Improvements in Prediction Performance of Ensemble Approaches for Genomic Prediction in Crop Breeding

Shunichiro Tomura[1,2,*], Melanie Wilkinson[1,2], Mark Cooper[1,2] and Owen Powell[1,2]

[1]Queensland Alliance for Agriculture and Food Innovation (QAAFI), Centre for Crop Science, the University of Queensland, St Lucia, QLD 4072, Australia
[2]ARC Centre of Excellence for Plant Success in Nature and Agriculture, the University of Queensland, St Lucia, QLD 4072, Australia

*Corresponding author: s.tomura@uq.edu.au

The refinement of prediction accuracy in genomic prediction is a key factor in accelerating genetic gain for crop breeding. The mainstream strategy for improving prediction performance has been to develop an individual genomic prediction model that performs better than others. However, this approach has limitations due to the absence of consistent outperformance by an individual genomic prediction model. This problem is aligned with the implication from the No Free Lunch Theorem, which states that the performance of an individual prediction model is expected to be equivalent to the others when averaged across all prediction scenarios. Hence, we applied an alternative method, combining multiple individual genomic prediction models into an ensemble, to investigate the potential for increased prediction performance. We developed a naïve ensemble-average model, which averages the predicted phenotypes of individual genomic prediction models with equal weight. The prediction performance of the genomic prediction models was evaluated over two traits that influence crop yield, days to anthesis and tiller number, in a publicly available Teosinte Nested Association Mapping (TeoNAM) dataset. Here we show an improvement in prediction accuracy for the ensemble approach compared with the six individual genomic prediction models in more than 95% of prediction scenarios. The advantage of the ensemble was derived from a more comprehensive view of the genomic architecture of these complex traits provided by contrasting genomic marker effects from the individual genomic prediction models. These results indicate the potential for ensemble approaches to enhance the performance of genomic prediction and therefore genetic gain, in crop breeding programs.

Ensemble; Genomic prediction; Machine learning; Interpretability; No Free Lunch Theorem; Crop breeding

Article summary

This research targets crop breeding industries and researchers developing prediction models that can accurately predict phenotypes using genetic marker information. We applied the concept of an ensemble, combining multiple individual genomic prediction models, to predict key traits (days to anthesis and tiller numbers) in crop breeding. The results showed that the prediction accuracy in the ensemble approach outperformed other models in over 95% of prediction cases. This indicates that the ensemble contributes to correctly selecting genotypes with desirable traits.

Introduction

Genomic selection has accelerated the rates of genetic gain in plant breeding programs (Voss-Fels *et al.* 2019) by using genomic prediction models to associate genetic markers with trait phenotypes (Meuwissen *et al.* 2001). The ability to predict and select plants based on genetic markers, instead of trait phenotypes, has enabled novel selection schemes and breeding program designs with the potential to accelerate rates of genetic gain (Heffner *et al.* 2009; Gaynor *et al.* 2017; Powell *et al.* 2020), primarily attributed to the reduction of breeding cycle time. As of today, genomic selection has accelerated the rates of genetic gain in the form of grain yield increase in commercial breeding programs (Cooper *et al.* 2014) and the integration of genomic selection in public breeding programs is underway (Dreisigacker *et al.* 2021; Prasanna *et al.* 2021).

The accuracy of genomic selection is dependent on the choice of a genomic prediction model. Therefore, research evaluating alternative genomic prediction models has received considerable attention. The primary goal is to develop a genomic prediction model that can consistently reach higher prediction performance compared with others. One major challenge in improving the performance of genomic prediction models is capturing complex genetic interactions that can often be demonstrated as a gene regulatory network (Mascher *et al.* 2024). Cooper *et al.* (2005) demonstrated that epistatic (nonlinear) marker interactions decreased the response to selection rate compared to the scenarios where only additive (linear) effects existed, emphasising the necessity to capture complex genetic interactions to maintain a high response rate. Mackay (2014) discussed the possibility of performance improvement on genomic prediction by including epistatic effects in prediction models. Previous research results (Montesinos-López *et al.* 2018; Pérez-Enciso and Zingaretti 2019; Washburn *et al.* 2021) implied the difficulty of finding an individual genomic prediction model that can consistently capture such complex interactions and outperform others after comparing various linear and nonlinear genomic prediction models.

Outside of plant breeding, the absence of a best prediction model has been explained by the No Free Lunch Theorem (Wolpert and Macready 1997). The No Free Lunch Theorem postulates that the average prediction performance of individual models becomes equal to that of all prediction problem scenarios. If we accept that this theorem applies generally, seeking superior individual genomic prediction models in plant breeding programs is unlikely to be successful. An alternative approach needs to be adopted and one potential could be to generate combinations of multiple, different genomic prediction models.

Ensemble approaches combine predictions from multiple models (Farooq *et al.* 2022). Several ensemble approaches have been proposed such as Bootstrap aggregating (an algorithm containing a number of prediction weak models in a parallel way) (Breiman 1996) and

AdaBoost (an algorithm leveraging a number of weak prediction models sequentially) (Freund and Schapire 1995). These ideas informed the development of approaches such as random forest (Breiman 2001a) and ensemble neural networks (Zhou *et al.* 2002; Li *et al.* 2004; Liu and Li 2008). Combining multiple distinctive prediction models is expected to cancel out errors derived from the gap between observed and actual values, leading to performance improvement (Kick and Washburn 2023).

Ensemble approaches have been leveraged in agricultural science. For instance, Wallach *et al.* (2018) demonstrated the higher prediction accuracy of ensemble approaches for forecasting climate change scenarios in crop yield prediction. Other examples have shown the superiority of ensemble approaches compared to approaches using individual genomic prediction models (McCormick *et al.* 2021; Huang and Wei 2022; Fradgley *et al.* 2023; Heilmann *et al.* 2023; Kick and Washburn 2023; Washburn *et al.* 2024). However, key factors leading to the performance increases of ensemble approaches for genomic selection have not been thoroughly discussed. Therefore, we examine the performance of ensemble approaches for genomic selection in comparison to that of the individual genomic prediction models and dissect the factors leading to performance increases.

