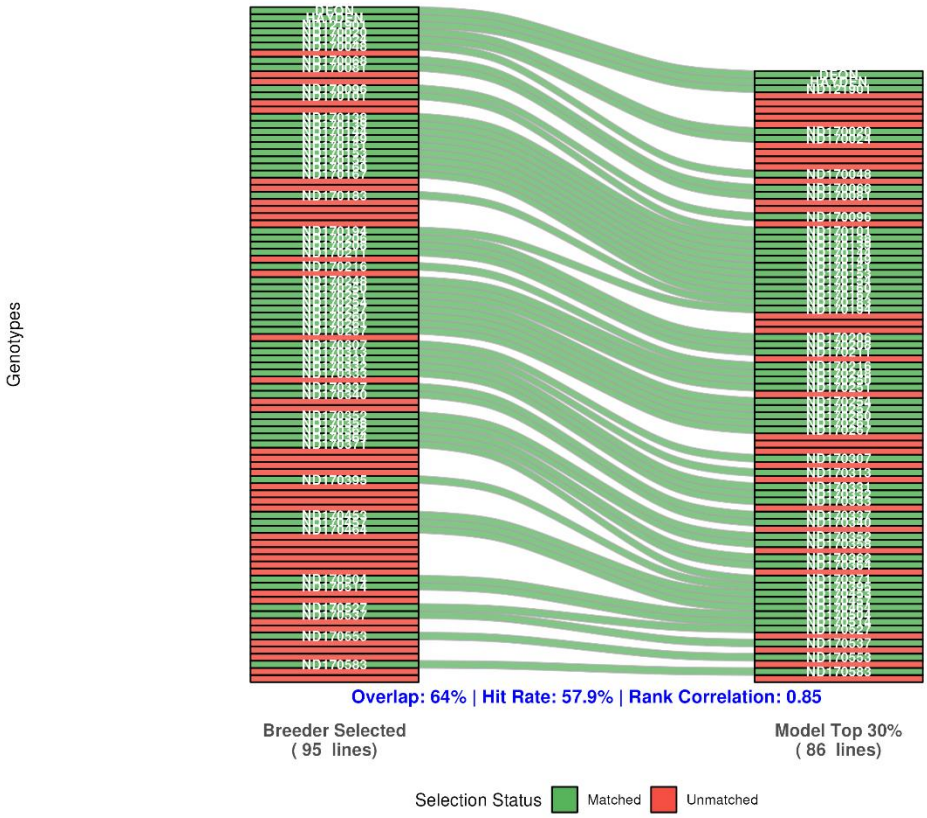
*Corresponding author Jason.fiedler@usda.gov



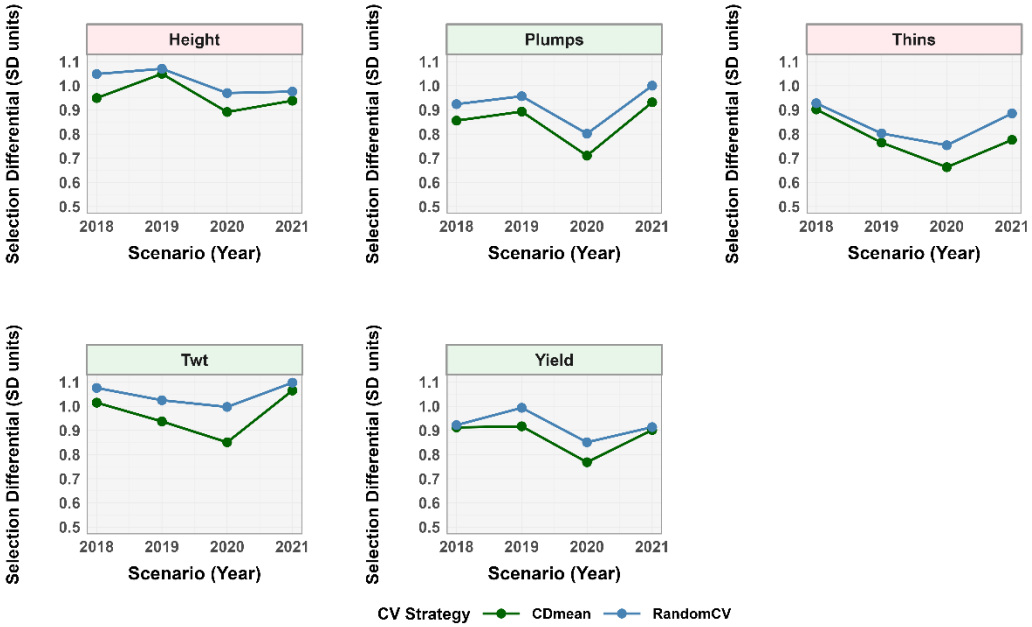
Supplementary Figure S1. *Principal component analysis of the genomic relationship matrix showing population structure across 1,941 oat breeding lines.*



