Multi-objective Evaluation and Optimization of Stochastic Gradient Boosting Machines for Genomic Prediction and Selection in Wheat (*Triticum aestivum*) Breeding

Henry Munroe[1], Bright Osatohanmwen[1,3], Reza Sharifi*[2,3]

[1] University of Göttingen, Division of Plant Breeding Methodology, Department of Crop Sciences Carl-Sprengel-Weg 1, 37075 Göttingen
[2] University of Göttingen, Division of Animal Breeding and Genetics, Albrecht-Thaer-Weg 3, 37075 Göttingen
[3]Center for Integrated Breeding Research, University of Göttingen, Albrecht-Thaer-Weg 3, Göttingen, 37075, Germany

*Corresponding author:
Reza Sharifi rsharif@gwdg.de

Running title: GBM Optimization for Wheat Genomic Prediction

Keywords: Genomic prediction; Gradient boosting machine; Hyperparameter optimization; Model Stochasticity; Cross-environment prediction; Wheat (*Triticum aestivum*) MAGIC population

Abstract

Machine learning (ML) models with stochastic and non-deterministic characteristics are increasingly used for genomic prediction in plant breeding, but evaluation often neglects important aspects like prediction stability and ranking performance. This study addresses this gap by evaluating how two hyperparameters of a Gradient Boosting Machine (GBM), learning rate (v) and boosting rounds (ntrees), impact stability and multi-objective predictive performance for cross-season, cross-environment prediction in a MAGIC wheat population. Using a grid search of 36 parameter combinations, we evaluated four agronomic traits with five metrics: Pearson's r, Area Under the Curve (AUC), Normalized Discounted Cumulative Gain (NDCG), and the Intraclass Correlation Coefficient (ICC) and Fleiss' κ for stability. Our findings show that a low learning rate combined with a high number of boosting rounds substantially improves prediction stability (ICC > 0.98) and selection stability (Fleiss' κ > 0.80), while reducing train-test performance gaps. This combination produced concurrent improvements for predictive accuracy (r) and ranking efficiency (NDCG), though optimal settings were trait-dependent. Conversely, classification accuracy (AUC) was poor and performed relatively better with higher learning rates, revealing a conflict in optimization hyperparameters. Despite moderate Pearson's r and poor AUC in this challenging cross-season, cross-environment prediction scenario, NDCG remained high (> 0.85), indicating strong ability to rank top-performing entries. Ultimately, prioritizing stability when tuning GBMs effectively yields reproducible cross-environment predictions with improved accuracy and top-end ranking performance.

Introduction

As machine learning (ML) becomes more prevalent in plant breeding, breeders face increasing demands from large genomic and phenotypic datasets, computational time constraints and pressing market and environmental challenges (Pieruschka and Schurr 2019; van Dijk et al. 2020; Calvin et al. 2023; Yoosefzadeh Najafabadi et al. 2023; Crossa et al. 2025). The effectiveness of ML models depends on how they are constructed, optimized, and evaluated (Domingos 2012; Feurer and Hutter 2019). Model evaluation often relies on a limited range of metrics emphasizing overall predictive accuracy, via Pearson's r (Meuwissen et al. 2001; Goddard and Hayes 2007). Although overall predictive accuracy is fundamental to assessing model performance, relying solely on accuracy metrics to evaluate ranking efficiency, selection efficiency, and stability, is inherently limited; this creates an evaluation gap and underscores the need for targeted metrics (Blondel et al. 2015; Dekkers et al. 2021; Lange et al. 2025).

A multi-metric evaluation is therefore justified to address this gap. Ranking ability can be assessed with Normalized Discounted Cumulative Gain (NDCG), which rewards the correct ordering of top-performing candidates (Järvelin and Kekäläinen 2002; Blondel et al. 2015). For binary "select/reject" decisions, the Area Under the Curve (AUC) measures a model's ability to distinguish between classes (Huang and Ling 2005; Robin et al. 2011). Yet, neither metric directly addresses the issue of non-determinism, or stochasticity, inherent in many algorithms.

In stochastic Gradient Boosting Machines (GBM), random subsampling introduces variability between runs (Friedman 2002). This requires explicit stability assessment. The Intraclass

Correlation Coefficient (ICC) can quantify the consistency of continuous predictions across iterations (Bartko 1966; McGraw and Wong 1996; Weir 2005; Koo and Li 2016; Lange et al. 2025). A high ICC value indicates that variability in predictions is mainly due to inherent differences among targets rather than random fluctuations, demonstrating stable model performance (Weir 2005). For categorical selections, Fleiss kappa (κ) measures the stability of agreement among "candidate" classifications across runs (Fleiss 1971; Lange et al. 2025).

GBM performance is influenced by hyperparameters that balance the bias-variance trade-off. The learning rate and the number of boosting rounds must be carefully tuned to capture data complexity without overfitting (Friedman 2001; Ridgeway and Developers 2003). This creates a multi-objective optimization problem: maximizing predictive, selection and ranking accuracy while ensuring high stability.

This study performs a parametric optimization of a stochastic GBM to address these challenges. We systematically evaluate 36 combinations of learning rate and boosting rounds using genomic and phenotypic data from a Multi-parent Advanced Generation Inter-Cross (MAGIC) wheat population. Cross-season predictions are assessed on four traits-Grain Yield (GY), Grain Protein Content (GPC), Thousand Grain Weight (TGW), and Height to Ear tip (HET). We aim to: (1) evaluate performance using Pearson's r, NDCG, and AUC; (2) quantify prediction and selection stability with ICC and Fleiss κ; (3) determine the effects of hyperparameters on all metrics; and (4) provide insights into model generalizability. Findings will help breeders select ML models that are not only accurate, but also stable enough to support confident, repeatable selection decisions.

Methods and Materials

A 500-line wheat (*Triticum aestivum*) Multi-parent Advanced Generation Inter-Cross (MAGIC) population was analyzed (Scott et al. 2021). Four agronomic traits were assessed from their study: GY, TGW, GPC, HET in a cross-season prediction scenario. Field experiments were conducted at Girton, UK (52.24144 °N, 0.08794 °E; clay-loam soil; 2016-17 season, training) and Whittlesford, UK (52.09909 °N, 0.13311 °E; loam soil; 2017-18 season, testing) (Scott et al. 2021).

SNP genotyping and quality control

Genotypes comprised 25,508 single-nucleotide polymorphisms (SNPs) from the Axiom 35k wheat breeders' SNP genotyping array (Winfield et al. 2016). Loci with > 5 % missing calls were excluded. Remaining missing values were imputed by Random Forest using the missRanger R package (v2.5.0, see File S1 - Package_Functions) (Mayer 2024).

Train-test splits

To evaluate performance in a cross-environment prediction scenario, models were trained using data from the 2016/17 season and evaluated against observed phenotypes from the 2017/18 season. Five random 80%/20% partitions of the 500 lines were generated using set.seed(123).

Gradient-boosting

Stochastic gradient-boosting machine was fitted with gbm v2.2.2 in R 4.4.3. A 6 × 6 tune grid, set manually with; learning rate = {0.0025, 0.042, 0.0815, 0.121, 0.1605, 0.20} and boosting rounds = {100, 1080, 2060, 3040, 4020, 5000}, while all other arguments remained at default

(including bag.fraction = 0.5) (File S1 - Package_Functions). (DOI: 10.5281/zenodo.16785772 - Data Availability).

The ensemble update at boost round m followed (Friedman 2002; Ridgeway and Developers 2003):

$$F_m(x) = F_{m-1}(x) + \nu \rho_m h(x; a_m)$$

Where:

$F_m(x)$ is the updated ensemble prediction

$F_{m-1}(x)$ is the prediction from the previous iteration

ρ_m is the optimal step length determined via line search

$h(x: a_m)$ is a regression tree fit to the pseudo residuals

ν is the learning rate (shrinkage, $0 < \nu \leq 1$)

Prediction pipeline

For each of the five train-test splits, each of the ten stochastic model fits was independently used to generate predictions on its corresponding 20% hold-out set. This yielded 50 distinct prediction vectors (5 splits × 10 iterations per split) for each hyperparameter combination, which were then compared to the observed phenotypes. Reproducibility ensured via set.seed(123 + split_number × 1000 + prediction_iteration) (DOI: 10.5281/zenodo.16785772 - Data Availability). See Fig. 1 for the schematic diagram of the pipeline.

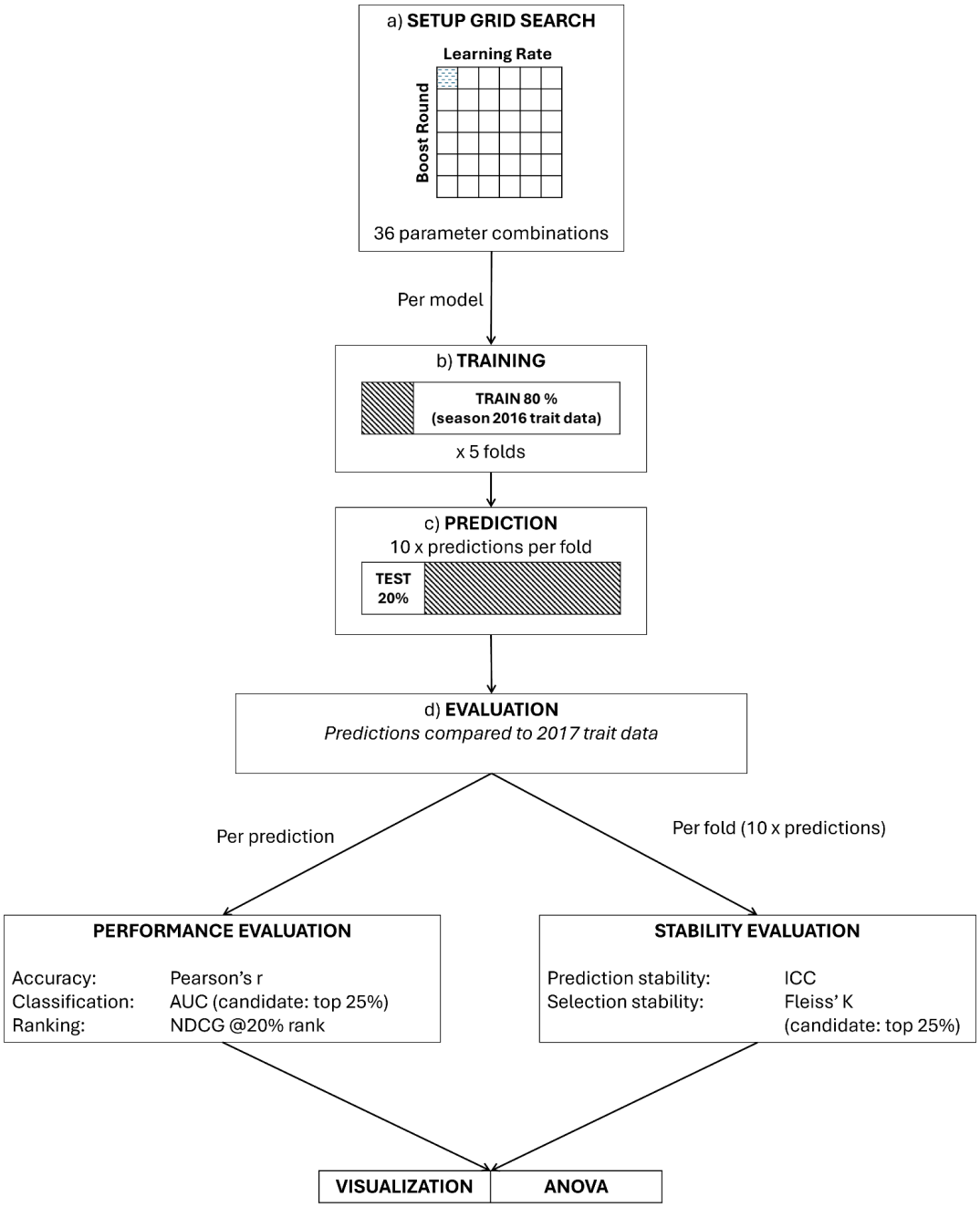


Figure 1. shows the pipeline (per trait): (a) grid search setup, where 36 tuning parameter combinations (6 learning rates x 6 boosting rounds) are specified; (b) training, where each combination is trained on 80 % of entries using the 2016 trait data with five-fold cross-validation; (c) prediction, in which ten iterations of predictions are produced per fold on the hold out 20 % test set; and (d) evaluation, where predictions are compared against 2017 trait values and divided into performance and stability evaluation performance metrics (Pearson's r, AUC for the top 25 % candidates, and NDCG at the top 20 % rank) are calculated for each prediction, and stability (Intraclass correlation coefficient and Fleiss' κ) is assessed per fold. All metrics are then plotted and analyzed via type 3 ANOVA.

Performance and stability metrics

Performance metrics, Accuracy - Pearson's r, ranking efficiency - Normalized Discounted Cumulative Gain for the top 20% of predicted lines (NDCG@20%), and classification ability - by labeling lines with observed phenotypes above the 75th percentile as Candidates and then computing the ROC AUC (File S1 - Formulae). Stability across the ten stochastic repeats was quantified by the single-measure, one-way intraclass correlation coefficient (ICC) for predictions and by Fleiss K for overall agreement of binary selection (File S1 - Formulae, DOI: 10.5281/zenodo.16785772 - Data Availability). Across the hyperparameter grid, each trait yielded n = 1,800 performance estimates (r, NDCG@20 %, AUC; 5 splits × 6 learning-rates × 6 boost-rounds 10 iterations) and n = 180 stability estimates (ICC, Fleiss K; 5 splits × 6 learning-rates × 6 boost-rounds) (File S4).

Data visualization

Figures were generated with ggplot2 3.4 and viridis 0.6. Error bars are the standard error (± SE) which are found in the table of summary statistics (mean, SD, SE, min, max, IQR) for every metric and hyperparameter combination in File S2. Min-max normalization of performance and stability metrics were applied and plotted as supplementary material (Fig. S2).

Statistical analysis

All statistics were run in R 4.4.3 to quantify how GBM hyperparameters influence predictive performance and stability within and across traits. Q-Q and residual-versus-fitted plots were made for observing normality of residuals (Fig. S4). Effect sizes were calculated as partial η^2 (η^2_p) for all Type III ANOVA Satterthwaite-approximated F-tests and as Kendall's W for Friedman tests, using the effectsize R package (v1.01).

Modelling framework

Trait-wise predictive performance was analyzed with random-intercept linear mixed models (LMM; lme4 v1.1-36), using learning rate and boosting rounds as fixed effects and random intercepts to account for Subset and Subset:Iteration nesting. Trait-wise stability was assessed non-parametrically via Friedman tests. To examine hyperparameter effects across traits, we fitted three-way LMMs (Trait × Learning Rate × Boost Rounds) with random intercepts for Trait and Subset (and, for performance only, Subset:Iteration). Cross-trait (aggregated) performance and stability were evaluated with two-way LMMs (Learning Rate × Boost Rounds), where performance models again included Subset:Iteration, but stability models retained only Trait and Subset nesting. Fixed effects were tested by, Type III, Satterthwaite-approximated F-tests (lmerTest v3.1-3), Tukey-Kramer (Tukey-adjusted pairwise) tests were made via emmeans v1.11.0. All p-values reported are raw unless otherwise specified, with post-hoc contrasts presented as Tukey-adjusted p-values. (DOI: 10.5281/zenodo.16785772 - Data Availability, File S3).