Here, we (1) investigate the prediction performance of six individual genomic prediction models for two traits controlled by a genetic architecture involving both linear and nonlinear interactions, (2) consider the implications of the No Free Lunch Theorem and investigate an ensemble approach to assess its effectiveness in comparison to the individual genomic prediction models and (3) outline a framework to investigate the contribution of individual genomic prediction models to an ensemble.

Materials and methods

Dataset

The TeoNAM dataset (Chen *et al.* 2019) consists of five recombinant inbred line (RIL) populations derived from five backcross hybrid crosses of the maize (*Z. Mays*) line W22 and five teosinte inbred lines; wild teosinte types TIL01, TIL03, TIL11 and TIL14 from *Z. mays ssp. parviglumis* and TIL25 from *Z.mays ssp. mexicana*. The maize inbred line W22 was used as the female plant for crossing. Following the F1 cross, the F1 was backcrossed once to W22. Each cross generated one population and thus five different populations were generated. After the backcross, each population was advanced four times by controlled self-pollination to produce the RILs used for the analyses.

Each population was measured for seven agronomic traits and fifteen domestication traits. From the full trait list, days to anthesis (DTA) and tiller number (TILN) were chosen as target traits for the investigations as both traits are expected to be a consequence of genetic interactions in a biologically complex network (Dong *et al.* 2012; DeWitt *et al.* 2021; Powell *et al.* 2022).

The traits were measured in two different environments. For W22TIL01, W22TIL03 and W22TIL11, the experiment was conducted in 2015 and 2016 summer. W22TIL14 was grown in 2016 and 2017 summer and W22TIL25 was developed in two blocks in 2017 summer. All the experiments were conducted at the University of Wisconsin West Madison Agricultural Research Station with a randomised complete block design.

The summary of genetic marker (SNP) and RIL numbers per cross are described in Table 1. Each population contains at least 200 RILs with more than 10,000 SNPs. More specifically, W22TIL03 possesses the highest SNP numbers, followed by W22TIL11, W22TIL01, W22TIL25 and W22TIL14. As per the number of RILs, W22TIL25 has the largest number at 308, followed by W22TIL03, W22TIL14, W22TIL01 and W22TIL11.

Table 1 The number of SNPs in each preprocessing phase and recombinant inbred lines (RILs) in each population

Population	SNPs			RIL
Name	Original	Imputation	LD Filtering	Number
W22TIL01	13,089	13,042	274 - 322	222
W22TIL03	16,110	16,076	295 - 342	270
W22TIL11	13,188	13,153	268 - 314	219
W22TIL14	11,396	11,375	270 - 320	230
W22TIL25	14,885	14,857	294 - 341	308

Data cleaning & preprocessing

Preliminary quality control on the TeoNAM dataset identified three potential issues that required further investigation prior to the application of prediction models; (1) missing genomic marker calls, (2) missing trait measurements/records and (3) a larger number of genomic markers compared to the number of RILs.

Missing marker calls was one noticeable problem affecting the quality of the datasets. Imputation was undertaken when possible. The missing information was imputed with the most frequent allele in each SNP. When SNPs contained more than 10% of missing marker calls, the imputation was not feasible and these SNPs were removed.

A missing target trait problem occurred when the phenotype of a target trait was missing. Imputation was an infeasible choice in this dataset due to a lack of external information enabling the reasonable imputation of missing phenotypes. Hence, RILs without target trait phenotypes were excluded in this study.

The high proportion of the marker numbers relative to the RIL numbers in each cross could negatively affect the performance of genomic prediction models. The number of markers in all the crosses was significantly higher than the number of RILs (Table 1). The RIL numbers may not have been sufficient to capture complex patterns among markers and phenotypes, resulting in the curse of dimensionality (Bellman

1957; Ramstein *et al.* 2019). Therefore, we reduced marker dimensionality by eliminating less informative markers based on their linkage disequilibrium (LD) relationships. When the squared correlation, calculated by Plink (v1.9) (Chang *et al.* 2015), was higher than the threshold of 0.8 after considering all the possible genetic marker pairs, the compared marker was removed from the datasets. A higher correlation indicated that two genetic markers provided similar information for prediction. Thus, removing either of the genetic markers could be beneficial by reducing the total number of genetic markers rather than losing critical information for better predictions.

Since each RIL population was grown and measured in two distinctive environments, the sub-datasets were concatenated into a single dataset with a dummy variable representing different environments to increase the available training data size.

The datasets after undertaking all the preprocessing steps were used as input for the six individual genomic prediction models. The final genetic marker (SNP) and RIL numbers used in this study are summarised in Table 1. The final number of SNPs varied depending on the combination of RILs in the training set for each sample.

Individual genomic prediction models

Six prediction models, three classical genomic prediction and three machine learning (ML) models, were applied to the TeoNAM dataset. The three classical genomic prediction models were rrBLUP (Meuwissen *et al.* 2001), BayesB (Meuwissen *et al.* 2001) and Reproducing Kernel Hilbert Space (RKHS) (Gianola and van Kaam 2008). The three ML models were random forest (RF) (Breiman 2001a), support vector regression (SVR) (Drucker *et al.* 1996) and graph attention network (GAT) (Brody *et al.* 2021).

Classical genomic prediction models

Parametric models: Parametric models have been widely leveraged for genomic prediction (Meuwissen *et al.* 2001). One major characteristic in parametric models is the requirement of assumptions in the input distribution. Such assumptions allow the model parameters to be determined within a finite value range.