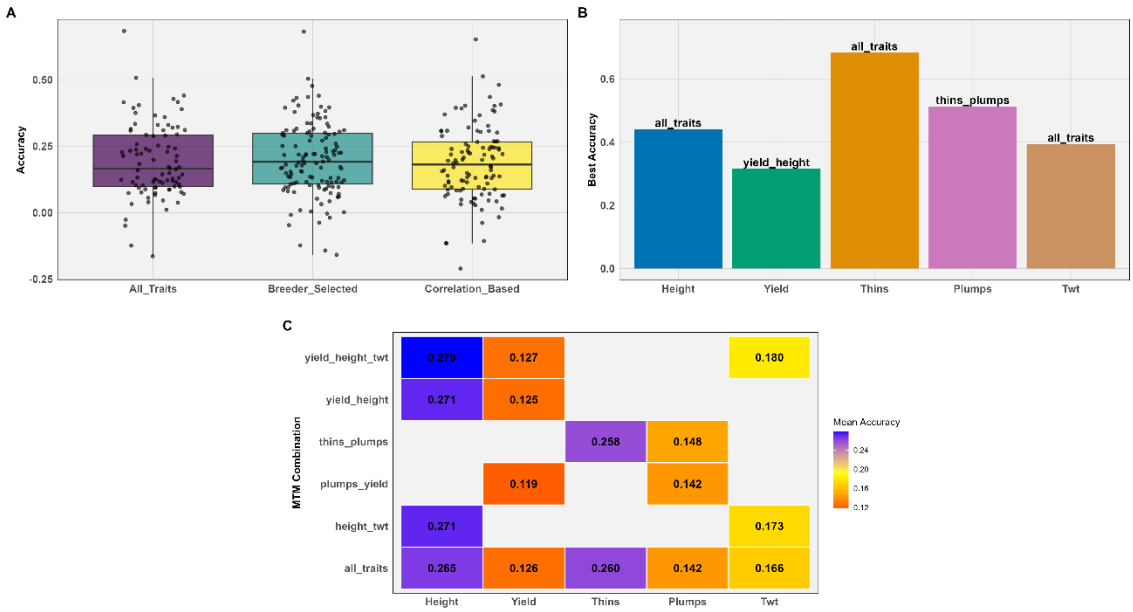
Supplementary Figure S2. *Effect of training data composition on multi-trait PA for 2021. PYT: preliminary yield trial; AYT: advanced yield trial.*