Results

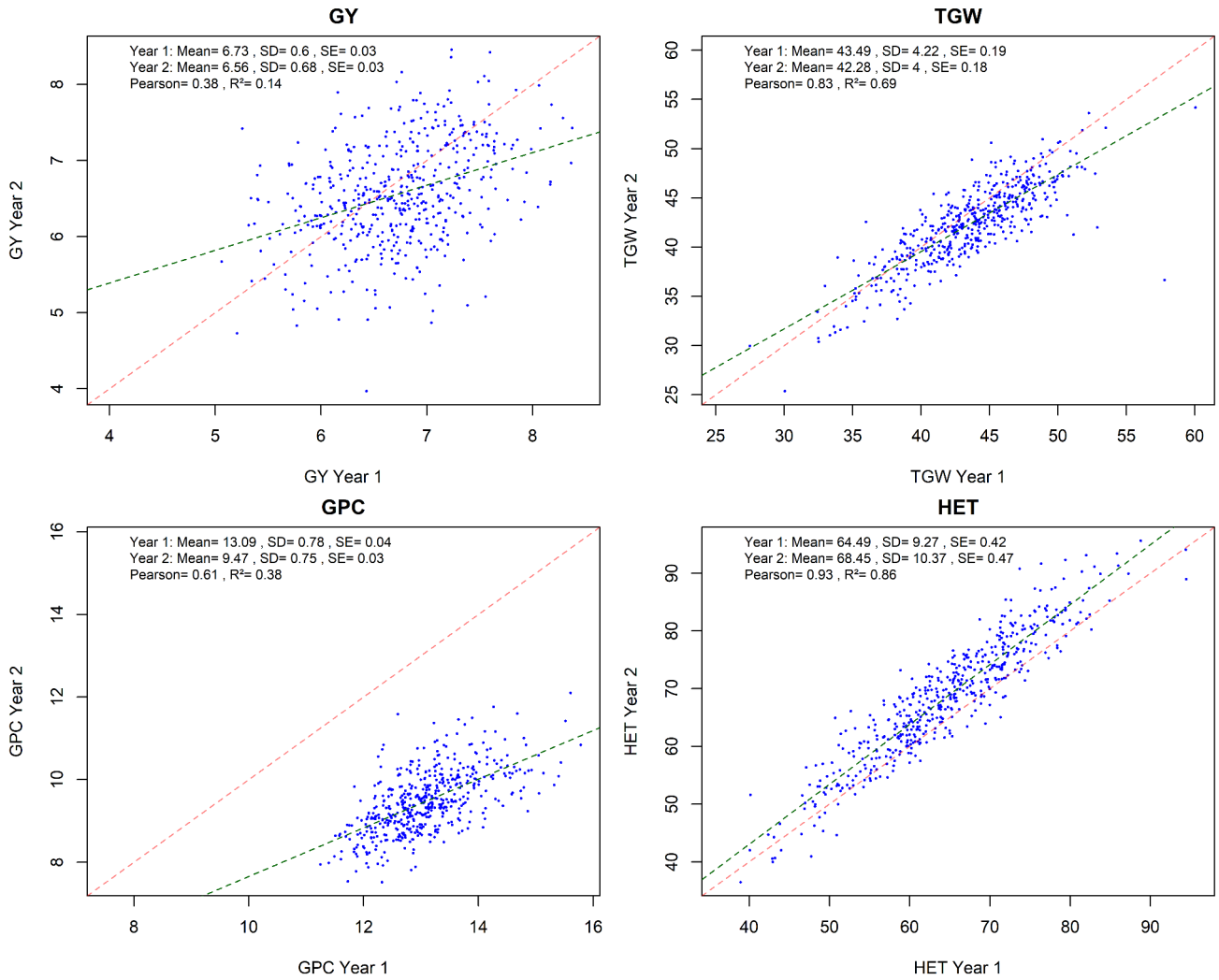


Figure 2. This figure displays four scatter plots of observed measurements, shows the relationship between Year 1 (x-axis) and Year 2 (y-axis) for four agronomic traits from the wheat dataset. Traits: GY (grain yield) - top left, TGW (thousand grain weight) - top right, GPC (grain protein content) - bottom left and HET (height to ear tip) - bottom right. Each blue point represents an individual wheat line's measurements in both years. The red diagonal line is the 45° (theoretical) perfect agreement between the two years. The green dashed line is the best-fit regression line. The annotations in each panel summarize the mean, standard deviation (SD), and standard error (SE) for both years, Pearson's r and R^2.

Analysis of raw agronomic data of the 500 entries across two seasons found that some traits were more stable year-to-year than others. TGW and HET showed strong Pearson's r between years (r = 0.83 and r = 0.93, respectively). In contrast, GY showed a weak cross-year correlation (r = 0.38), whereas GPC was moderate (r = 0.61). The mean values for GPC also varied largely between the first year (13.09% SE = 0.03) and the second year (9.47% SE = 0.03), unlike other traits (Fig. 2).

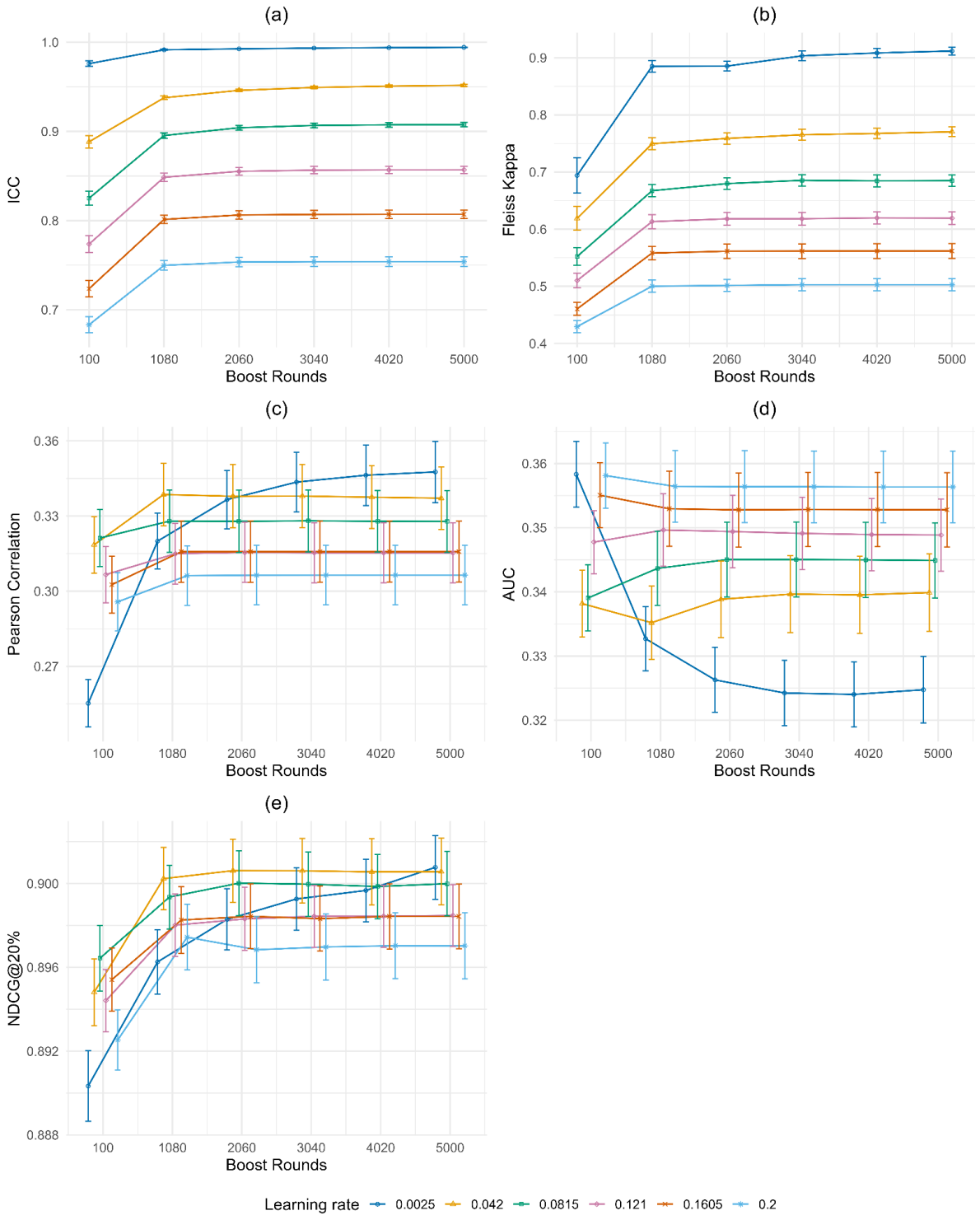


Figure 3. Displays trait aggregated values for five performance metrics: Intra-class Correlation Coefficient (ICC, top left), Fleiss' κ (top right), Pearson's r (center left), Area Under the Curve (AUC, center right), and Normalized Discounted Cumulative Gain at the top 20% ranking (NDCG@20%, bottom left). Each plot shows aggregated results across all four traits, with the x-axis representing boosting rounds and the y-axis indicating the respective evaluation metric. Colored lines connect the mean metric values for each learning rate, and error bars reflect ±1 standard error. This figure provides a view of model performance across all metrics. Each color corresponds to a distinct Learning Rate; Dark Blue (0.0025), Orange (0.042), Green (0.0815), Purple (0.121), Red-Orange (0.1605), and Light Blue (0.2).

To determine the optimal hyperparameters for genomic prediction, we evaluated the effects of learning rate and boosting rounds on overall model performance and stability by analyzing metrics aggregated across all four traits. Elapsed time was not affected by learning rate but increased linearly with boosting rounds: models with 100 rounds finished in under 30 minutes per core, while 5,000 boost rounds required ≈ 10 hours per core (Fig. S3).

The aggregated performance of the 36 hyperparameter combinations across all traits is presented in Figure 3. Models combining low learning rates (v = 0.0025 to 0.0815) with moderate-to-high boosting rounds (> 2000) achieved the highest ICC, Fleiss' κ, Pearson's r and NDCG@20% (Fig. 3a-c, 3e). Predictive accuracy (Pearson's r) and ranking performance (NDCG@20 %) showed moderate improvements, while substantial improvements in stability metrics (ICC and Fleiss' κ) were observed (Fig. 3a-c, 3e). A Type III, three-way ANOVA, with fixed-effect significance determined by Satterthwaite-approximated F-tests, showed that the interaction of trait × learning rate × boosting rounds was a significant source of variation for all three performance metrics: Pearson's r ($F_{(75, 6860)} = 5.06$, $\eta^2_p = 0.052$, $p < 0.001$), AUC ($F_{(75, 6860)} = 1.95$, $\eta^2_p = 0.020$, $p < 0.001$), and NDCG@20% ($F_{(75, 6860)} = 2.60$, $\eta^2_p = 0.028$, $p < 0.001$). This indicates that optimal hyperparameter settings were trait-dependent. The two-way interaction between learning rate × boosting rounds was also significant across all three metrics: Pearson's r ($F_{(25, 6860)} = 22.18$, $\eta^2_p = 0.075$, $p < 0.001$), AUC ($F_{(25, 6860)} = 8.26$, $\eta^2_p = 0.029$, $p < 0.001$), NDCG@20% ($F_{(25, 6860)} = 1.71$, $\eta^2_p = 0.006$, $p < 0.001$) (File S3 - ANOVA 3-way).

Pairwise comparisons of the estimated marginal means (Tukey-adjusted) for the aggregated results identified two distinct performance trends. For Pearson's r and NDCG@20%, models with lower learning rates (v = 0.0025 to 0.0815) combined with a moderate-to-high number of

boosting rounds (ntrees > 2000) significantly outperformed models with higher learning rates (v > 0.1) ($p < 0.001$) (Fig. 3c, 3e). Within the low learning-rate group, at ≥ 2000 boosting rounds, the estimated marginal means for Pearson's r and NDCG did not differ between v = 0.0025 and v = 0.042 (e.g Pearson's r: $\Delta = -0.00134$, SE = 0.004, $t(6965) = -0.323$, $p > 0.99$) (File S3 - Pairwise_overallmodel_Adj_Pears). In contrast, classification accuracy (AUC) found models with low boosting rounds (ntrees = 100) and higher learning rates (v > 0.01) consistently and significantly outperformed those with lower learning rates (v < 0.01) ($p < 0.001$) (Fig. 3d, File S3 - Pairwise_overallmodel_Adj_AUC).

The type III three-way ANOVA showed that hyperparameters, learning rate and boosting rounds had significant main effects on all performance metrics (r, AUC, NDCG@20%) although the magnitude of their effects varied by metric (File S3 - ANOVA 3-way). For Pearson's r, both the learning rate and boosting rounds showed moderate, statistically significant effects: learning rate ($F_{(5, 6860)} = 97.28$, $\eta^2_p = 0.066$, $p < 0.001$) and boosting rounds ($F_{(5, 6860)} = 78.00$, $\eta^2_p = 0.054$, $p < 0.001$). For classification performance (AUC) the learning rate was the main driver ($F_{(5, 6860)} = 128.73$, $\eta^2_p = 0.086$, $p < 0.001$) with boosting rounds provided significant but low effect ($F_{(5, 6860)} = 5.55$, $\eta^2_p = 0.004$, $p < 0.001$). Ranking performance (NDCG@20%) showed boosting rounds as the main driver ($F_{(5, 6860)} = 38.74$, $\eta^2_p = 0.027$, $p < 0.001$) compared to learning rate ($F_{(5, 6860)} = 14.23$, $\eta^2_p = 0.010$, $p < 0.001$) (File S3 - ANOVA 3-way).

The effect of hyperparameter tuning for model stability showed substantial improvements (Figure 3a, 3b). A Type III, three-way ANOVA showed that the trait × learning rate × boosting rounds interaction was not significant for either stability metric (ICC: $F_{(75, 560)} = 0.59$, $\eta^2_p = 0.073$,

$p = 0.99$, Fleiss' κ: $F_{(75, 560)} = 1.21$, $\eta^2_p = 0.140$, $p = 0.11$), indicating consistent stability tuning effects across traits. Learning rate made the greatest contribution to stability (ICC: $F_{(5, 560)} = 7304.27$, $\eta^2_p = 0.985$, $p < 0.001$, Fleiss' κ: $F_{(5, 560)} = 1793.30$, $\eta^2_p = 0.941$, $p < 0.001$) followed by boosting rounds with substantial secondary effects (ICC: $F_{(5, 560)} = 602.76$, $\eta^2_p = 0.843$, $p < 0.001$, Fleiss' κ: $F_{(5, 560)} = 257.11$, $\eta^2_p = 0.697$, $p < 0.001$) (Table S3 - ANOVA 3-way). Significant interaction of learning rate × boosting rounds confirm that joint, co-optimization provide large effects in improvements to stability (ICC: $F_{(25, 560)} = 15.15$, $\eta^2_p = 0.404$, $p < 0.001$, Fleiss' κ: $F_{(25, 560)} = 5.86$, $\eta^2_p = 0.207$, $p < 0.001$) (File S3 - ANOVA 3-way).

Pairwise comparisons of the estimated marginal means (Tukey-adjusted) confirmed a consistent trend: lower learning rates produced significantly higher stability scores for both ICC and Fleiss' κ, with nearly all learning rate combinations being significantly different from one another when compared at the same boosting round level. For example, at 100 boosting rounds, the ICC score with a learning rate of 0.0025 was significantly higher than a rate of 0.042 (Mean Difference = 0.088, SE = 0.0056, $t(677) = 15.76$, $p < 0.001$). Full pairwise results are provided in File S3.