The most well-known parametric models in genomic prediction are linear mixed models. They have been leveraged since the initial developmental stage in genomic prediction (Ray *et al.* 2022). A linear mixed model, in general, can be formulated as Equation 1 (Pérez and de Los Campos 2014):

$$\boldsymbol{\eta} = 1\boldsymbol{\mu} + \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j + \sum_{l=1}^{L} \mathbf{u}_l \tag{1}$$

where $\boldsymbol{\eta} = \{\eta_1, \eta_2, ..., \eta_n\}$ is a set of the predicted phenotypes for true values $\mathbf{y} = \{y_1, y_2, ..., y_n\}$, $\boldsymbol{\mu}$ is intercept, $\mathbf{X}_j$ is design matrices representing marker values, $\boldsymbol{\beta}_j$ is coefficient matrices indicating each genomic marker effects and $\mathbf{u}_l = \{u_{l1}, u_{l2}..., u_{ln}\}$ is random effects in a vector format. The assumed value distribution for $\boldsymbol{\beta}_j$ differs among models. The linear mixed models aim to determine $\boldsymbol{\beta}_j$ in a way that the gap between $\boldsymbol{\eta}$ and $\mathbf{y}$ is minimum.

Ridge regression best linear unbiased prediction (rrBLUP) is a commonly used mixed linear model in genomic prediction, assuming that the effect of each marker is small and normally distributed with the same variance regardless of the size of the actual effect size in each genetic marker (Meuwissen *et al.* 2001; Bernardo and Yu 2007). Hence, $\boldsymbol{\beta}_j$ is distributed as $\boldsymbol{\beta}_j \sim N(0, \mathbf{I}, \sigma^2_{\beta_j})$ where $\mathbf{I}$ is an identity matrix and $\sigma^2_{\beta_j}$ is the variance of genomic marker effects (Endelman 2011; Endelman and Jannink 2012). Genomic best linear unbiased prediction (GBLUP) (VanRaden 2008) is another linear mixed model emphasising the relationship between genetic markers for predictions by developing the genotypic relationship (kinship) matrix (Lipka *et al.* 2012; Wang *et al.* 2015). This method is demonstrated to be equivalent to rrBLUP and both models share the same mechanisms and assumptions (Habier *et al.* 2007). Hence, only the rrBLUP model implementation was chosen as one of the genomic prediction models for the ensemble investigations in this study.

Another group of linear mixed models used for genomic prediction analyses is the Bayesian methods often characterised by the Bayesian alphabet series (Gianola *et al.* 2009). In contrast to rrBLUP assuming that all the markers have normally distributed small effects, Bayesian methods allow each genomic marker effect $\boldsymbol{\beta}_j$ to have different distributions (Wang *et al.* 2018). This variation in the distribution is expected to capture more realistic genomic marker effects whenever it is unrealistic for all the genomic marker effects to have the same distribution with small effects, as is assumed in rrBLUP. Another noticeable characteristic of the Bayesian models is the application of prior distributions, constructed from prior knowledge, assumptions and statistics, iteratively updated by the accumulation of information from each sampling to capture the properties of the true distribution (Kruschke 2010). Among various Bayesian models, BayesB was selected in this study because it has been widely used as one of the standard models in genomic prediction (Abdollahi-Arpanahi *et al.* 2020; Farooq *et al.* 2022; John *et al.* 2022; Meher *et al.* 2022; Plavšin *et al.* 2022). For BayesB some genomic marker effects are expected not to be influential on target phenotypes and their genomic marker effects are set as zero.

We developed rrBLUP and BayesB models using the library BGLR (Pérez and de Los Campos 2014) in R. For parameter setting, we set the number of iterations and burn-in as 12,000 and 2,000 respectively for both models. Other parameters were set as default throughout this study.

Semiparametric models: Semiparametric models possess both parametric and nonparametric properties, indicating that the models can capture both linear and nonlinear relationships from two different approaches. One of the semiparametric models that has been widely leveraged for genomic prediction is a Reproducing Kernel Hilbert Space (RKHS) regression model introduced by Gianola and van Kaam (2008). The original idea of RKHS was proposed by Aronszajn (1950), which mapped given data into Hilbert space to capture complex nonlinear relationships among the data points. Any relationship that may not be captured in the original dimension can be observed on complex hyperplanes. Gianola and van Kaam (2008) leveraged RKHS to detect nonlinear effects represented as dominance and epistatic genetic effects, while maintaining the parametric components to capture the additive genetic effects. RKHS can be formulated as Howard

et al. (2014) denoted with the base of Equation 1:

$$\boldsymbol{\eta} = 1\boldsymbol{\mu} + \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j + \sum_{j=1}^{J} g(\mathbf{X}_j) + \sum_{l=1}^{L} \mathbf{u}_l \tag{2}$$

where $g(\mathbf{X}_j)$ represents genetic effects from nonlinear genomic marker effects such as dominance and epistatic effects. $g(\mathbf{X}_j)$ is expressed as

$$g(.) = \alpha_0 + \sum_{n=1}^{N} \alpha_n K(., x_{jn}) \tag{3}$$

where $X_j = \{x_{j1}, x_{j2}, ...x_{jn}\}$, α_0 is a fixed term, α_n is coefficient accompanied to x_n and K is reproducing kernel. These equations imply that the RKHS model consists of both additive and non-additive components, aimed at capturing genomic marker effects in a more comprehensive way. Similarly, the BGLR library was leveraged for the RKHS model. The number of iterations and burn-in are also set as 12,000 and 2,000 and the rest of the parameters remained as default as well.