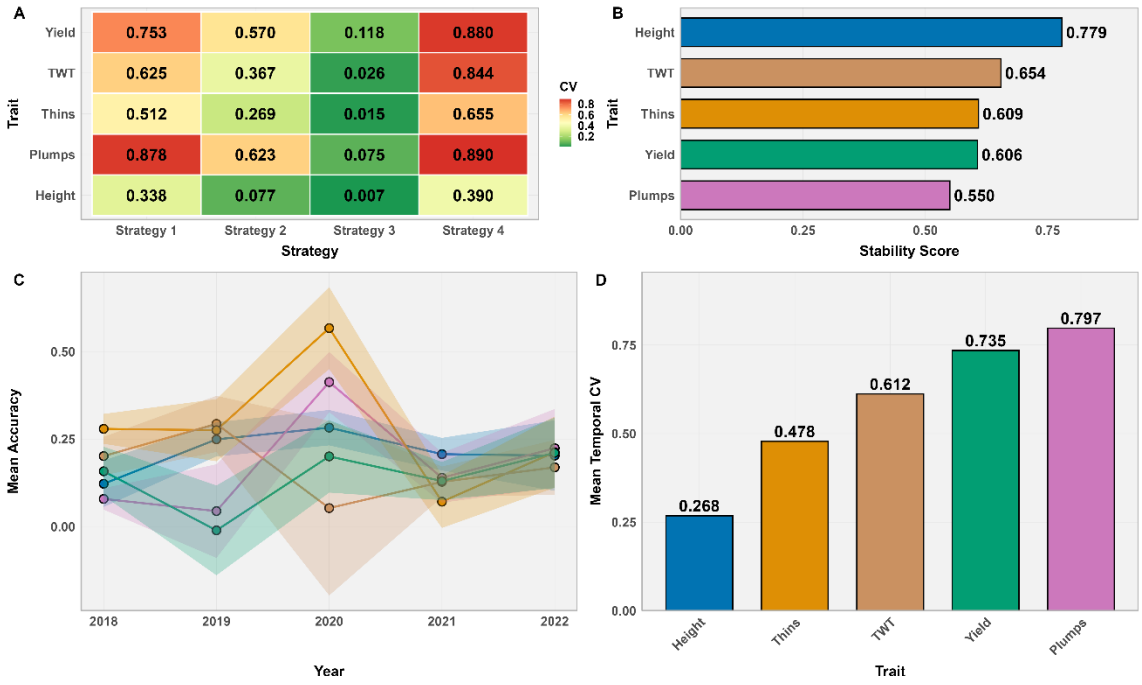
Supplementary Figure S3. *Selection efficiency comparison between breeder selections and model predictions for yield in 2018 under RandomCV (SM model). Sankey diagram showing the overlap between breeder-selected genotypes (left) and the top 30% of genotypes predicted by the model (right).*



Supplementary Figure S4. *Single-trait Selection Differential trends across prediction scenarios under different cross-validation strategies. Light coral strips indicate traits to minimize (Height, Thins); light green strips indicate traits to maximize (Plumps, Twt, Yield).*



Supplementary Figure S5. Multi-Trait Model (MTM) Trait Combinations Analysis. *(A) Performance comparison across MTM approaches, (B) best MTM combination by individual trait, (C) performance matrix of specific trait combinations.*



Supplementary Figure S6. Environmental and Temporal Stability Analysis of Genomic Prediction Models Across Five Oat Traits. *(A) Temporal stability by trait showing coefficients of variation across four validation strategies (lower CV = more stable), (B) environmental stability by trait with higher scores indicating greater stability (Strategy 4), (C) accuracy trends across years (2018-2022) for Strategy 4 showing temporal patterns, and (D) mean temporal stability by trait with lower values indicating greater stability (Strategy 1 - 4).*

Supplementary Methods S1: Phenotypic Data Processing

Stage 1: Within-Environment Spatial Adjustment

Spatial analysis was conducted separately for each trait within each environment (unique combination of year, location, field, and experiment number) using the SpATS package in R (Rodríguez-Álvarez et al., 2018). Data quality control was implemented by requiring at least 90% valid (non-missing) observations per trait-environment combination.

The spatial model incorporated both experimental design factors and spatial trends:

$$y = \mu + X_b \beta_b + Z_g g + Z_r r + Z_c c + f(row, column) + \varepsilon$$

where:

- y represents phenotypic observations
- μ is the overall mean
- β_b represents fixed effects for incomplete blocks
- $g \sim N(0, \sigma^2_g I)$ represents random genotype effects
- $r \sim N(0, \sigma^2_r I)$ and $c \sim N(0, \sigma^2_c I)$ represent random row and column effects to account for discrete spatial patterns
- $f(row, column)$ is a smooth bivariate P-spline surface (PSANOVA) capturing continuous spatial trends beyond the discrete row/column structure
- $\varepsilon \sim N(0, \sigma^2_\varepsilon I)$ represents residual error
- Design matrices X_b, Z_g, Z_r, and Z_c link observations to their respective effects

The discrete random effects (rows and columns) and the smooth spatial surface are identifiable because they target different spatial scales: the former captures discrete striping patterns across entire rows or columns, while the latter models residual continuous spatial gradients.