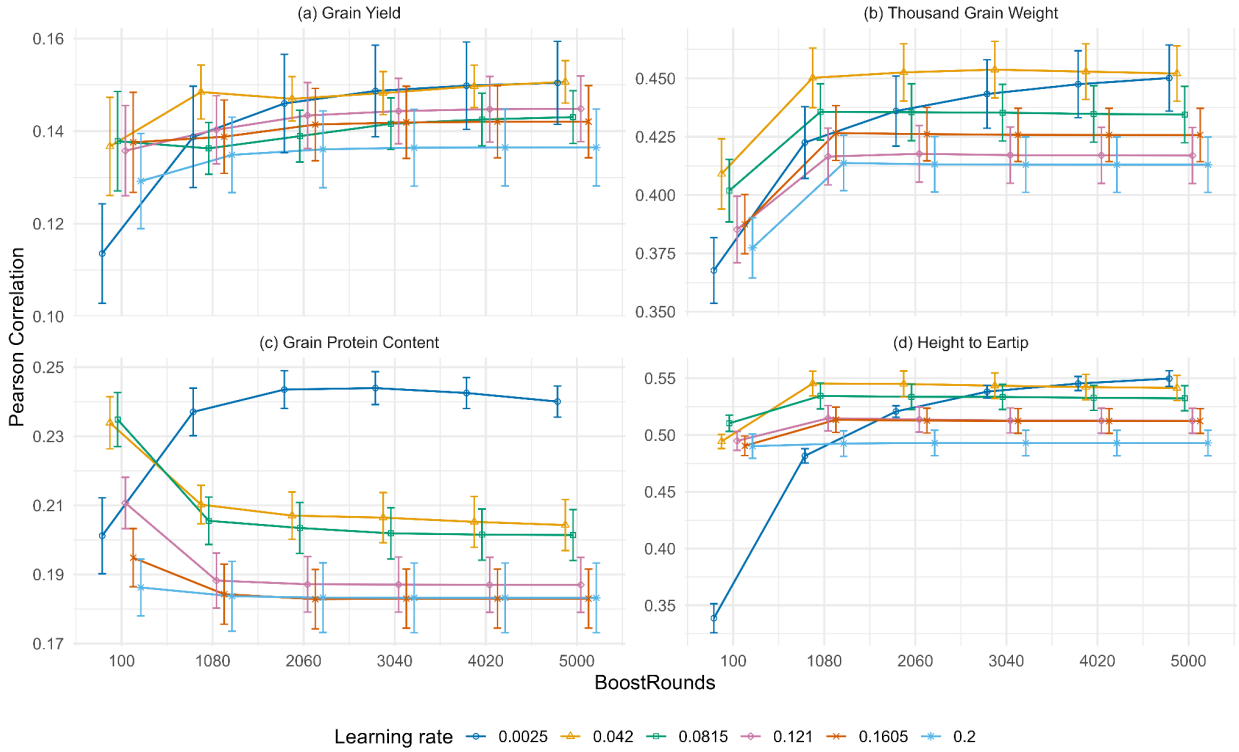


Figure 4. Displays the Pearson's r coefficient varies with boosting rounds and learning rate across four traits Grain Yield (GY - top left), Thousand Grain Weight (TGW - top right), Grain Protein Content (GPC - bottom left) , Height to Ear Tip (HET - bottom right). The x-axis represents boosting rounds and the y-axis indicates the mean Pearson's r for each model combination, with error bars showing ±1 standard error. Coloured lines connect mean values for each learning rate. Each color corresponds to a distinct Learning Rate; Dark Blue (0.0025), Orange (0.042), Green (0.0815), Purple (0.121), Red-Orange (0.1605), and Light Blue (0.2).

Model accuracy performance via Pearson's r varied by trait. GPC showed the greatest increase in Pearson's r at an intermediate number of boosting rounds (ntrees > 2000) when using learning rate v = 0.0025 (Fig. 4c). The mean Pearson's r across hyperparameter combinations was 0.42 (SD = 0.091) for TGW and 0.51 (SD = 0.080) for HET (Fig. 4a-d, File S5). A Type III, two-way ANOVA for each trait confirmed that hyperparameter responses differed markedly among traits (File S3 - ANOVA Trait Specific 2-way). For GY, both learning rate ($F_{(5, 1715)} = 2.34$, $\eta^2_p = 0.007$, $p < 0.05$) and boosting rounds ($F_{(5, 1715)} = 3.78$, $\eta^2_p = 0.011$, $p < 0.01$) had small yet significant main effects, whereas their interaction was not significant ($F = 0.63$, $\eta^2_p = 0.009$, $p = 0.92$). In TGW, both hyperparameters showed large, highly significant effects for learning rate (F

= 61.39, $\eta^2_p = 0.152$, $p < 0.001$) and boosting rounds (F = 98.87, $\eta^2_p = 0.224$, $p < 0.001$) and the learning rate × boosting rounds interaction also contributed low but significant effect (F = 2.98, $\eta^2_p = 0.042$, $p < 0.001$). Pearson's r for GPC was driven primarily by learning rate (F = 58.94, $\eta^2_p = 0.147$, $p < 0.001$) and boosting rounds showed a low but significant effect (F = 2.36, $\eta^2_p = 0.007$, $p < 0.05$), and the interaction remained significant with low effect (F = 3.07, $\eta^2_p = 0.043$, $p < 0.001$). HET displayed the strongest response to hyperparameters with large, significant main effects of learning rate (F = 77.71, $\eta^2_p = 0.185$, $p < 0.001$) and boosting rounds (F = 113.42, $\eta^2_p = 0.249$, $p < 0.001$), and a strong interaction effect (F = 38.82, $\eta^2_p = 0.361$, $p < 0.001$). Full trait-specific ANOVA tables are provided in File S3 - ANOVA Trait-Specific 2-Way.

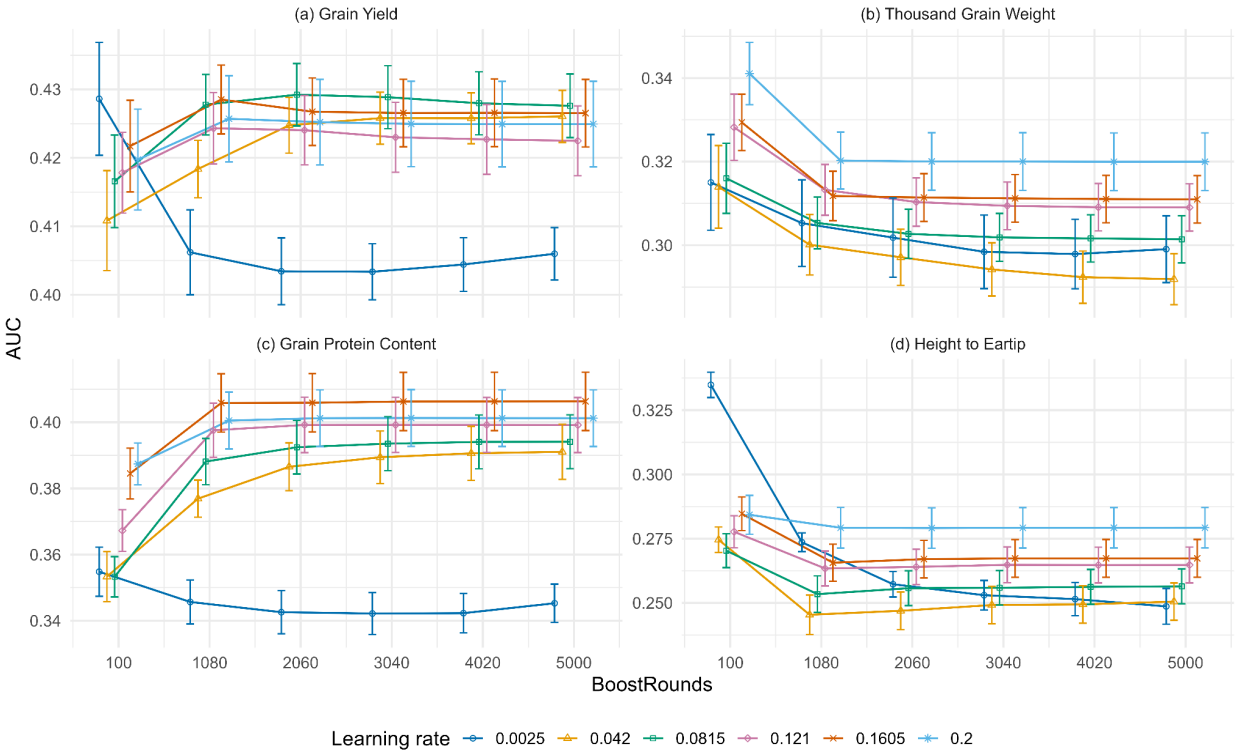


Figure 5. Displays Area Under the Curve (AUC) by boosting rounds across learning rate across four traits Grain Yield (GY - top left), Thousand Grain Weight (TGW - top right), Grain Protein Content (GPC - bottom left) , Height to Ear Tip (HET - bottom right). The x-axis represents boosting rounds and the y-axis indicates the mean AUC for each model combination, with error bars showing ±1 standard error. Colored lines connect mean values for each learning rate. Data are presented in separate panels for each trait with free y-axis scales. Each color corresponds to a distinct Learning Rate; Dark Blue (0.0025), Orange (0.042), Green (0.0815), Purple (0.121), Red-Orange (0.1605), and Light Blue (0.2).

Classification accuracy, measured by Area Under the Curve (AUC), was poor across all hyperparameter combinations, with values below the 0.5 threshold of random chance for all four traits (Fig. 5a-d). Visually, models with higher learning rates ($v \geq 0.042$) resulted in higher AUC values, often peaking with 1000 - 2000 boosting rounds. In contrast, the lowest learning rate ($v = 0.0025$) consistently produced lower AUC values across all boosting rounds. With a learning rate of $v = 0.0025$, AUC decreased during moderate-to-high numbers of boosting rounds but began to increase from 4,000 boost rounds.

ANOVA for each trait confirmed that AUC responses to hyperparameters varied among traits (File S3 - ANOVA Trait-Specific 2-Way). For GY, learning rate showed a significant effect ($F_{(5, 1715)} = 16.41$, $\eta^2_p = 0.046$, $p < 0.001$), boosting rounds had no independent influence ($F_{(5, 1715)} = 0.53$, $\eta^2 p = 0.002$, $p = 0.755$), and the interaction was significant but small ($F_{(25, 1715)} = 2.08$, $\eta^2_p = 0.030$, $p < 0.01$). In TGW, both learning rate ($F_{(5, 1715)} = 41.65$, $\eta^2_p = 0.108$, $p < 0.001$) and boosting rounds ($F = 25.56$, $\eta^2_p = 0.070$, $p < 0.001$) had moderate effects, whereas the interaction was not significant ($F = 0.24$, $\eta^2_p = 0.004$, $p > 0.99$). For GPC, learning rate had the largest main effect across traits ($F_{(5, 1715)} = 111.83$, $\eta^2_p = 0.246$, $p < 0.001$), boosting rounds also contributed ($F = 20.31$, $\eta^2_p = 0.056$, $p < 0.001$), and a small but significant interaction was detected ($F = 2.66$, $\eta^2_p = 0.037$, $p < 0.001$). HET displayed large, significant effects for learning rate ($F_{(5, 1715)} = 40.22$, $\eta^2_p = 0.105$, $p < 0.001$) and boosting rounds ($F = 48.09$, $\eta^2_p = 0.123$, $p < 0.001$), alongside a substantial interaction ($F_{(25, 1715)} = 9.47$, $\eta^2_p = 0.121$, $p < 0.001$). Full ANOVA results are available in File S3 - ANOVA Trait-Specific 2-Way.

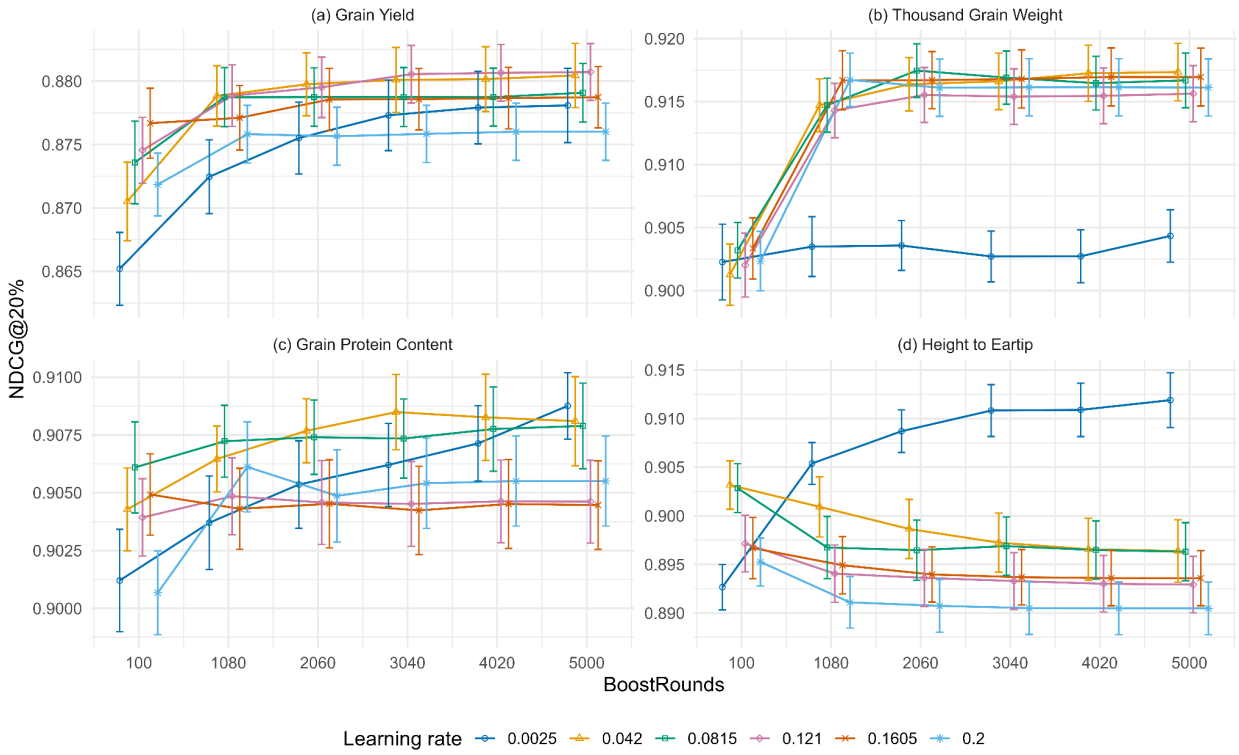


Figure 6. The plot displays Normalized Discounted Cumulative Gain (NDCG) evaluated at the cutoff corresponding to the top 20% position of ranked entries (denoted, NDCG@20%) compared to different combinations of boosting rounds and learning rate across four traits Grain Yield (GY - top left), Thousand Grain Weight (TGW - top right), Grain Protein Content (GPC - bottom left) , Height to Ear Tip (HET - bottom right). The x-axis shows boosting rounds and the y-axis indicates the mean NDCG@20% for each model combination, with error bars showing ±1 standard error. Colored lines connect mean values for each learning rate. Data are presented in separate panels for each trait with free y-axis scales. Each color corresponds to a distinct Learning Rate; Dark Blue (0.0025), Orange (0.042), Green (0.0815), Purple (0.121), Red-Orange (0.1605), and Light Blue (0.2).