Machine learning models

Machine learning (nonparametric) models do not require any assumptions about the underlying distributions of the model terms. Instead, the parameters of the models are determined by an iterative training process. From the various machine learning (ML) methods, we selected random forest (Breiman 2001a), support vector regression (Drucker *et al.* 1996) and graph attention network (Veličković *et al.* 2017) for our investigation of ensemble prediction.

A random forest (RF) contains a collection of decision trees (Belson 1959) trained by sub-train sets sampled from the original training set (Liu *et al.* 2012). Decision trees develop a tree-like decision mechanism flow, consisting of nodes and edges. After two edges are released from the top node called the root, each layer consists of nodes with one incoming and two outgoing edges except the end layer nodes called leaves which have no outgoing edges (Rokach and Maimon 2005). Each node except the leaves holds a condition based on values in a specific feature from the given data. The algorithm starts traversing from the root, and if the target data point (a set of features) satisfies the condition of the root, the algorithm traverses the tree to the left bottom adjacent node and the right bottom adjacent node in vice-versa. This process repeats until it reaches the leaves determining the class or value of the target data point (de Ville 2013). When a target data point is given, each decision tree returns a prediction value, and the final prediction value is determined by aggregating the prediction results from all the trees. Using different sub-train sets for training respective decision trees can cancel out prediction noise from each tree, resulting in more stable prediction results (Qi 2012). RF was chosen as a model because it is one of the commonly used ML models in genomic prediction (González-Camacho *et al.* 2018; Sandhu *et al.* 2021a; Farooq *et al.* 2022; John *et al.* 2022).

A support vector regression (SVR) is another ML approach that draws a hyperplane between data points for continuous target values. A hyperplane is drawn in a way that the distance between the hyperplane and the closest data points (support vector) becomes maximum with minimum prediction errors that include the largest number of data points within the range of the decision boundary. This is formalised by minimising the following objective function under several constraints (Drucker *et al.* 1996):

$$\begin{aligned} & U(\sum_{i=1}^{N} \xi_i^* + \sum_{i=1}^{N} \xi_i) + \frac{1}{2}(\mathbf{w}^t \mathbf{w}) \\ s.t. \quad & y_i - (\mathbf{w}^t \mathbf{v}_i) - b \leq \epsilon + \xi_i \\ & (\mathbf{w}^t \mathbf{v}_i) + b - y_i \leq \epsilon + \xi_i^* \\ & \xi_i^* \leq 0 \\ & \xi_i \leq 0 \end{aligned} \tag{4}$$

where U is an objective function targeted for the constraints, b is an intercept, $\mathbf{w}$ is a coefficient vector for a feature vector, ϵ is the distance between a hyperplane and a decision boundary, ξ_i^* and ξ_i are slack variables functioned to make the decision boundary "soft" by allowing some data points to be outside the upper and lower boundary. SVR can be equipped with a kernel, mapping data points to another dimension for capturing nonlinear interactions. SVR has also been leveraged for various experiments in genomic prediction as a standard method (An *et al.* 2021; Yu *et al.* 2021; John *et al.* 2022; Li *et al.* 2023) and is selected as one prediction model in this study.

A graph attention network (GAT) is a graph neural network that applies a self-attention mechanism for predictions. We apply GAT by converting the relationship between markers and phenotypes into a graphical format. Genetic markers and phenotypes can be represented as nodes, and the connections between genetic marker nodes and phenotype nodes are represented as edges. The edges are directed from the genetic marker nodes to phenotype nodes, showing that genetic markers explicitly affect phenotypes. In this study, we do not add connections between marker nodes because allowing the edges between them did not improve the prediction performance and resulted in the exponential increase of computational time. We leverage the GAT model proposed by Brody *et al.* (2021). The attention mechanism can be written as below:

$$\alpha_{ij} = \frac{LeakyReLU(\mathbf{a}^\top [\mathbf{W}\boldsymbol{h}_i || \mathbf{W}\boldsymbol{h}_j])}{\sum_{j' \in \mathcal{N}_i} LeakyReLU(\mathbf{a}^\top [\mathbf{W}\boldsymbol{h}_i || \mathbf{W}\boldsymbol{h}_{j'}])} \tag{5}$$

where $\mathbf{a}^\top \in R^{2d'}$ is a transposed weight vector, d' is the number of features in each node, $\mathbf{W}$ is a weight matrix, $\boldsymbol{h}_i = \{h_1, h_2, ..., h_N\}$ is a set of node features of node i, $||$ concatenates vectors, $j' \in \mathcal{N}_i$ is a partial neighbour of nodes i sampled from the entire neighbours. The feature

values of neighbours and trainable weights are nonlinearly activated to generate the attention values, and the calculated attention value is normalised at the end. This attention calculation mechanism is repeated for K times to stabilise the calculation result. It is called multi-head attention and is formulated as below (Veličković *et al.* 2017):

$$\boldsymbol{h}_i^{'} = \mathop{\Big\|}_{k=1}^{K} \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \boldsymbol{h}_j \right) \tag{6}$$

where K is the number of total attention layers, σ is a nonlinear activate function and $\boldsymbol{h}_i^{'}$ is the updated node features. At the final layer, the result from each attention layer is averaged instead of concatenation:

$$\boldsymbol{h}_i^{'} = \sigma \left(\frac{1}{K} \sum_{k=1}^{K} \sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \boldsymbol{h}_j \right) \tag{7}$$

This final value $\boldsymbol{h}_i^{'}$ is used as a predicted phenotype from the model. GAT was chosen among graph neural networks due to its attention mechanism which can more precisely capture key prediction patterns underlying the given data to enhance prediction performance.