The effective dimension (ED) of spatial components was used to quantify the importance of spatial variation. The ED ratio was calculated as the ratio of the effective dimension of the spatial component to the total effective degrees of freedom in the model. Environments were classified as having:

- High spatial heterogeneity: ED ratio > 0.3
- Medium spatial heterogeneity: $0.1 <$ ED ratio ≤ 0.3
- Low spatial heterogeneity: ED ratio ≤ 0.1

Spatially adjusted genotype means were extracted as empirical best linear unbiased predictors (E-BLUPs) from the fitted models.

Stage 2: Multi-Environment Analysis

Spatially adjusted means from Stage 1 were combined across environments using mixed models fitted with the lme4 package:

Normalized ~ 1 + (1|Name) + (1|EnvID) + (1|EnvID:Name)

where:

- Observations were weighted by $1/SE^2$ (inverse squared standard errors from Stage 1)
- Name represents genotype
- EnvID represents environment
- The interaction term captures genotype-by-environment (G×E) effects

Genotype was modeled as a random effect to obtain BLUPs and estimate variance components. Simplified models without the G×E interaction term were fitted when singular fits occurred due to insufficient data structure (e.g., when genotypes were evaluated in very few environments).

Heritability Estimation

Broad-sense heritability (reliability) was calculated from variance components as:

$h^2 = \sigma^2\_g / (\sigma^2\_g + \sigma^2\_ge/n\_e + \sigma^2\_\varepsilon/n\_e)$

where:

- σ^2_g is the genotypic variance
- σ^2_ge is the genotype-by-environment interaction variance
- σ^2_ε is the residual variance
- n_e is the mean number of environments per genotype

This formulation accounts for both the direct genetic variance and the interaction effects scaled by environmental representation. Variance components were estimated using restricted maximum likelihood (REML). The resulting BLUPs served as response variables for all subsequent genomic prediction analyses.

Supplementary Methods S2: Genotypic Data and G-Matrix Construction

Genotyping-by-Sequencing (GBS) Library Preparation and Sequencing

Breeding lines (n=1,941) were genotyped using a dual-enzyme genotyping-by-sequencing (GBS) approach as described by Ardayfio et al. (2025). Approximately 2.5 million paired-end 2×150 bp reads per line were acquired on an Illumina sequencing platform.

Sequence Quality Control and Alignment

Raw sequence reads were quality filtered using bbduk from the BBTools suite (https://jgi.doe.gov/data-and-tools/bbtools) to remove adapters and low-quality bases. Filtered sequences were aligned to the *Avena sativa* OT3098 v2 reference genome (https://wheat.pw.usda.gov/jb?data=/ggds/oat-ot3098v2-pepsico) using Bowtie 2 (Langmead and Salzberg, 2012) with default parameters. Duplicate alignments were identified and marked using samblaster (Faust and Hall, 2014).

SNP Calling and Genotype Quality Control

Reads with a mapping quality score ≥30 were retained and processed with samtools/bcftools (Li, 2011) to identify single nucleotide polymorphisms (SNPs) and small insertions/deletions (InDels). Initial genotype calls underwent stringent quality control filtering using VCFtools with the following parameters:

- Minimum depth per sample (minDP): 2
- Maximum proportion of missing data (max-missing): 0.5 (i.e., ≥50% call rate required)
- Minimum minor allele frequency (maf): 0.01

Additional filtering was performed using TASSEL v5.2.50 (Bradbury et al., 2007) to remove SNP sites with >10% heterozygous calls, as oat is a predominantly self-pollinating species and high heterozygosity indicates potential genotyping errors or paralogs.

Missing Data Imputation

Missing genotype data were imputed using the LD-KNNi (linkage disequilibrium k-nearest neighbor imputation) method (Money et al., 2015) implemented in TASSEL. This approach uses linkage disequilibrium patterns among markers and genotypic similarity among lines to infer missing genotypes.

Final Marker Set Optimization

Following imputation, an additional filtering step removed:

- Taxa (breeding lines) with >75% missing data

- Taxa with >20% heterozygous calls

To reduce marker redundancy due to linkage disequilibrium, SNPs were thinned using PLINK v1.9 (Chang et al., 2015) with an r^2 threshold of 0.8 within 1 Mb sliding windows. When SNPs were in high LD ($r^2 > 0.8$), only one representative SNP was retained.