Ranking performance, measured by NDCG@20%, was high across all traits, with values consistently above 0.85 (Fig. 6, File S2, S4). Figure 6 shows that for the lowest learning rate (v = 0.0025), NDCG@20% values increased across all boosting rounds. In contrast, for higher learning rates, performance typically peaked with 1000-3000 boosting rounds before plateauing or declining (Fig. 6a-d).

ANOVA for each trait confirmed that NDCG@20 % responses to hyperparameters varied among traits (File S3 - ANOVA Trait-Specific 2-Way). For GY, learning rate ($F_{(5, 1715)} = 6.97$, $\eta^2_p =$

0.020, $p < 0.001$) and boosting rounds ($F_{(5, 1715)} = 13.02$, $\eta^2_p = 0.037$, $p < 0.001$) had small but significant effects, while their interaction was not significant ($F_{(25, 1715)} = 0.84$, $\eta^2_p = 0.012$, $p = 0.70$). In TGW, both learning rate ($F = 50.75$, $\eta^2_p = 0.129$, $p < 0.001$) and boosting rounds ($F = 59.49$, $\eta^2_p = 0.148$, $p < 0.001$) showed substantial effects, and a modest interaction was present ($F = 2.12$, $\eta^2_p = 0.030$, $p < 0.001$). For GPC, learning rate ($F = 6.05$, $\eta^2_p = 0.017$, $p < 0.001$) and boosting rounds ($F = 4.14$, $\eta^2_p = 0.012$, $p < 0.001$) had small but significant effects, whereas the interaction was not significant ($F = 0.82$, $\eta^2_p = 0.012$, $p = 0.72$). HET differed, showing a strong main effect of learning rate ($F = 62.35$, $\eta^2_p = 0.154$, $p < 0.001$), no main effect of boosting rounds ($F = 0.35$, $\eta^2_p = 0.001$, $p = 0.88$), and a moderate interaction ($F = 5.27$, $\eta^2_p = 0.071$, $p < 0.001$). Full ANOVA results are provided in File S3 - ANOVA Trait-Specific 2-Way.

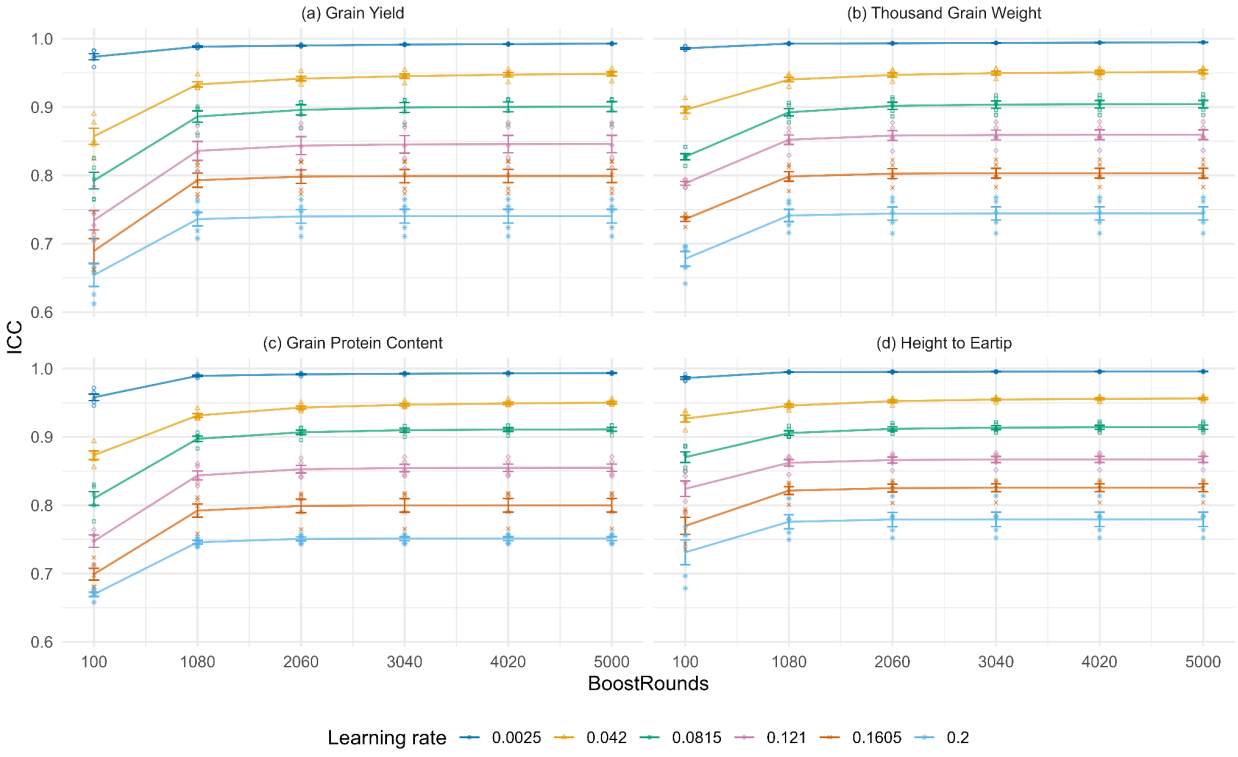


Figure 7. The plot displays the intra class correlation coefficient (ICC) compared to combinations of boost rounds and learning rate across four traits Grain Yield (GY - top left), Thousand Grain Weight (TGW - top right), Grain Protein Content (GPC - bottom left) , Height to Ear Tip (HET - bottom right). The x-axis represents boost rounds and the y-axis shows ICC. For each trait and learning rate, mean ICC values are connected by dashed lines with error bars indicating ±1 standard error. Individual ICC observations appear as points. Each color corresponds to a distinct Learning Rate; Dark Blue (0.0025), Orange (0.042), Green (0.0815), Purple (0.121), Red-Orange (0.1605), and Light Blue (0.2).

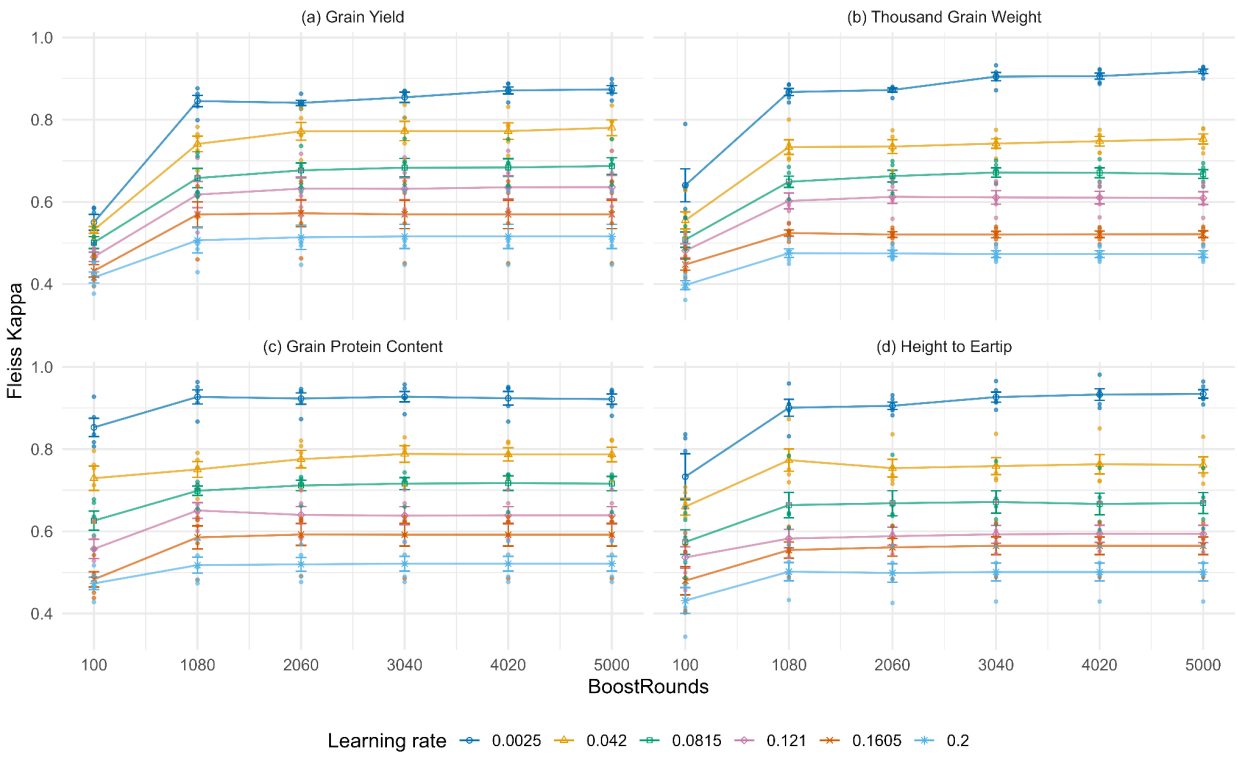


Figure 8. The plot displays the Fleiss' κ compared to combinations of boost rounds and learning rate across four traits Grain Yield (GY - top left), Thousand Grain Weight (TGW - top right), Grain Protein Content (GPC - bottom left) , Height to Ear Tip (HET - bottom right). The x-axis represents boost rounds and the y-axis shows Fleiss' κ. For each trait and learning rate, mean Fleiss' κ values are connected by dashed lines with error bars indicating ±1 standard error. Individual Fleiss' κ observations appear as points. Each color corresponds to a distinct Learning Rate; Dark Blue (0.0025), Orange (0.042), Green (0.0815), Purple (0.121), Red-Orange (0.1605), and Light Blue (0.2).

Prediction stability (ICC) and selection stability (Fleiss Kappa) showed highly consistent and clear trends across all four traits (Fig. 7, 8). For both metrics, stability values systematically increased as the learning rate decreased. The lowest learning rate (ν = 0.0025) consistently resulted in the highest stability, with ICC values approaching 1.0 for all traits at 5,000 boost rounds (M = 0.994, SD = 0.001, N = 20; File S4, S2). For any given learning rate, stability also increased with more boosting rounds, typically plateauing after approximately 2,000 rounds (Fig. 7, 8). Friedman tests confirmed significant differences among the 36 hyperparameter combinations for both stability metrics across all traits. For ICC, $\chi^2(35) = 172.75 - 174.47$, $p < 0.001$, with a strong level of agreement among the rankings (Kendall's $W = 0.987 - 0.997$). For

Fleiss' κ, $\chi^2(35) = 160.48 - 168.10$, $p < 0.001$, with very strong agreement (Kendall's $W = 0.917 - 0.961$). Full trait-specific tables are provided in File S3 - Trait Specific Friedman Stab.

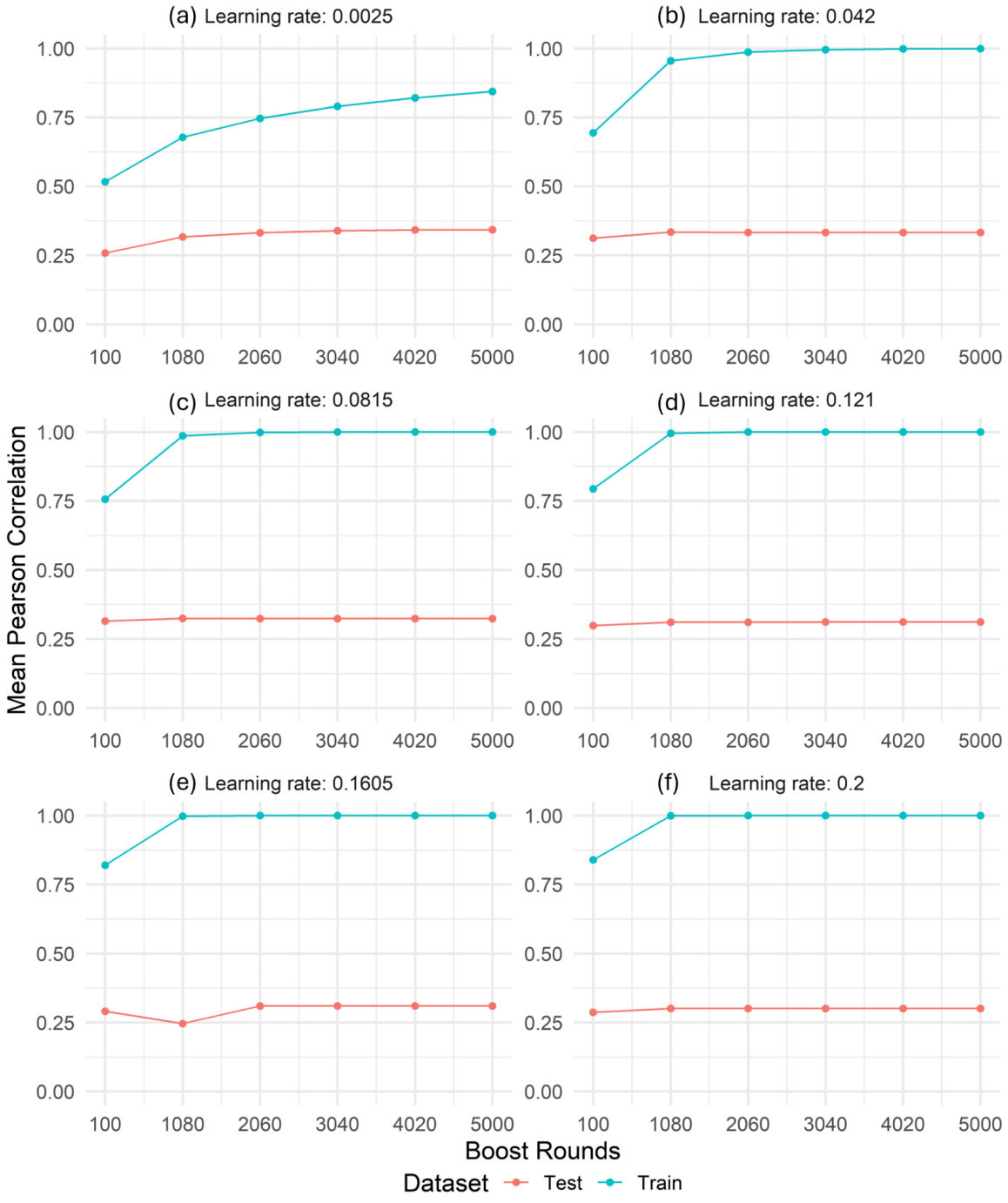


Figure 9. Shows the relationship between correlation coefficient and boosting rounds across different learning rates with traits aggregated. The y-axis shows the mean Pearson's r coefficient. The x-axis shows the number of boosting rounds. Each facet represents a different learning rate. In each plot, the blue line depicts model performance on the training data, while the orange-red line shows performance on unseen test data, allowing the visualization of overfitting.

Model generalizability was assessed by comparing the mean Pearson's r on the training and test datasets for each learning rate across boosting rounds (Figure 9a-f). A performance gap was observed for all hyperparameter combinations, with training set correlations consistently higher than test set correlations. For learning rates of 0.042 and higher, training performance rapidly approached a correlation of 1.0, creating a large gap relative to the test performance (Fig. 9b-f). In contrast, for the lowest learning rate (v = 0.0025), training performance increased more gradually, resulting in the smallest performance gap between the training and test datasets among all tested learning rates (Fig. 9a).