For RF and SVR, Sklearn (v1.2.2) was used for the model implementation in Python. The number of trees in RF was set as 1,000 while the default setting was used for other hyperparameters. In SVR, the radial basis function (RBF) was used for the kernel and the default setting was used for the remaining parameters. For GAT, Pytorch Geometric (v2.3.0) was used. Since phenotype and genetic marker nodes contained different types of information, they needed to be identified as different node types. Hence, the graph was converted into a heterogeneous graph. The model was three-layered with one hidden layer with 20 channels and a dropout rate of 0.9 was applied to every layer. The Exponential Linear Unit (ELU) function was used for the activation functions. The total number of heads was set as 8. The model was trained with 150 epochs by mini-batched graphs with a batch size of 8. AdamW was selected for the optimiser with a learning rate of 0.005 and weight decay of 0.00005.

Naïve ensemble-average model

Ensemble models combine the predicted phenotypes from multiple individual genomic prediction models. Firstly, each individual genomic prediction model output predicted phenotypes from the given genetic markers. These predicted phenotypes from the individual genomic prediction models were assembled as a predicted phenotype matrix. Each row contained predicted phenotypes from each individual genomic prediction model for a specific target RIL. This matrix was used as input data for the ensemble model. The ensemble can be formulated as below:

$$\boldsymbol{\eta}^{'} = f(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, ..., \boldsymbol{\eta}_m) \tag{8}$$

where $\boldsymbol{\eta}^{'} = \{\eta_1^{'}, \eta_2^{'}, ..., \eta_n^{'}\}$ is a vector of the final predicted phenotypes, each $\boldsymbol{\eta}$ represents predicted phenotypes from each individual genomic prediction model and $f(\cdot)$ is the ensemble function. Hence, six diverse predicted phenotypes were generated per RIL and were assembled to develop a predicted phenotype matrix for the ensemble model in our study.

As per the actual ensemble function f(·) a naïve averaging method was used in this study, by taking the arithmetic average of the predicted phenotypes from all the individual genomic prediction models with equal weights. The naïve ensemble can be formulated as below:

$$f(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, ..., \boldsymbol{\eta}_m) = \frac{\sum_{m=1}^{M} \boldsymbol{\eta}_m}{M} \tag{9}$$

where M is the total number of models. Hence, M is set as 6 for this study.

Genomic marker effect estimation

The extraction of estimated genomic marker effect values from each model can suggest how each model estimates the genomic marker effect of respective SNPs for predicting target phenotypes, allowing a comparison of the genomic prediction models at the genomic level. For rrBLUP and BayesB, the genomic marker effects were extracted using $\boldsymbol{\beta}$ which represents allele substitution effect.

For RKHS and SVR, the Shapley value (Shapley *et al.* 1953) was employed to estimate genomic marker effects. The Shapley value is a metric, originally in the field of game theory, to assess the equitable distribution of resources and rewards to players based on their contribution level under cooperative scenarios (Winter 2002). A larger Shapley value is allocated to SNPs causing larger changes in predicted values. The SNP Shapley value is calculated as below (Lundberg and Lee 2017):

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S \bigcup \{i\}}(x_{S \bigcup \{i\}}) - f_S(x_S)] \tag{10}$$

where ϕ_i is a Shapley value for feature i (SNP_i in this case), S is a subset of features F, $f_{S \bigcup \{i\}}$ is a model predicting a target value with a feature subset including i, f_S is a model predicting the target value without including i. In this study, the Shapley value, a genomic marker effect of an SNP, is estimated as the average gap in predicted phenotypes between the cases where the target SNP is included and the case where the target SNP is excluded. For RKHS, the Shapley value was implemented using iml (v0.11.3) (Molnar *et al.* 2020) in R whereas SHAP (v0.42.1) (Lundberg and Lee 2017) was applied to SVR in Python. Since the Shapley value returns element-wise feature effects, the final Shapley value of each SNP is calculated by averaging the Shapley values for the target SNP from RILs in a test set.

For RF, genomic marker effects were estimated by extracting impurity-based feature importance values, measuring the importance of features X by splitting the data using a value s from X at node t formulated as below (Ishwaran 2015):

$$\hat{\Delta}(s,t) = \hat{\Delta}(t) - [\frac{N_L}{N}\hat{\Delta}(t_L) + \frac{N_R}{N}\hat{\Delta}(t_R)] \tag{11}$$

where $\hat{\Delta}(t)$ represents impurity at node t, N is the total number of data points in daughter node t_L and t_R, $\frac{N_L}{N}$ is the weight of left daughter node t_L, $\frac{N_R}{N}$ is the weight of right daughter node t_R, $\hat{\Delta}(t_L)$ is impurity value of left daughter node t_L and $\hat{\Delta}(t_R)$ is impurity value of right daughter node t_R. The impurity is measured by the summed squared value at the gap between the mean value of the prediction target and the actual value of each element in each node. Since this denotes only the impurity gap at node t, the summation of the impurity gap across all nodes leads to the total impurity gap for a target feature SNP. The importance of the feature was extracted using Sklearn (v1.2.2) in Python.

The genomic marker effects of GAT were estimated by leveraging an interpretability method called Integrated Gradients (Sundararajan *et al.* 2017), integrating the gradient of a line drawn between two points; the baseline point shows the prediction value without the effect of the target feature SNP and the other indicates the prediction value with the feature SNP value. Using the baseline point, the true gradient can be calculated by eliminating the effect of the initial value. Integrated Gradients is calculated as below (Sundararajan *et al.* 2017):

$$\mathrm{IG}(x_i) \approx (x_i - x_i^{'}) * \sum_{k=1}^{m} \frac{\partial F(x_i^{'} + \frac{k}{m} * (x_i - x_i^{'}))}{\partial x_i} * \frac{1}{m} \tag{12}$$

where x is a feature, $x^{'}$ is a baseline value of x and m is the total number of interpolation steps of the integral. Similar to the Shapley value, Integrated Gradients also returns element-wise marker effect, and thus the final genomic marker effect was estimated by averaging the effect of the target SNP across RILs in a test set. Pyg (v2.4.0) was used for the implementation.