Final Marker Dataset

The final optimized marker dataset consisted of 5,126 SNPs distributed across the hexaploid oat genome.

Genomic Relationship Matrix Construction

The additive genomic relationship matrix (G) was constructed using the VanRaden method (VanRaden, 2008) implemented in the rrBLUP package (Endelman, 2011):

$$G = ZZ' / (2\Sigma p_i(1-p_i))$$

where:

- Z is the centered genotype matrix (coded as -1, 0, 1 for homozygous minor, heterozygous, and homozygous major genotypes, respectively, then centered by subtracting $2p_i$)
- p_i is the allele frequency at marker i
- The denominator $2\Sigma p_i(1-p_i)$ standardizes the matrix

This matrix represents the realized genomic relationships among breeding lines based on genome-wide marker information.

Population Structure Assessment

Principal component analysis (PCA) was performed on the G-matrix to assess population structure (Supplementary Figure S1). The first five principal components (PCs) were retained as fixed effects in all genomic prediction models to account for population stratification and avoid confounding between population structure and trait associations.

Supplementary Methods S3: Validation Strategy

Overview

A four-strategy validation framework was designed to address distinct practical questions relevant to implementing genomic selection in operational breeding programs. Each strategy tests different aspects of model performance and provides complementary information about prediction reliability under varying data conditions.

Data Leakage Prevention

To ensure temporal separation and avoid data leakage across validation strategies, trait-year-type-specific BLUPs (TYTs) were computed separately for each validation strategy. This was accomplished by including only the appropriate training years' data in the Stage 2 BLUP calculation (the multi-environment analysis step described in Supplementary Methods S1). Specifically:

- **Forward prediction**: Only data from preceding years were used to compute training set BLUPs
- **Backward prediction**: Only recent years' data (as specified per scenario) were used
- **Historical prediction**: All available prior observations per genotype were included to maximize training information
- **Cross-validation**: BLUPs were computed from randomly partitioned data, ensuring unbiased estimation through either random sampling or CDmean optimization

This approach ensures that predictions reflect realistic breeding scenarios where training and validation sets are truly independent.

Strategy 1: Forward Prediction

Purpose: Evaluate the ability to predict future breeding populations using historical data, mimicking the practical breeding scenario where models trained on past years are used to select among new breeding lines.

Design: Training data consisted of genotypes evaluated in years 2018-2021, with predictions made for new genotypes in subsequent years. Three temporal scenarios were tested:

1. **2018 → 2019**: Train on 2018 data, predict 2019 lines
2. **2018-2019 → 2020**: Train on 2018-2019 data, predict 2020 lines
3. **2018-2019-2020 → 2021**: Train on 2018-2020 data, predict 2021 lines

Interpretation: This strategy tests whether historical data can effectively predict performance of genetically distinct future populations, which is critical for operational implementation of genomic selection in early breeding stages.

Strategy 2: Backward Temporal Prediction

Purpose: Assess the effect of temporal proximity and training set size on predictive ability by varying the duration of the temporal window used to predict a fixed target year (2021).

Design: Three scenarios with decreasing temporal distance between training and validation were evaluated, all predicting 2021 performance:

1. **2018-2019-2020 → 2021**: Three-year training window (maximum temporal span)
2. **2019-2020 → 2021**: Two-year training window
3. **2020 → 2021**: One-year training window (minimum temporal distance)

Interpretation: This strategy tests whether training on a larger temporal window (more data but potentially less genetically relevant) improves predictive ability compared to training on only the most recent data (less data but more genetically similar to the validation population). It also assesses the effect of training population size when temporal proximity varies.

Strategy 3: Historical-to-Future Prediction

Purpose: Test the value of maintaining extensive datasets across trial types (PYT and AYT) for maximizing prediction accuracy when forecasting future performance.

Design: Two training population strategies were compared for predicting 2021 performance:

1. **PYT-only training**: Only preliminary yield trial (PYT) data from 2018-2020 were used
2. **PYT+AYT combined training**: Both PYT and advanced yield trial (AYT) data from 2018-2020 were combined to form a larger, more diverse training population

Interpretation: This strategy assesses whether including advanced trials (AYT), which represent more elite germplasm but smaller population sizes, improves prediction of future populations compared to using only preliminary trials. It addresses resource allocation questions about whether maintaining historical data across multiple trial stages is beneficial.