Discussion

The stochastic characteristic of many modern genomic prediction models produces stochastic outputs; evaluating them solely on predictive accuracy ignores this critical instability and undermines their reliability for supporting selection and ranking decisions in breeding programs. Our multi-objective hyperparameter search tuning for model stability (via ICC and Fleiss' κ) is a highly effective strategy to obtain reproducible predictions and is associated with smaller train - test gaps. A GBM configuration with the lowest learning rate (v = 0.0025) and > 2,000 boosting rounds maximized stability (Fig. 3a-b,7,8) and provided moderate but significant concurrent improvements in Pearson's r and ranking (NDCG) performance (Fig. 3c, 3e); pairwise contrasts indicate that very low learning rates

(0.0025 - 0.042) perform comparably for r and NDCG at higher rounds (Fig. 4,6, File S3). While model output stability (ICC, Fleiss' κ) was consistent across traits, performance metrics (Pearson's r, AUC and NDCG) showed trait dependent responses, as indicated by significant trait × learning rate × boosting rounds interactions for Pearson's r, AUC and NDCG in the aggregated analysis. This complexity supports the use of a multi-objective approach, as no single optimization strategy can meet all the diverse objectives within breeding programs.

Low learning rates increased stability to excellent levels (e.g, mean ICC ≈ 0.994 at 5,000 boosting rounds across traits), supporting consistent predictions (Fig. 3a, 7). Low learning rates also increased selection stability with Fleiss' κ frequently exceeding 0.8 (Fig. 3b, 8). To interpret these coefficients, we adopt a unified framework. Following Shrout, who notes conceptual links between chance-corrected agreement statistics such as κ and reliability coefficients like the ICC, we apply a single set of modern benchmarks to both measures (Shrout 1998). Accordingly, we use the thresholds recommended by Koo and Li, defining reliability as good for values between 0.75 and 0.90 and excellent for values greater than 0.90 (Koo and Li 2016). Under this framework, both results indicate good to excellent stability, implying highly reliable model predictions across continuous and categorical outputs.

Aggregated analysis identified a practical hyperparameter window: the lowest learning rate (v = 0.0025) maximized stability, while low learning rates up to v = 0.042, when combined with more than 2,000 boost rounds delivered comparable Pearson's r and NDCG performance (Fig. 3, File S3 - Pairwise_overallmodel_Adj_NDCG and Pairwise_overallmodel_Adj_Pears). This suggests moderately boosted models with low learning rates can achieve similar performance for Pearsons's r and NDCG with substantially less computation (Friedman 2002) (Fig. S3). Within

this near optimal window, the learning rate of 0.0025 with 5,000 boosting rounds achieved the highest combined stability, predictive accuracy, and ranking efficiency (Friedman 2002).

In contrast to other metrics, classification accuracy (AUC) revealed a conflict in optimization objectives: it not only performed worse than random chance across all models and traits but also favored a different hyperparameter configuration, where fewer boosting rounds and higher learning rates yielded better relative discriminatory ability. Since the total AUC metric evaluates performance across all decision thresholds, it can obscure a model's utility for breeders by giving undue weight to irrelevant portions of the ROC curve (Bradley 1997; Peng et al. 2024). We therefore support that a partial AUC (pAUC), by restricting the analysis to a relevant window of high True Positive and low-to-moderate False Positive Rates (TPR, FPR), should be explored for a more relevant metric for gauging the discriminatory power of models for candidate selection (McClish 1989; Peng et al. 2024). A pAUC can adjust the FPR tolerance to match the decision-making context. For example, in pre-breeding where a common aim is to maximize diversity, breeders may tolerate higher FPR (e.g, > 0.2) to avoid prematurely discarding potentially elite candidates, whereas late stage multi-environment performance trials may have stricter FPR (e.g, FPR < 0.1) due to the cost per entry (Simmonds 1993). Our findings suggest that model optimization requires a strategic choice based on the primary objective, allowable tradeoffs and trait of interest, as a single set of hyperparameters could not deliver optimal performance across all evaluation metrics. The poor absolute values for AUC are likely a consequence of the challenging cross-environment scenario where candidate assignment was based on the top 25% of observed phenotypes within the test environment, creating a 'moving target' and, due to this fixed upper threshold, resulting in class imbalance. The unmodeled

environmental covariates were likely an important factor in the poor cross-environment prediction. The absence of these covariates inhibited the modeling of G × E interactions, which can degrade the cross-season performance (Jarquín 2017; Tomar et al. 2021; Montesinos-López et al. 2022).

Despite the model's modest prediction and classification accuracy, its consistently high ranking performance (NDCG@20% > 0.85) indicates practical value for elite-line discovery, where correctly ranking top candidates is more valuable than precise estimation of their breeding values. This high ranking performance, in contrast to the lower Pearson's r values, is explained by the nature of the NDCG metric, which focuses specifically on rewarding the correct ordering of top-performing entries rather than penalizing errors across the entire distribution (Blondel et al. 2015; Pan et al. 2024). The suitability of this metric for breeding decisions, where correctly ranking and selecting elite lines is a primary goal, has led to its increasing adoption (Gianola et al. 2018; Negussie et al. 2022; Li et al. 2023; Pan et al. 2024).

The ANOVA results confirm that model performance is influenced mostly by main effects of learning rate and boosting rounds but also by their significant interaction, which supports the need for co-optimization. This hierarchy of effects is consistent with previous research where learning rate was a dominant hyperparameter for XGBoost model with a general support for lower learning rates (Westhues et al. 2021). The learning rate provides regularization that prevents overfitting to subsample artifacts by shrinking the contribution of each new boost round (Elith et al. 2008). The non-significant three-way interaction for prediction stability (ICC) is an important result as it indicates that the relationship between learning rate and boosting rounds

was consistent across all traits, suggesting that an optimized hyperparameter combination for stability may be broadly applicable across traits.

An important finding was the trait-specific response to hyperparameter tuning. While the overall pattern of the hyperparameters explored (favoring low learning rates and high boost rounds) was consistent, Grain Protein Content (GPC) displayed the clearest and most substantial gain in model accuracy (Pearson's r) when using the lowest learning rate (v = 0.0025), when combined with sufficient boosting rounds (ntrees > 1000). The findings for GPC show that small incremental error corrections in each boosting round may better capture the cumulative contribution of many small-effect loci, which are more effectively modelled through slower learning rates and high boosting rounds (Azodi et al. 2019; Sandhu et al. 2021; Paina and Gregersen 2023).

While our results demonstrate the benefits of GBM hyperparameter tuning for genomic prediction, we acknowledge the limitations of this study. First, the performance metrics, Pearson’s r and AUC, remained low-moderate in absolute terms across all quantitative traits. This poor performance is supported by a clear train-test gap, suggesting overfitting (Fig. 9). Lower learning rates were associated with smaller train-test gaps (Fig. 9). By shrinking each boost round’s contribution, lower learning rates reduce the influence of subsample artifacts introduced by stochastic subsampling and reduce noise amplification across iterations, improving generalization (Friedman 2001; Friedman 2002). This result aligns with established principles, confirming that learning rates are useful for maintaining generalizability and mitigating

overfitting, especially when using high boosting rounds and datasets with large numbers of predictor variables (Friedman 2001; Elith et al. 2008).

Differences between training and test environments were not modeled via environmental covariates or multi-environment-trial (MET) terms. The inclusion of MET data presents an opportunity to improve model performance, as including environmental covariates has been shown to benefit generalizability by better capturing G × E interactions in MET studies (Jarquín 2017; Scott et al. 2021; Westhues et al. 2021; Montesinos-López et al. 2022).

The small n size and low haplotype diversity of the dataset (500 lines) are another area for future expansion. The Multi-parent Advanced Generation Inter Cross (MAGIC) wheat dataset (Scott et al. 2021) offered broad allelic diversity but still could be improved by increasing haplotype variation (van de Wouw et al. 2010; Fradgley et al. 2019; Scott et al. 2021). A larger population with higher haplotype diversity could further improve the reliability and performance of the evaluated models (Norman et al. 2018).

Another direction for exploration is to expand our hyperparameter grid to tune other important GBM hyperparameters including tree depth (interaction.depth) and regularization parameters like minimum node size and colsample rate (Westhues et al. 2021). Additionally, tuning stochastic subsampling values in the range 0.4 - 0.6 have been shown to improve performance, as suggested in the foundational work (Friedman 2002). To determine global optima, we suggest using Bayesian optimization for its adaptability and efficiency, or expanding a manual grid search to

cover broader ranges and additional types of hyperparameters for a more thorough model optimization (Shahriari et al. 2016).

Future studies could address the stability of feature contributions or the performance trade-offs when optimizing model structural stability. One study identified a trade-off showing a ~2% reduction in predictive power for a 30% gain in structural stability (Bertsimas et al. 2025). Their work examined structural stability during model retraining, while our study focused on output stability across iterations. Applying a mixed integer optimization approach, which constrains variation in model structure, such as split variables and leaf values across successive retraining, could show whether the loci driving predictions across separate model training remain consistent, support feature selection efforts and provide evidence of predictive performance tradeoffs when structural stability is improved (Bertsimas et al. 2025).

Computation time increases consistently with higher boosting rounds, and adopting efficient implementations can reduce costs, which is particularly useful in time-sensitive applications (Fig. S3). An effective model strategy to optimize computational costs is to utilize early stopping, allowing the model to find parameters that generalize well and mitigate excessive epochs that can degrade generalizability (Prechelt 2012). The gbm.fit function lacks early stopping; however, alternative gradient boosting packages, XGBoost and LightGBM, contain an early_stopping_rounds argument that can be directly used to set the number of boosting iterations the model can perform without improvement on a specified evaluation metric, after which the training automatically stops (Chen et al. 2014; Ke et al. 2017; lightgbm package - RDocumentation).

This study focussed on one modeling approach (GBM). In future work, benchmarking different algorithms (GBM vs. random forests vs. GBLUP, etc.) using the same multi-objective evaluation under the same cross-year prediction scenario would be useful for model selection, with each algorithm being individually optimized.

Our study shows that model stability can be greatly improved without sacrificing ranking performance and correlation, although classification (AUC) favored different hyperparameters. These findings support the use of multi-objective optimization in genomic prediction models, where stability is evaluated alongside predictive performance metrics. This work outlines a practical framework for tuning stochastic models to deliver reliable and consistent predictions suitable for breeding decisions.

Data Availability

File S1 contains, Packages and Functions, Formulae of metrics and models, Hyperparameter Grid, Computing Environment.
File S2 contains, evaluation metric descriptive statistics used in data visualization.
File S3 contains, pairwise comparisons of the estimated marginal means and ANOVA tables
File S4 contains the complete evaluation metric results for each model configuration (stability metric per subset, performance metrics per iteration and subset)
File S5 contains the summary statistics for each trait-metric combination (aggregated across configurations)

The R scripts and datasets used for modelling, statistical analysis, and figure generation are available in the GitHub repository (https://github.com/HenryOmix/Multi-objective-Evaluation-and-Optimization-of-Stochastic-Gradient-Boosting-Machine).

The specific version of the code used to produce the results in this manuscript is permanently archived on Zenodo:

HenryOmix. (2025).

HenryOmix/Multi-objective-Evaluation-and-Optimization-of-Stochastic-Gradient-Boosting-Machine: v1.03 (v1.03). Zenodo. https://doi.org/10.5281/zenodo.16785772

ORCID

0000-0002-0356-5045 (Henry Munroe)
0009-0006-0921-3143 (Bright Enogieru Osatohanmwen)

Acknowledgements

We are grateful to Professor Dr. Wolfgang Link for his valuable input and for the discussions within the Department of Crop Sciences, Plant Breeding Methodology Division (University of Göttingen) for fostering a supportive research environment. We thank the University of Göttingen for providing computing resources and IT support used for genotype imputation in 2024. We acknowledge that OpenAI's ChatGPT assisted in text grammar and code refactoring, which were reviewed and verified by the authors. We especially thank the authors from the Scott et al. 2021 study for their genomic and phenotypic data.

Conflict of Interest

We declare that there are no conflicts of interest.

References

1. Pieruschka R, Schurr U. 2019. Plant Phenotyping: Past, Present, and Future. Plant Phenomics. 2019:7507131. https://doi.org/10.34133/2019/7507131

2. van Dijk ADJ, Kootstra G, Kruijer W, de Ridder D. 2020. Machine learning in plant science and plant breeding. iScience. 24(1):101890. https://doi.org/10.1016/j.isci.2020.101890

3. Calvin K et al. 2023. IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland. Intergovernmental Panel on Climate Change p 1–34. IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental

Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland. [accessed 2025 July 16]. https://www.ipcc.ch/report/ar6/syr/. https://doi.org/10.59327/ipcc/ar6-9789291691647.001

4. Yoosefzadeh Najafabadi M, Hesami M, Eskandari M. 2023. Machine Learning-Assisted Approaches in Modernized Plant Breeding Programs. Genes. 14(4):777. https://doi.org/10.3390/genes14040777

5. Crossa J et al. 2025. Machine learning algorithms translate big data into predictive breeding accuracy. Trends Plant Sci. 30(2):167–184. https://doi.org/10.1016/j.tplants.2024.09.011

6. Domingos P. 2012. A few useful things to know about machine learning. Commun ACM. 55(10):78–87. https://doi.org/10.1145/2347736.2347755

7. Feurer M, Hutter F. 2019. Hyperparameter Optimization. In: Hutter F, Kotthoff L, Vanschoren J, editors. Automated Machine Learning: Methods, Systems, Challenges. Springer International Publishing; p 3–33 [accessed 2025 July 16]. https://doi.org/10.1007/978-3-030-05318-5_1. https://doi.org/10.1007/978-3-030-05318-5_1

8. Meuwissen TH, Hayes BJ, Goddard ME. 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

9. Goddard M e., Hayes B j. 2007. Genomic selection. J Anim Breed Genet. 124(6):323–330. https://doi.org/10.1111/j.1439-0388.2007.00702.x

10. Blondel M, Onogi A, Iwata H, Ueda N. 2015. A Ranking Approach to Genomic Selection. PLOS ONE. 10(6):e0128570. https://doi.org/10.1371/journal.pone.0128570

11. Dekkers JCM, Su H, Cheng J. 2021. Predicting the accuracy of genomic predictions. Genet Sel Evol. 53(1):55. https://doi.org/10.1186/s12711-021-00647-w

12. Lange TM, Gültas M, Schmitt AO, Heinrich F. 2025. optRF: Optimising random forest stability by determining the optimal number of trees. BMC Bioinformatics. 26(1):95. https://doi.org/10.1186/s12859-025-06097-1

13. Järvelin K, Kekäläinen J. 2002. Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst. 20(4):422–446. https://doi.org/10.1145/582415.582418

14. Huang J, Ling CX. 2005. Using AUC and Accuracy in Evaluating Learning Algorithms. IEEE Trans Knowl Data Eng. 17(3):299–310. https://doi.org/10.1109/TKDE.2005.50

15. Robin X et al. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 12(1):77. https://doi.org/10.1186/1471-2105-12-77

16. Friedman JH. 2002. Stochastic gradient boosting. Comput Stat Data Anal. 38(4):367–378 (Nonlinear Methods and Data Mining). https://doi.org/10.1016/S0167-9473(01)00065-2