Assessment criteria

Experimental flow

The experimental framework for evaluation of the individual genomic prediction and the ensemble models is shown in Figure 1. This

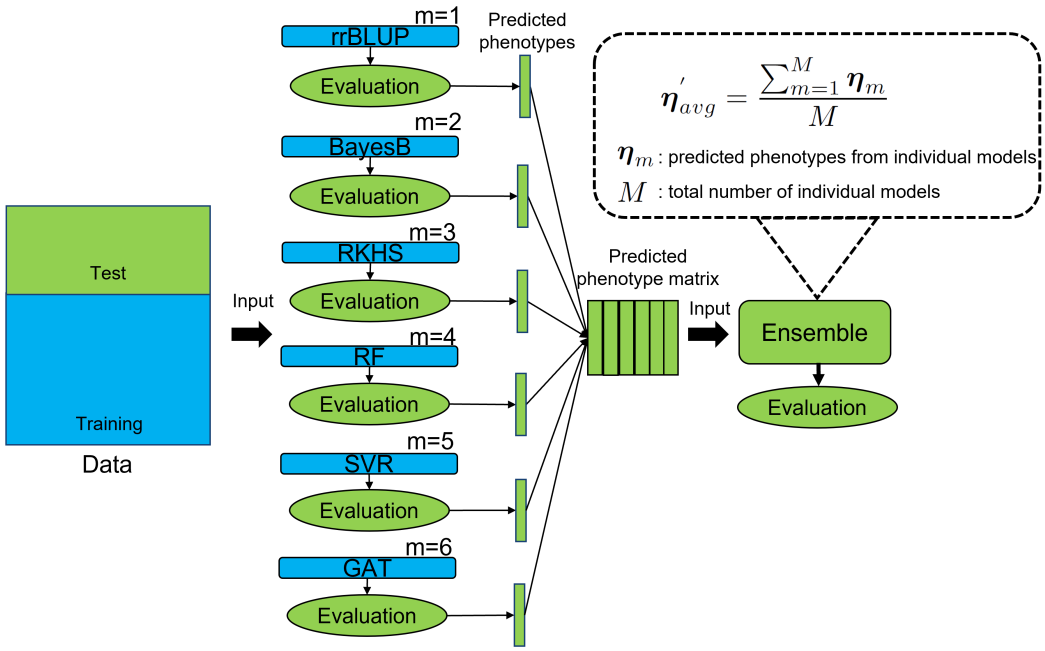

Figure 1 A diagram of the experimental flow used in this study. The blue and green colours represent components related to a training and test set respectively. Each individual genomic prediction model was trained using genetic markers. The performance of the trained individual genomic prediction models was evaluated using the test set. Predicted phenotypes for RILs in the test set from the individual genomic prediction models were used as input data for the ensemble model.

study evaluated the performance of the genomic prediction models under within-population prediction scenarios. Each population data in the TeoNAM dataset was randomly split into training and test sets. After SNPs were filtered based on the LD filtering, the training set was used to train the six individual genomic prediction models. The performance of the trained individual genomic prediction models was evaluated using the test set. Vectors of predicted phenotypes for RILs in the test set, derived from each individual genomic prediction model, were assembled to construct a predicted phenotype matrix. The ensemble genomic prediction model leveraged the predicted phenotype matrix as the input. After evaluating the prediction performance of the ensemble approach, the genomic marker effects from each of the six individual genomic prediction models were extracted by the interpretable approaches explained in Section Genomic marker effect estimation.

To ensure the generalisability of the observed result, we repeatedly implemented the genomic prediction models for several different settings. Three different training-test ratios (0.8-0.2, 0.65-0.35 and 0.5-0.5) were leveraged with a random sampling of 500. Hence, each genomic prediction model was evaluated over 1,500 prediction results (3 ratios*500 samples) in each population-trait combination.

Metrics

The Pearson correlation between the predicted and observed phenotypes for the test set is used as one metric to evaluate the performance of the individual and ensemble genomic prediction models, formulated as below:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{13}$$

where x_i is the predicted phenotype, y_i is the observed phenotype for a RIL in the test set and $\bar{x}$ and $\bar{y}$ are the mean values of x and y, respectively. Pearson correlation ranges from 1 to -1. The value 1 indicates that all the ranks of predicted phenotypes are concordant with that of observed phenotypes. Thus, the correlation value of a genomic prediction model with higher prediction performance should be close to 1.

The second metric used to assess the individual and ensemble genomic prediction models is the mean squared error (MSE) between the predicted and observed phenotypes. The MSE measures the average of squared differences between the predicted and observed phenotypes. MSE can be formulated as below:

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{14}$$

where y_i is the observed phenotypes and $\hat{y}_i$ is the predicted phenotypes. The difference between predicted and observed phenotypes is squared to assess the cumulative effect of positive and negative deviation values. A model with a smaller MSE is regarded as a better model because it suggests that on average there is a smaller difference between the predicted and observed phenotypes.

Results

Ensemble of models outperformed the best individual genomic prediction models

The naïve ensemble-average model demonstrated higher prediction accuracies (measured in Pearson correlation) compared with individual genomic prediction models for both traits (Figure 2). In contrast, the lowest prediction errors (measured in MSE) were achieved by either BayesB or the ensemble, depending on the target traits. These results demonstrated the overall outperformance of the ensemble over the individual genomic prediction models. There was no individual genomic prediction model that was consistently the "best" across all the training-test sets. The same trend was observed at the per-population level as well (Figure S1); Across the five RIL populations, the naïve ensemble-average model reached the highest prediction accuracy while either BayesB or the naïve ensemble-average model had the lowest prediction errors.