Strategy 4: Within-Year Cross-Validation

Purpose: Assess within-year predictive ability and compare two cross-validation approaches that represent different validation philosophies commonly employed in breeding programs.

Design: Within-year predictive ability was evaluated for each year (2018-2022) separately. Both approaches maintained 80/20 training-validation proportions but differed fundamentally in partitioning methodology:

Random Cross-Validation (RandomCV)

- Implemented as 5-fold cross-validation repeated 10 times
- In each repetition, genotypes were randomly partitioned into 5 folds
- Each fold served as the validation set once per repetition, with the remaining 4 folds as training
- Systematic partitioning ensured each genotype appeared in validation sets exactly 10 times (once per repetition)
- This yielded 50 total predictions per genotype (10 repetitions × 5 folds), ensuring equal representation across all individuals
- Represents an unbiased sampling approach designed to estimate average prediction accuracy

CDmean Optimization (CDmean)

- Implemented as 10 independent optimizations of training set composition
- In each optimization, 80% of genotypes were selected for training by maximizing the coefficient of determination (CD) criterion (Rincent et al., 2012)
- The CD criterion identifies representative training sets that capture population diversity and minimize prediction error variance
- The remaining 20% comprised the validation set for that optimization
- Unlike RandomCV's systematic sampling, CDmean optimization-based training set selection resulted in variable validation coverage across genotypes
- Validation frequency for each genotype was determined by its relationship to the optimized training sets rather than systematic assignment
- This yielded 10 total predictions per genotype
- Represents a training set optimization strategy used in operational breeding contexts to maximize prediction accuracy for a given training set size

Model Fitting For both approaches, genomic estimated breeding values (GEBVs) were obtained through repeated model fitting:

- Each iteration used genomic best linear unbiased prediction (GBLUP) as described in Supplementary Methods S3
- Relationship-based kriging was employed to predict all genotypes in each iteration
- Final GEBVs for each genotype were calculated as the mean across all predictions (50 for RandomCV, 10 for CDmean)

Interpretation: This strategy compares unbiased accuracy estimation (RandomCV) with training set optimization (CDmean) to determine whether sophisticated training set selection improves prediction performance in practice. It also establishes baseline within-year prediction accuracy for comparison with temporal prediction strategies (Strategies 1-3).

Supplementary Methods S4: Advanced Analytical Framework

Overview

To provide a comprehensive evaluation of genomic selection performance beyond simple predictive ability, four complementary analyses were conducted using standardized methodological approaches. Each analysis addresses a distinct aspect of model performance relevant to operational breeding program implementation.

Analysis 1: Trait Performance Across Validation Strategies

Purpose: Provide an integrated assessment of how individual traits perform across different validation contexts.

Methods: Mean predictive ability (PA), hit rate, and standardized selection differential were summarized across all four validation strategies (Strategies 1-4 described in Supplementary Methods S4) using multi-trait model outputs as the primary comparison basis. Multi-trait models were selected because they leverage cross-trait genetic covariance structures and provide the most representative assessment of trait performance across validation contexts.

For each trait, performance metrics were averaged across:

- All temporal prediction scenarios (Strategies 1, 2, and 3)
- All within-year cross-validation iterations (Strategy 4)
- Both cross-validation approaches (RandomCV and CDmean in Strategy 4)

Interpretation: This analysis identifies which traits are most reliably predictable and which yield the greatest selection gains, informing breeding program prioritization decisions.

Analysis 2: Multi-Trait Model Strategy Comparison

Purpose: Identify optimal trait-combination approaches for multi-trait genomic prediction by comparing three fundamentally different strategies.