17. Bartko JJ. 1966. The Intraclass Correlation Coefficient as a Measure of Reliability. Psychol Rep. 19(1):3–11. https://doi.org/10.2466/pr0.1966.19.1.3

18. McGraw KO, Wong SP. 1996. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1(1):30–46. https://doi.org/10.1037/1082-989X.1.1.30

19. Weir JP. 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 19(1):231–240. https://doi.org/10.1519/15184.1

20. Koo TK, Li MY. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 15(2):155–163. https://doi.org/10.1016/j.jcm.2016.02.012

21. Fleiss JL. 1971. Measuring nominal scale agreement among many raters. Psychol Bull. 76(5):378–382. https://doi.org/10.1037/h0031619

22. Friedman JH. 2001. Greedy Function Approximation: A Gradient Boosting Machine. Ann Stat. 29(5):1189–1232

23. Ridgeway G, Developers G. 2003. gbm: Generalized Boosted Regression Models. 2.2.2 [accessed 2025 Apr 4]. https://CRAN.R-project.org/package=gbm. https://doi.org/10.32614/CRAN.package.gbm

24. Scott MF et al. 2021. Limited haplotype diversity underlies polygenic trait architecture across 70 years of wheat breeding. Genome Biol. 22(1):137. https://doi.org/10.1186/s13059-021-02354-7

25. Winfield MO et al. 2016. High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. Plant Biotechnol J. 14(5):1195–1206. https://doi.org/10.1111/pbi.12485

26. Mayer M. 2024. missRanger: Fast Imputation of Missing Values. [accessed 2025 July 6]. https://cran.r-project.org/web/packages/missRanger/index.html

27. Shrout PE. 1998. Measurement reliability and agreement in psychiatry. Stat Methods Med Res. 7(3):301–317. https://doi.org/10.1177/096228029800700306

28. Bradley AP. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30(7):1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

29. Peng H et al. 2024. Optimizing differential expression analysis for proteomics data via high-performing rules and ensemble inference. Nat Commun. 15(1):3922. https://doi.org/10.1038/s41467-024-47899-w

30. McClish DK. 1989. Analyzing a Portion of the ROC Curve. Med Decis Making. 9(3):190–195. https://doi.org/10.1177/0272989X8900900307

31. Simmonds NW. 1993. Introgression and Incorporation. Strategies for the Use of Crop Genetic Resources. Biol Rev. 68(4):539–562. https://doi.org/10.1111/j.1469-185X.1993.tb01243.x

32. Jarquín. 2017. [accessed 2025 May 7]. https://acsess.onlinelibrary.wiley.com/doi/full/10.3835/plantgenome2016.12.0130

33. Tomar V et al. 2021. Increased Predictive Accuracy of Multi-Environment Genomic Prediction Model for Yield and Related Traits in Spring Wheat (Triticum aestivum L.).

Front Plant Sci. 12 [accessed 2025 May 7]. https://www.frontiersin.orghttps://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2021.720123/full. https://doi.org/10.3389/fpls.2021.720123

34. Montesinos-López OA et al. 2022. Multi-trait genome prediction of new environments with partial least squares. Front Genet. 13 [accessed 2025 May 7]. https://www.frontiersin.orghttps://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2022.966775/full. https://doi.org/10.3389/fgene.2022.966775

35. Pan S et al. 2024. Using the Pearson's correlation coefficient as the sole metric to measure the accuracy of quantitative trait prediction: is it sufficient? Front Plant Sci. 15 [accessed 2025 May 6]. https://www.frontiersin.orghttps://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2024.1480463/full. https://doi.org/10.3389/fpls.2024.1480463

36. Gianola D, Cecchinato A, Naya H, Schön C-C. 2018. Prediction of Complex Traits: Robust Alternatives to Best Linear Unbiased Prediction. Front Genet. 9 [accessed 2025 May 6]. https://www.frontiersin.orghttps://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2018.00195/full. https://doi.org/10.3389/fgene.2018.00195

37. Negussie E et al. 2022. Integrating heterogeneous across-country data for proxy-based random forest prediction of enteric methane in dairy cattle. J Dairy Sci. 105(6):5124–5140. https://doi.org/10.3168/jds.2021-20158

38. Li T et al. 2023. IP4GS: Bringing genomic selection analysis to breeders. Front Plant Sci. 14 [accessed 2025 May 6]. https://www.frontiersin.orghttps://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2023.1131493/full. https://doi.org/10.3389/fpls.2023.1131493

39. Westhues CC et al. 2021. Prediction of Maize Phenotypic Traits With Genomic and Environmental Predictors Using Gradient Boosting Frameworks. Front Plant Sci. 12 [accessed 2025 Apr 29]. https://www.frontiersin.orghttps://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2021.699589/full. https://doi.org/10.3389/fpls.2021.699589

40. Elith J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. J Anim Ecol. 77(4):802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

41. Azodi CB et al. 2019. Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. G3 GenesGenomesGenetics. 9(11):3691–3702. https://doi.org/10.1534/g3.119.400498

42. Sandhu KS et al. 2021. Genomic Selection and Genome-Wide Association Studies for Grain Protein Content Stability in a Nested Association Mapping Population of Wheat. Agronomy. 11(12):2528. https://doi.org/10.3390/agronomy11122528

43. Paina C, Gregersen PL. 2023. Recent advances in the genetics underlying wheat grain protein content and grain protein deviation in hexaploid wheat. Plant Biol. 25(5):661–670. https://doi.org/10.1111/plb.13550

44. van de Wouw M et al. 2010. Genetic diversity trends in twentieth century crop cultivars: a meta analysis. Theor Appl Genet. 120(6):1241–1252.

https://doi.org/10.1007/s00122-009-1252-6

45. Fradgley N et al. 2019. A large-scale pedigree resource of wheat reveals evidence for adaptation and selection by breeders. PLOS Biol. 17(2):e3000071. https://doi.org/10.1371/journal.pbio.3000071

46. Norman A, Taylor J, Edwards J, Kuchel H. 2018. Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. G3 GenesGenomesGenetics. 8(9):2889–2899. https://doi.org/10.1534/g3.118.200311

47. Shahriari B et al. 2016. Taking the Human Out of the Loop: A Review of Bayesian Optimization. Proc IEEE. 104(1):148–175. https://doi.org/10.1109/JPROC.2015.2494218

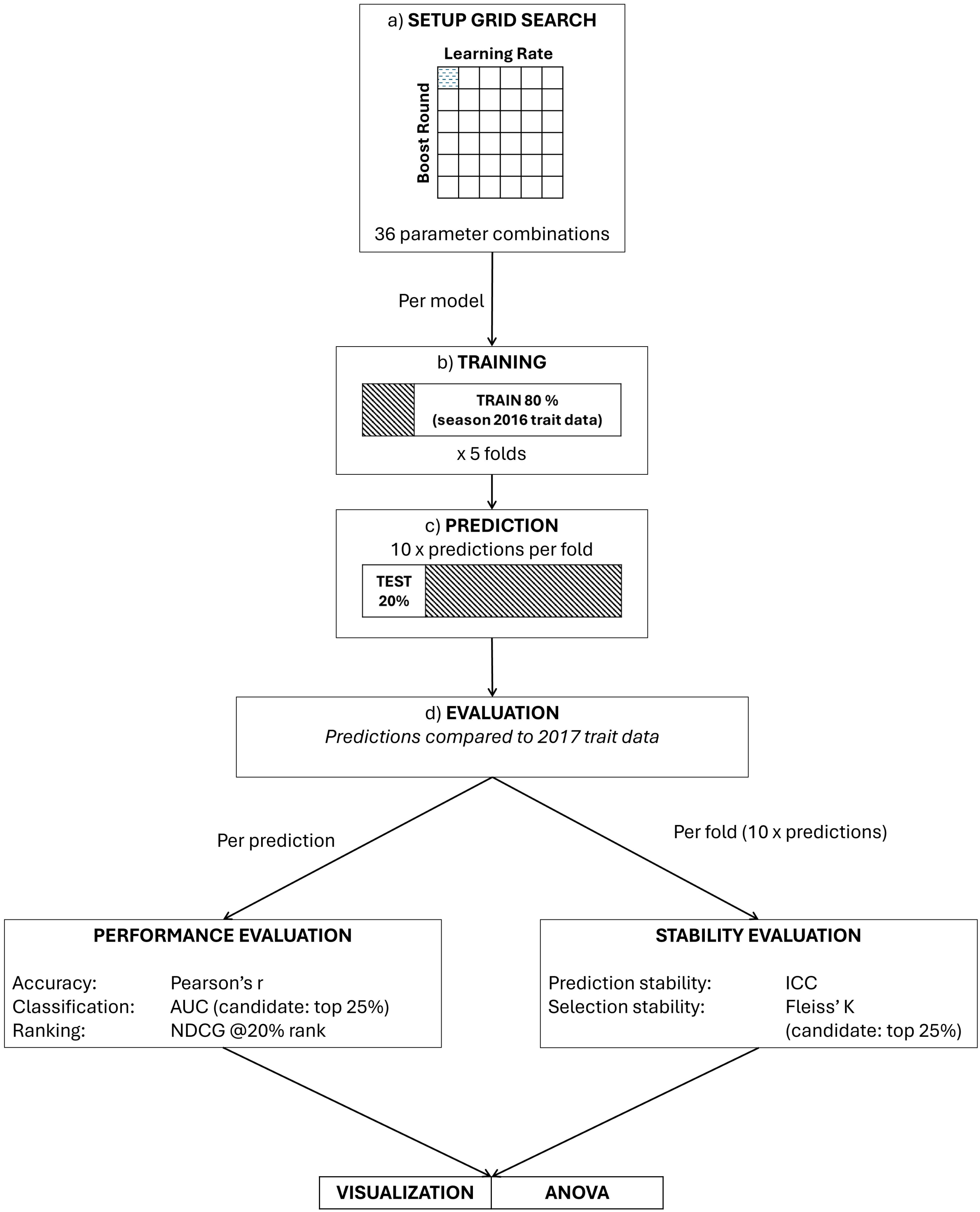
48. Bertsimas D, Jr VD, Ma Y, Paschalidis P. 2025. Towards Stable Machine Learning Model Retraining via Slowly Varying Sequences. [accessed 2025 Apr 4]. http://arxiv.org/abs/2403.19871. https://doi.org/10.48550/arXiv.2403.19871

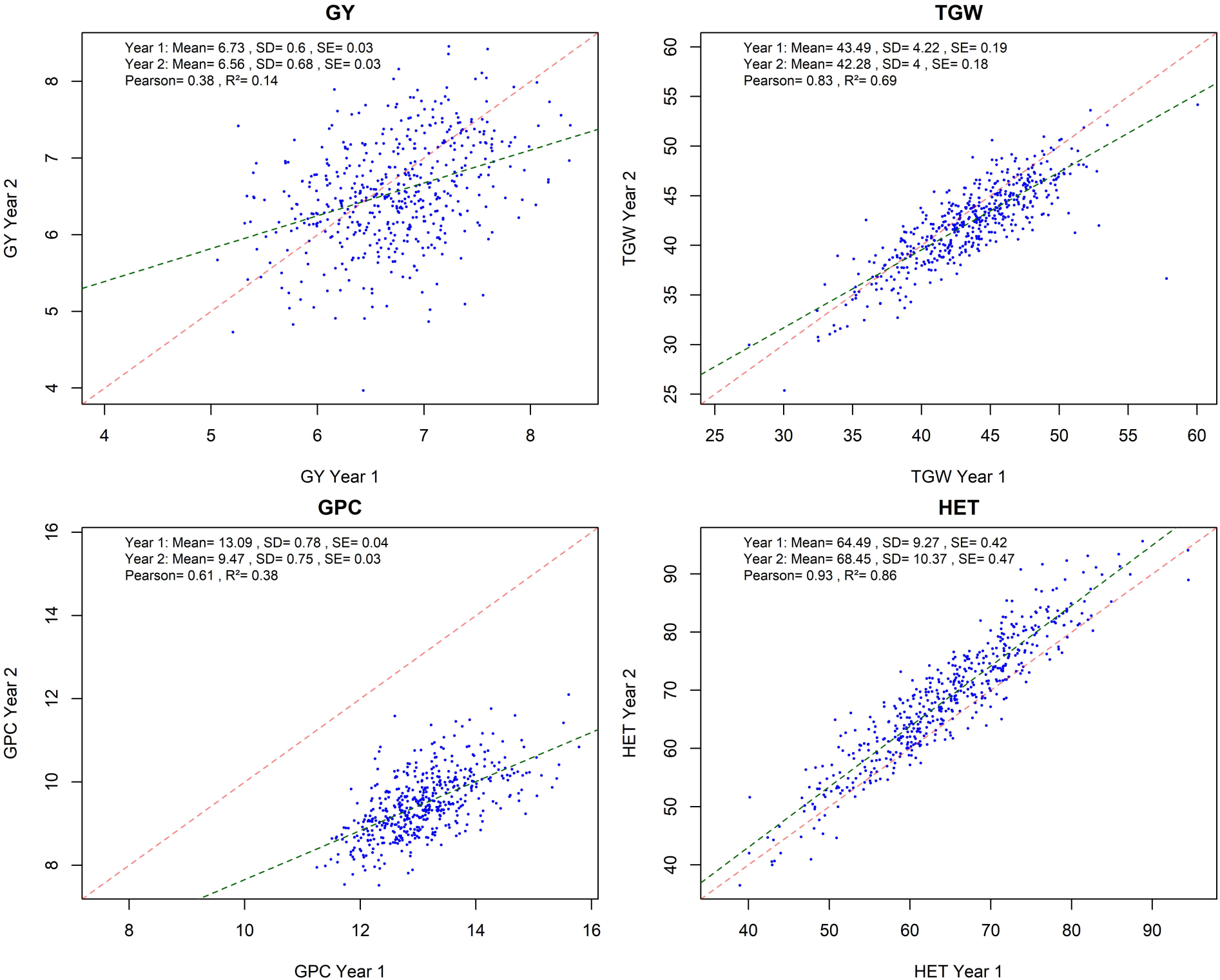
49. Prechelt L. 2012. Early Stopping — But When? In: Montavon G, Orr GB, Müller K-R, editors. Neural Networks: Tricks of the Trade: Second Edition. Springer; p 53–67 [accessed 2025 July 16]. https://doi.org/10.1007/978-3-642-35289-8_5. https://doi.org/10.1007/978-3-642-35289-8_5