For the DTA trait, the best genomic prediction model depended on which metric (prediction accuracy or prediction error) was used for the performance evaluation. Prediction accuracy was highest for the naïve ensemble-average model (median = 0.920) with the best individual genomic prediction model being BayesB (median = 0.888). In contrast, BayesB reached the lowest prediction errors (median = 7.765) among all the genomic prediction models including the naïve ensemble-average model (median = 9.340). These results indicate that the naïve ensemble-average model can be the preferable genomic prediction model when prediction accuracy is the primary metric to optimise while BayesB can be selected when prediction error is the main target metric.

For the TILN trait, the highest prediction accuracies and lowest prediction errors were observed with the naïve ensemble-average model. The highest prediction accuracy was observed in BayesB (median = 0.709) within the individual genomic prediction models and the naïve ensemble-average model surpassed it (median = 0.797). BayesB reached the lowest prediction error among the individual genomic prediction models (median = 0.297), but the lower prediction error was observed in the naïve ensemble-average model (median = 0.257). These results indicate the naïve ensemble-average model was the most suitable genomic prediction model for the TILN trait.

Improved prediction performance with an ensemble compared to the average of individual genomic prediction models

The naïve ensemble-average model outperformed the average of individual genomic prediction models across both traits. Figure 3 demonstrated higher prediction accuracies and lower prediction errors of the ensemble. For the DTA trait, the average median prediction accuracy of the individual genomic prediction models (rrBLUP, BayesB, RKHS, RF, SVR and GAT) was 0.719 while the median prediction accuracy of the naïve ensemble-average model was 0.919. The average of the median MSE of the individual genomic prediction models was 16.838 while the median MSE of the naïve ensemble-average model was 10.167. For the TILN trait, the average median prediction accuracy of the individual genomic prediction models was 0.671 while the median prediction accuracy of the naïve ensemble-average model was 0.790. The average of the median MSE of the individual genomic prediction models was 0.356 while the median MSE of the naïve ensemble-average model was 0.277.

Classical genomic prediction models outperform ML models

Figure 2 indicates that classical genomic prediction models (rrBLUP, BayesB and RKHS) demonstrated higher prediction accuracies and lower prediction errors than ML models (RF, SVR and GAT). However, the magnitude of differences was trait-dependent. For the DTA trait, SVR had considerably lower median Pearson correlations (0.360) and higher median MSE (29.811) than other individual genomic prediction models. RF and GAT demonstrated comparable prediction accuracies and errors to RKHS with median Pearson correlations (0.778, 0.818 and 0.806) and median prediction errors (14.274, 14.537 and 12.976). rrBLUP and BayesB demonstrated the highest median prediction accuracies (0.878 and 0.888) and lowest median prediction errors (8.308 and 7.765) of all the individual genomic prediction models. For the TILN trait, smaller performance differences between the classical genomic prediction models and ML models were observed in both prediction accuracy and error. SVR demonstrated the lowest median Pearson correlations (0.650) and higher median MSE errors (0.368)

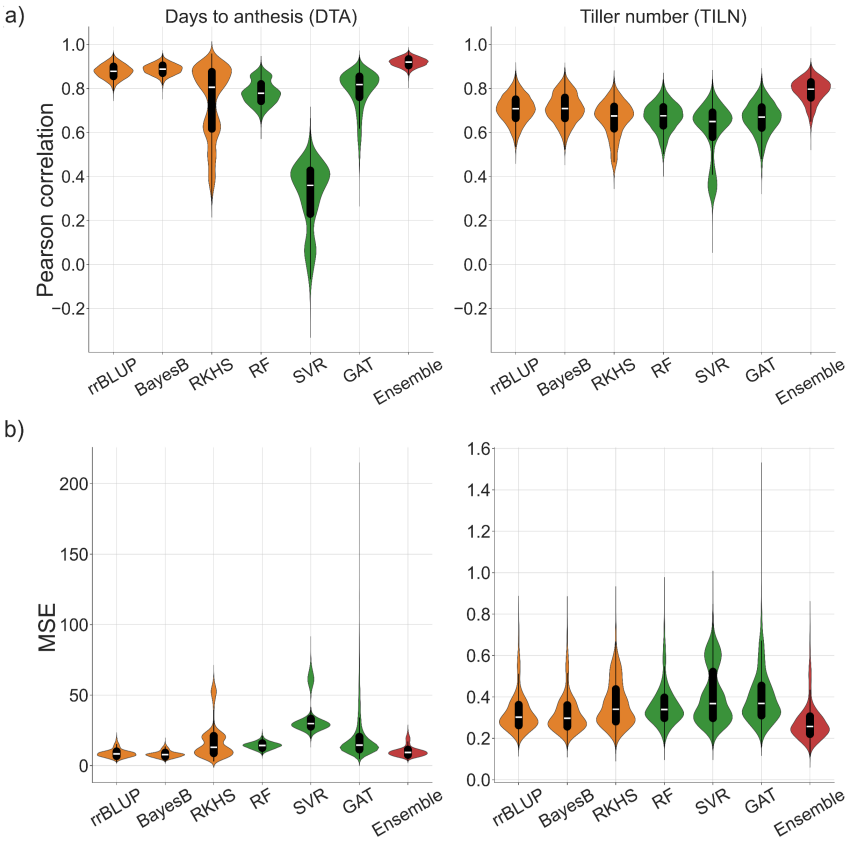

Figure 2 A comparison of genomic prediction performance of the ensemble (naïve ensemble-average) model versus each of the individual genomic prediction models in violin plots. The width of the violins indicates the distribution of the metric values. The performance of genomic prediction models was measured with a) the Pearson correlation and b) mean squared error (MSE). The orange represents the performance of classical models (rrBLUP, BayesB and RKHS) while the green represents ML models (RF, SVR and GAT). The red is the performance of the ensemble. The black vertical bars with whiskers illustrate box plots and the white horizontal lines on the box plots are the median value.

compared to the other individual genomic prediction models. RF and GAT demonstrated comparable prediction accuracies and errors to RKHS with median Pearson correlations (0.676, 0.670 and 0.675) and median MSE errors (0.339, 0.369 and 0.340). rrBLUP and BayesB demonstrated the highest median prediction accuracies (0.708 and 0.709) and lowest prediction errors (0.302 and 0.297) of all the individual genomic prediction models.