Methods: Three multi-trait modeling strategies were implemented throughout all validation analyses (Strategies 1-4) and subsequently compared:

1. All_Traits Strategy

- Configuration: All five traits (yield, height, test weight, plumps, thins) included in each multi-trait model
- Rationale: Maximizes information utilization by leveraging all genetic correlations simultaneously
- Trade-off: Increased model complexity and computational demands

2. Breeder_Selected Strategy

- Configuration: Domain knowledge-informed trait combinations based on breeding program priorities:
    - Yield + Height + Test weight (comprehensive agronomic performance)
    - Yield + Height (key agronomic traits)
    - Thins + Plumps (grain quality traits)
- Rationale: Reflects how breeders conceptualize trait relationships based on biological understanding
- Trade-off: May miss empirically strong but non-intuitive correlations

3. Correlation_Based Strategy

- Configuration: Traits grouped based on observed genetic correlation strength:
    - Thins + Plumps (strong negative correlation)
    - Height + Test weight (moderate correlation)
    - Plumps + Yield (strong positive correlation)
- Rationale: Data-driven approach leveraging empirically observed genetic architecture
- Trade-off: May combine traits with limited breeding relevance

Evaluation: Relative performance of these three strategies was evaluated across all four validation strategies to determine which approach maximized PA overall and for individual traits. This provides practical guidance for operational multi-trait model implementation—whether breeders should rely on domain knowledge, empirical correlations, or comprehensive models.

Analysis 3: Prediction Stability Metrics

Purpose: Assess model consistency and robustness beyond average accuracy, as stable predictions are critical for operational breeding decisions.

Methods: Two complementary stability metrics were calculated:

Temporal Stability

- **Definition**: Quantifies prediction consistency across validation scenarios within each validation strategy

- **Calculation**: Coefficient of variation (CV) of predictive ability across scenarios per trait: CV_temporal = SD(PA) / mean(PA) where PA values are from different scenarios within the same strategy (e.g., for Strategy 1: scenarios 2018→2019, 2018-2019→2020, 2018-2019-2020→2021)
- **Interpretation**: Lower CV indicates methodologically robust models maintaining consistent performance regardless of validation framework. High CV suggests prediction quality is sensitive to specific data partitioning choices.

Environmental Stability

- **Definition**: Assesses consistency across crop years (2018-2022), where each year represents distinct environmental conditions
- **Data source**: Strategy 4 (within-year cross-validation) exclusively, as this provides PA estimates for each year
- **Calculation**: S_env = 1 / (1 + CV_year) where CV_year = SD(PA_year) / mean(PA_year) across the five years
- **Interpretation**: Higher scores (approaching 1.0) indicate greater resilience to inter-annual environmental variation. Low scores suggest genotype-by-environment interactions substantially impact prediction quality.

Rationale: These metrics address the practical question: "Will this model perform consistently across different breeding cycles and environmental conditions?" Average PA may be misleading if performance varies drastically across contexts.

Analysis 4: Predictive Ability and Selection Differential Association

Purpose: Evaluate whether higher predictive accuracy translates to superior parent selection and genetic gain, testing the fundamental assumption that improving PA leads to better breeding outcomes.

Methods: The association between PA and realized selection differential was assessed across prediction scenarios. For each unique prediction scenario (defined by trait × year × trial type × model approach × validation strategy combination):

1. **Selection differential calculation**: Mean observed phenotypic value (BLUP) of the top 30% highest-predicted individuals minus the population mean
2. **Correlation analysis**: Pearson correlations between PA and selection differential were computed across prediction scenarios from:
    - Strategy 1 (forward prediction): Tests relationship under realistic temporal prediction
    - Strategy 4 (within-year cross-validation): Tests relationship under ideal within-population conditions

3. **Trait-specific analysis**: Correlations were calculated separately for each trait to assess whether the PA-selection differential relationship varied by trait architecture (e.g., highly heritable vs. low heritability traits)

Interpretation: Strong positive correlations indicate that improving PA reliably translates to capturing superior genotypes. Weak or inconsistent correlations suggest that PA alone may be insufficient for evaluating model utility, supporting the study's "beyond predictive ability" framework. Trait-specific differences reveal whether certain traits benefit more from accuracy improvements than others.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Burgueño, J., Crossa, J., Cotes, J. M., San Vicente, F., & Das, B. (2011). Prediction assessment of linear mixed models for multienvironment environments. Crop Science, 51(3), 944-954. https://doi.org/10.2135/cropsci2010.07.0403

Pérez, P., & de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. Genetics, 198(2), 483-495. https://doi.org/10.1534/genetics.114.164442