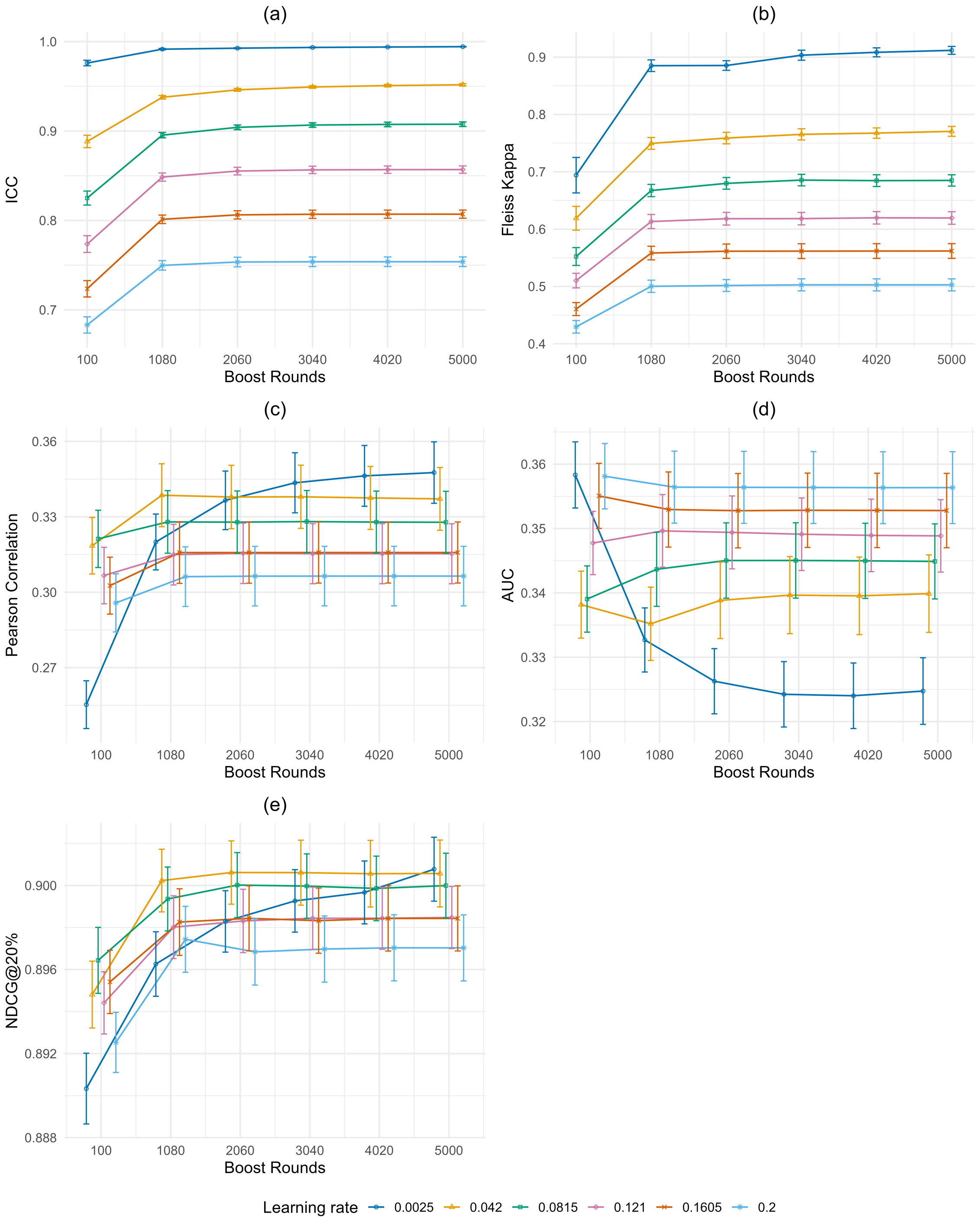
50. Chen T et al. 2014. xgboost: Extreme Gradient Boosting. 1.7.10.1 [accessed 2025 May 5]. https://CRAN.R-project.org/package=xgboost. https://doi.org/10.32614/CRAN.package.xgboost

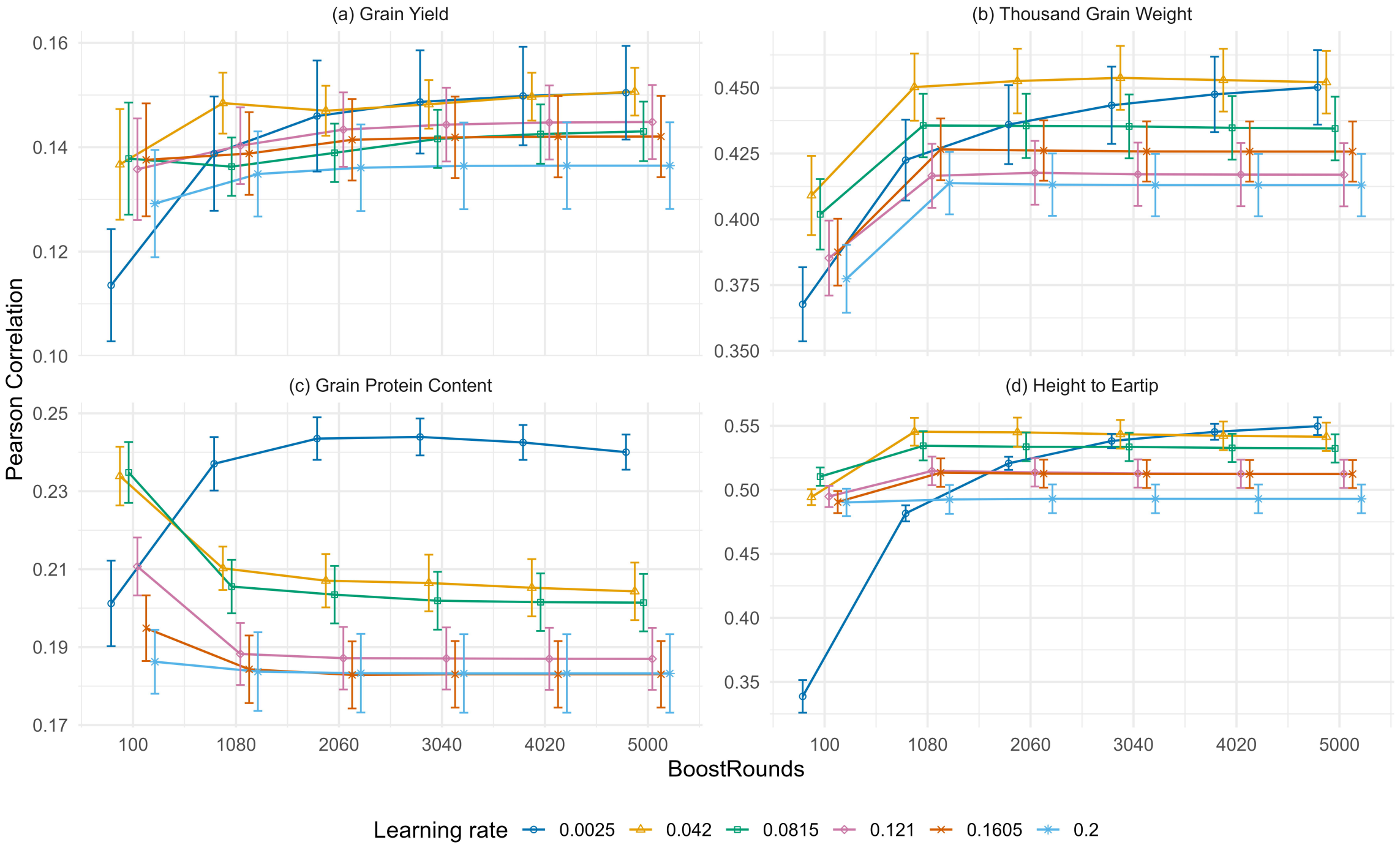
51. Ke G et al. 2017. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc.; p 3149–3157 (NIPS'17). [accessed 2025 Apr 4]

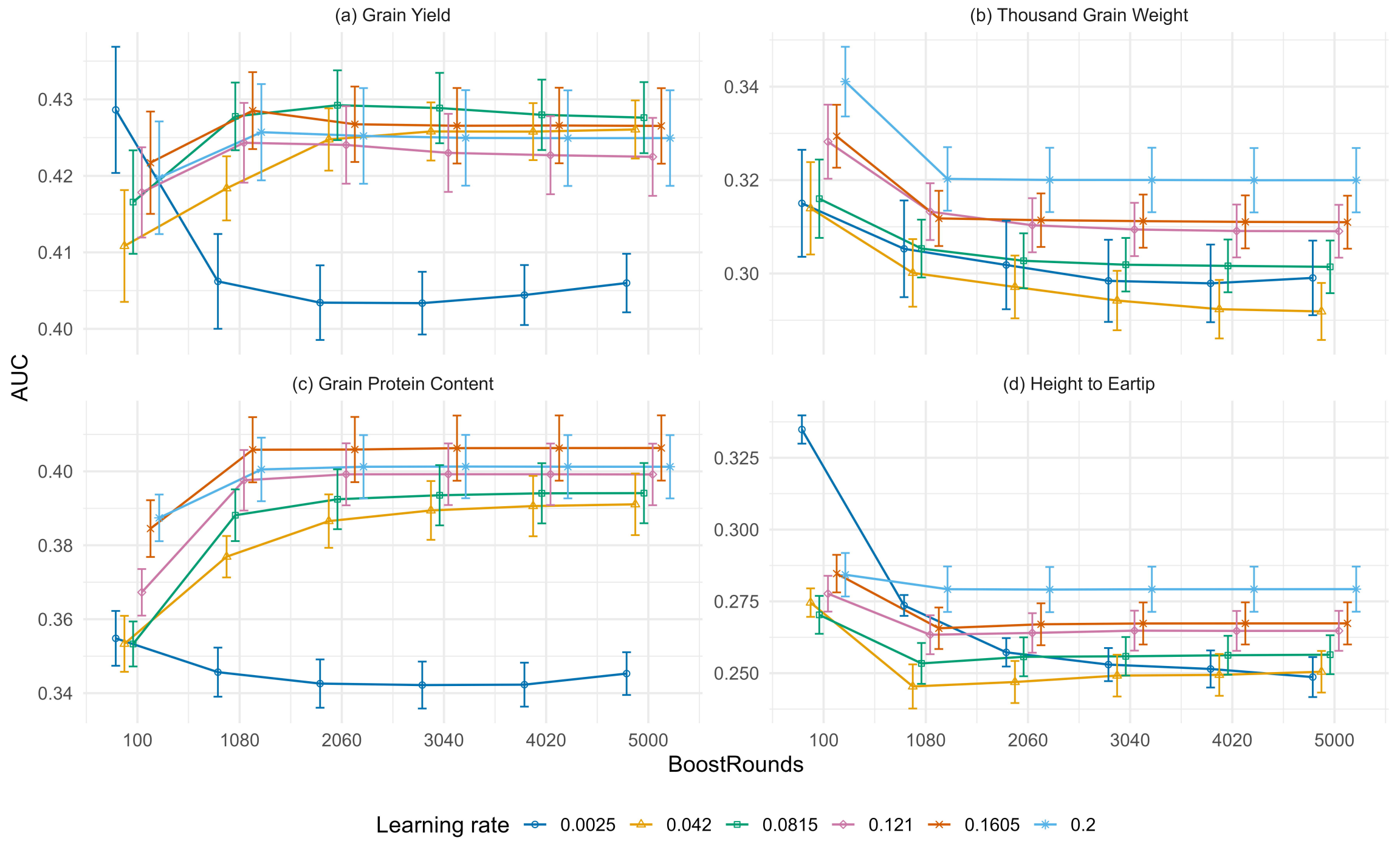
52. lightgbm package - RDocumentation. [accessed 2025 May 5]. https://www.rdocumentation.org/packages/lightgbm/versions/4.6.0

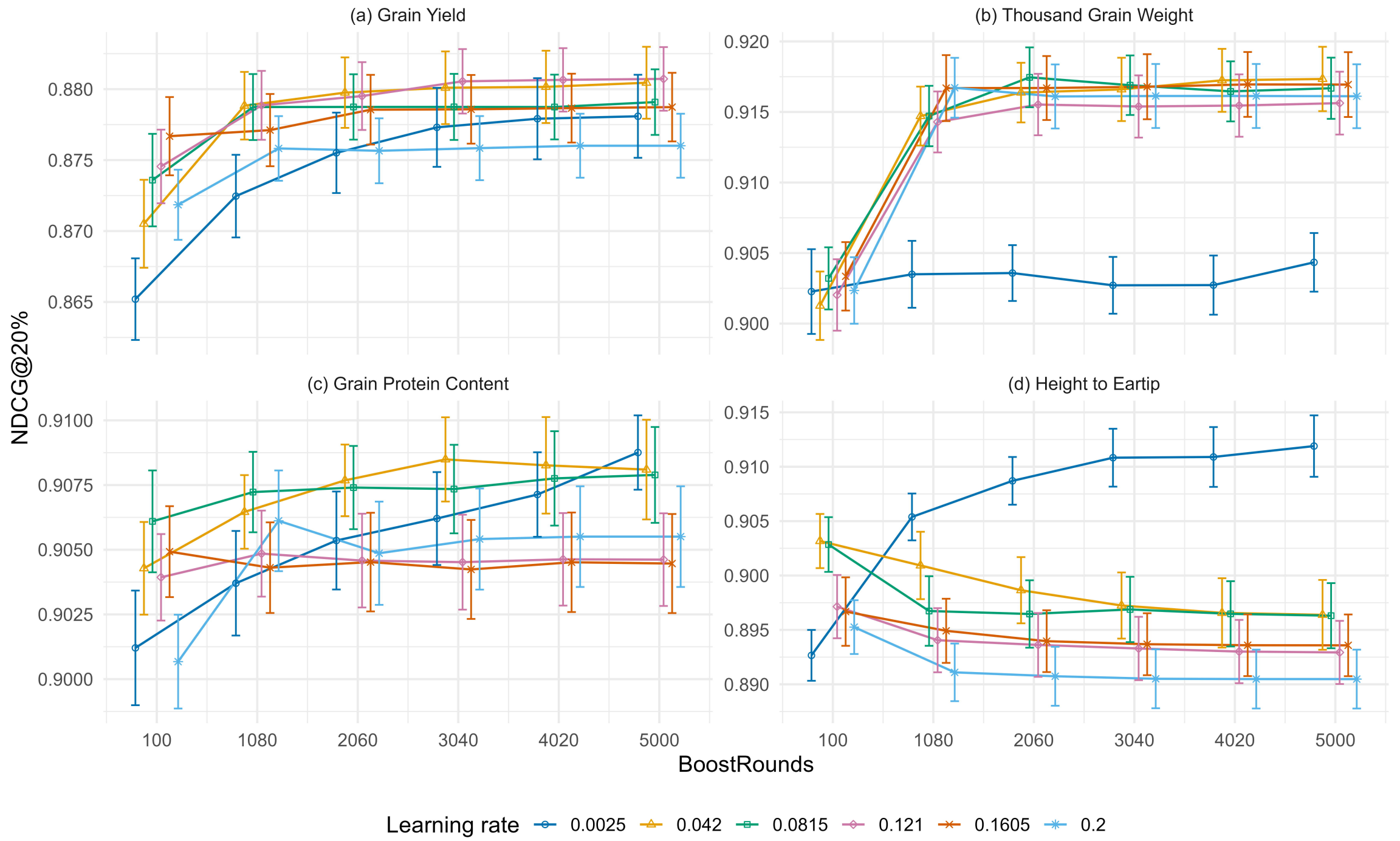

a) SETUP GRID SEARCH
Learning Rate
Boost Round
36 parameter combinations
Per model
b) TRAINING
TRAIN 80 %
(season 2016 trait data)
x 5 folds
c) PREDICTION
10 x predictions per fold
TEST
20%
d) EVALUATION
Predictions compared to 2017 trait data
Per prediction
Per fold (10 x predictions)
PERFORMANCE EVALUATION
Accuracy: Pearson's r
Classification: AUC (candidate: top 25%)
Ranking: NDCG @20% rank
STABILITY EVALUATION
Prediction stability: ICC
Selection stability: Fleiss' K
(candidate: top 25%)
VISUALIZATION
ANOVA

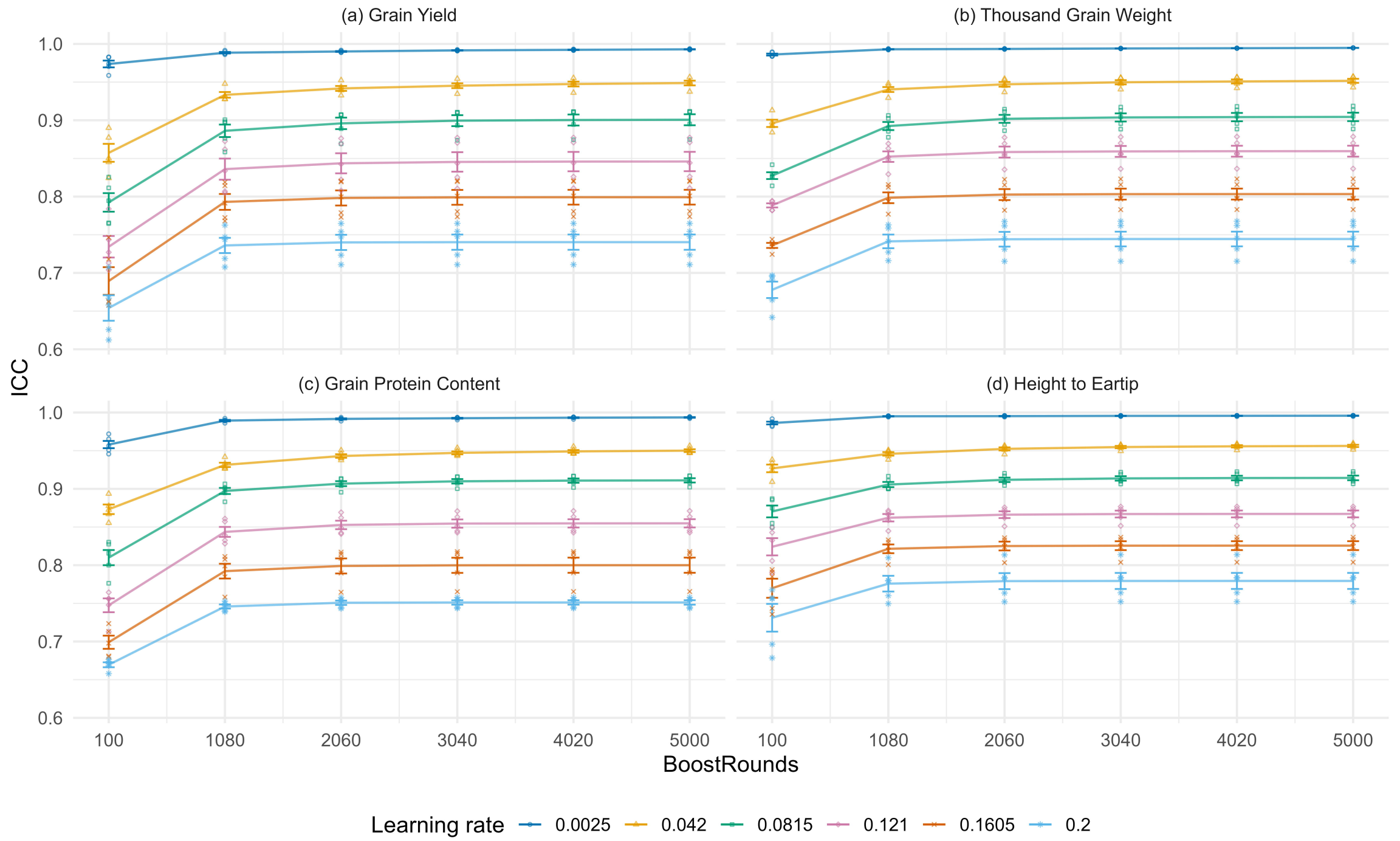

GY
Year 1: Mean= 6.73 , SD= 0.6 , SE= 0.03
Year 2: Mean= 6.56 , SD= 0.68 , SE= 0.03
Pearson= 0.38 , R²= 0.14
GY Year 1
GY Year 2
TGW
Year 1: Mean= 43.49 , SD= 4.22 , SE= 0.19
Year 2: Mean= 42.28 , SD= 4 , SE= 0.18
Pearson= 0.83 , R²= 0.69
TGW Year 1
TGW Year 2
GPC
Year 1: Mean= 13.09 , SD= 0.78 , SE= 0.04
Year 2: Mean= 9.47 , SD= 0.75 , SE= 0.03
Pearson= 0.61 , R²= 0.38
GPC Year 1
GPC Year 2
HET
Year 1: Mean= 64.49 , SD= 9.27 , SE= 0.42
Year 2: Mean= 68.45 , SD= 10.37 , SE= 0.47
Pearson= 0.93 , R²= 0.86
HET Year 1
HET Year 2

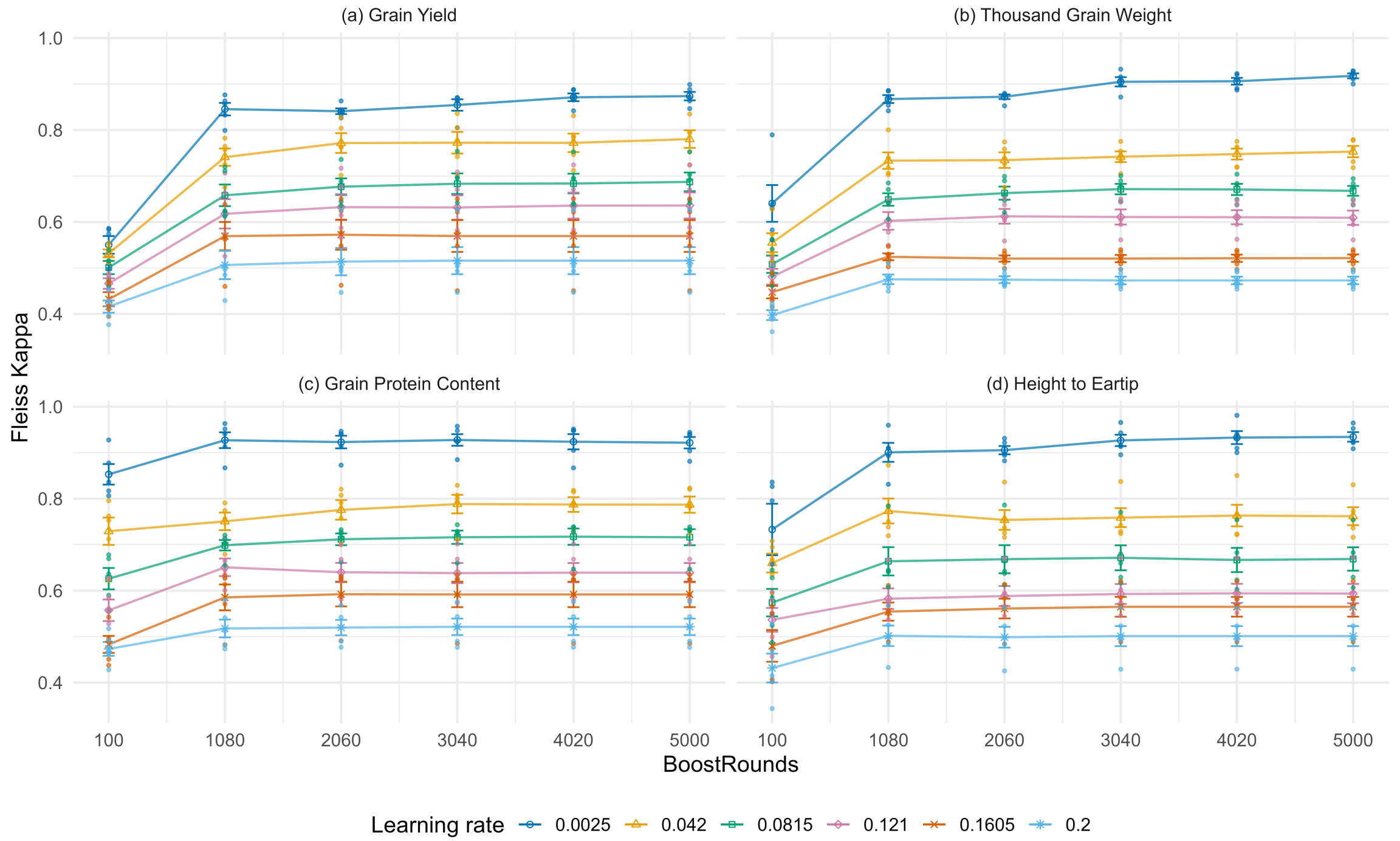

(a)
ICC
1.0
0.9
0.8
0.7
100
1080
2060
3040
4020
5000
Boost Rounds
(b)
Fleiss Kappa
0.9
0.8
0.7
0.6
0.5
0.4
100
1080
2060
3040
4020
5000
Boost Rounds
(c)
Pearson Correlation
0.36
0.33
0.30
0.27
100
1080
2060
3040
4020
5000
Boost Rounds
(d)
AUC
0.36
0.35
0.34
0.33
0.32
100
1080
2060
3040
4020
5000
Boost Rounds
(e)
NDCG@20%
0.900
0.896
0.892
0.888
100
1080
2060
3040
4020
5000
Boost Rounds
Learning rate
0.0025
0.042
0.0815
0.121
0.1605
0.2

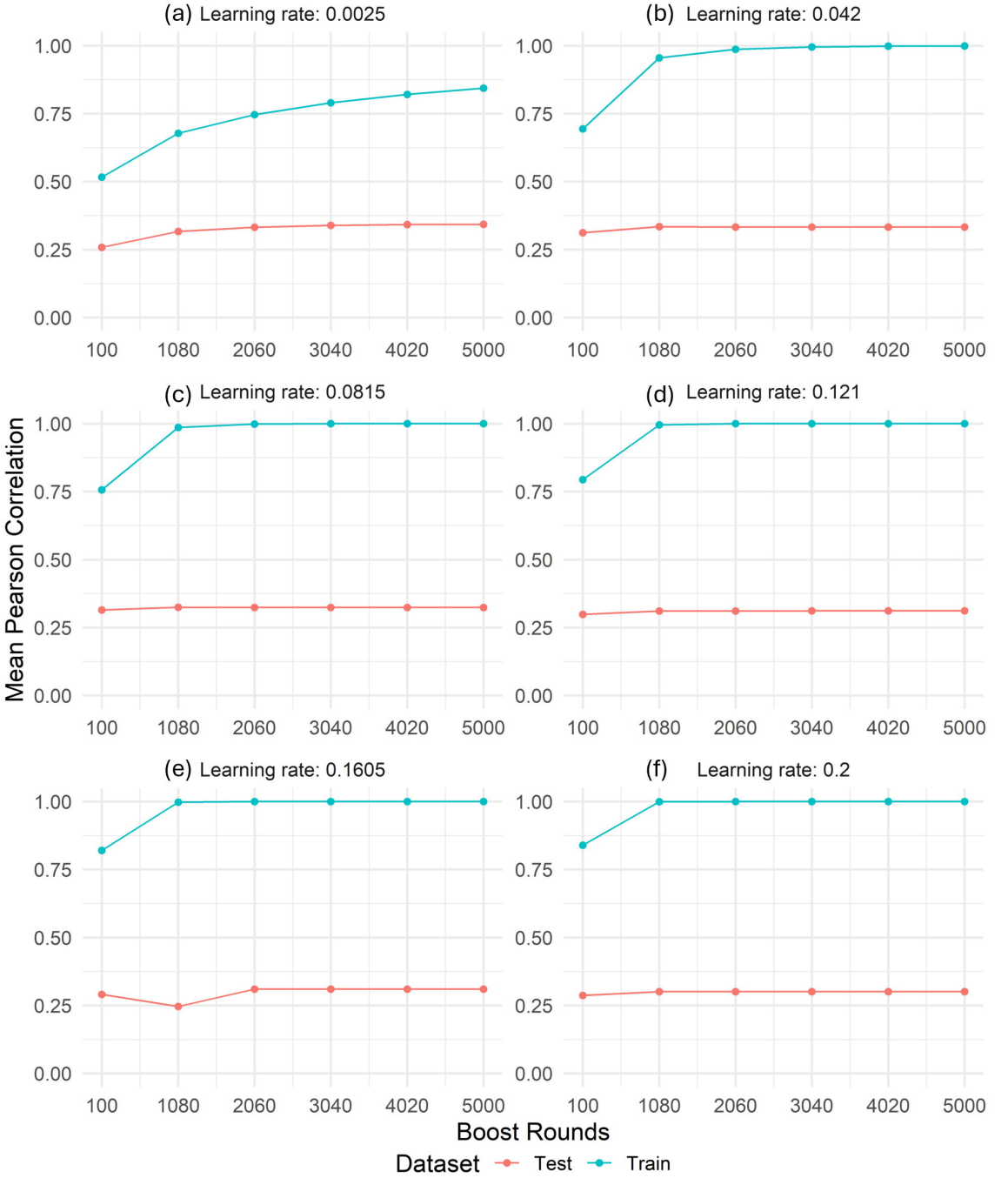

(a) Grain Yield
(b) Thousand Grain Weight
(c) Grain Protein Content
(d) Height to Eartip
Pearson Correlation
BoostRounds
0.16
0.14
0.12
0.10
0.450
0.425
0.400
0.375
0.350
0.25
0.23
0.21
0.19
0.17
0.55
0.50
0.45
0.40
0.35
100
1080
2060
3040
4020
5000
Learning rate
0.0025
0.042
0.0815
0.121
0.1605
0.2


(a) Grain Yield
(b) Thousand Grain Weight
(c) Grain Protein Content
(d) Height to Eartip
AUC
0.43
0.42
0.41
0.40
0.34
0.32
0.30
0.40
0.38
0.36
0.34
0.325
0.300
0.275
0.250
100
1080
2060
3040
4020
5000
BoostRounds
Learning rate
0.0025
0.042
0.0815
0.121
0.1605
0.2


(a) Grain Yield
(b) Thousand Grain Weight
(c) Grain Protein Content
(d) Height to Eartip
NDCG@20%
BoostRounds
0.880
0.875
0.870
0.865
0.920
0.915
0.910
0.905
0.900
0.9100
0.9075
0.9050
0.9025
0.9000
0.915
0.910
0.905
0.900
0.895
0.890
100
1080
2060
3040
4020
5000
Learning rate
0.0025
0.042
0.0815
0.121
0.1605
0.2


(a) Grain Yield
(b) Thousand Grain Weight
(c) Grain Protein Content
(d) Height to Eartip
ICC
1.0
0.9
0.8
0.7
0.6
100
1080
2060
3040
4020
5000
BoostRounds
Learning rate
0.0025
0.042
0.0815
0.121
0.1605
0.2


(a) Grain Yield
(b) Thousand Grain Weight
(c) Grain Protein Content
(d) Height to Eartip
Fleiss Kappa
1.0
0.8
0.6
0.4
100
1080
2060
3040
4020
5000
BoostRounds
Learning rate
0.0025
0.042
0.0815
0.121
0.1605
0.2


(a) Learning rate: 0.0025
(b) Learning rate: 0.042
(c) Learning rate: 0.0815
(d) Learning rate: 0.121
(e) Learning rate: 0.1605
(f) Learning rate: 0.2
Mean Pearson Correlation
1.00
0.75
0.50
0.25
0.00
100
1080
2060
3040
4020
5000
Boost Rounds
Dataset
Test
Train