No consistent winner amongst individual genomic prediction models

Individual genomic prediction models performed similarly overall across traits and samples of training-test sets. The individual genomic prediction model with the highest prediction accuracies varied across the three training-test set sizes while the individual genomic prediction model with the lowest prediction errors varied across the two traits (Table 2). These results suggest the absence of a "best" individual genomic prediction model.

Table 2 Percentage of best performance achieved by each genomic prediction model in the respective training-test ratio (0.8, 0.65, 0.5) and trait (DTA, TILN) combinations. The performance was measured by Pearson correlation and MSE. The value reaching the highest percentage in each combination is highlighted in bold.

Ratio	Trait	Pearson correlation							MSE						
		rrBLUP	BayesB	RKHS	RF	SVR	GAT	Ensemble	rrBLUP	BayesB	RKHS	RF	SVR	GAT	Ensemble
0.8	DTA	1.24	0.44	0.40	0.00	0.00	0.00	**97.92**	**39.32**	31.92	15.64	0.08	0.00	0.00	13.04
	TILN	0.00	0.40	0.00	0.00	0.04	0.00	**99.56**	2.84	9.28	0.00	0.04	0.24	0.04	**87.56**
0.65	DTA	0.24	1.08	0.00	0.00	0.00	0.00	**98.68**	18.12	**72.00**	0.12	0.04	0.00	0.00	9.72
	TILN	0.00	0.24	0.00	0.00	0.08	0.00	**99.68**	1.00	13.64	0.00	0.00	0.08	0.04	**85.24**
0.5	DTA	0.04	2.08	0.00	0.00	0.00	0.00	**97.88**	4.44	**94.12**	0.00	0.20	0.00	0.00	1.24
	TILN	0.00	0.24	0.00	0.00	0.00	0.00	**99.76**	0.24	16.04	0.00	0.00	0.00	0.00	**83.72**

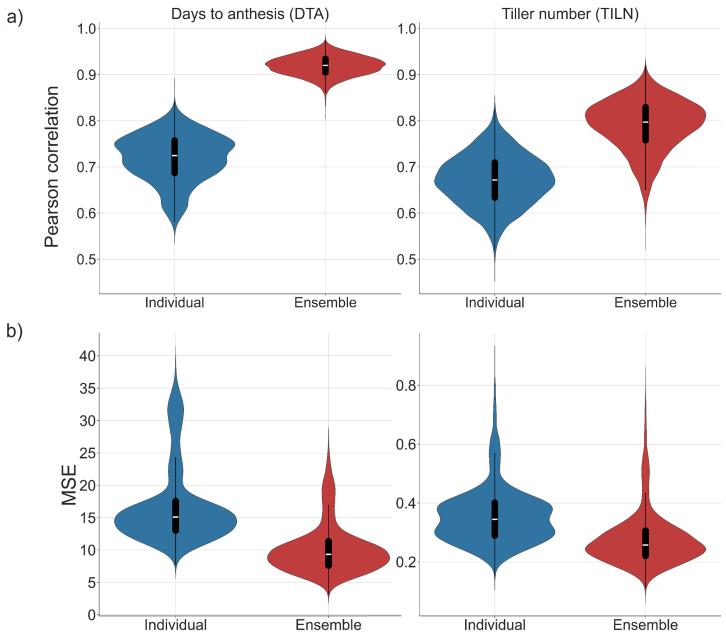

Figure 3 Violin plots comparing genomic prediction performance of the individual (average of individual genomic prediction models) in the blue versus the ensemble (naïve ensemble-average model) in the red. The performance of genomic prediction models was measured with a) the Pearson correlation and b) mean squared error (MSE). The width of the violins represents the distribution of metric values and the black boxes with whiskers are box plots. Horizontal lines within the boxes represent the median value.

In the individual genomic prediction models, BayesB maintained the highest prediction accuracy percentage across training-test sets for the three training-test sizes (DTA ranged between 0.44% and 2.08% and TILN between 0.24% and 0.40%) while the second highest prediction accuracy percentage in the individual genomic prediction models was rrBLUP with the range of 0.04% and 1.24% for the DTA trait and 0.00% for the TILN trait. The highest prediction accuracy percentage was observed with the other individual genomic prediction models (RKHS, RF, SVR and GAT) in less than 1.00% of training-test set samples.

The individual genomic prediction model with the highest percentage of lowest prediction errors depended on the combination of training-test set size and traits. For the DTA trait, the lowest prediction error percentage was the highest in rrBLUP (39.32%) when the training-test set ratio was 0.8, but BayesB was the highest when the size of the training set became smaller (72.00% for 0.65 and 94.12% for 0.5). For the TILN trait, the lowest prediction errors among the individual genomic prediction models were observed with BayesB across all ratios of training-test set sizes (from 9.28% to 16.04%).

Large variation in the genomic marker effects estimated by individual genomic prediction models

Large magnitudes of variation were observed in genomic marker effects across the classical and ML genomic prediction models. Figure 4 presents pairwise comparisons of predicted phenotypes (top right triangle) and genomic marker effects (bottom left triangle) between the individual genomic prediction models. Positive associations were observed in pairwise comparisons between individual genomic prediction models in several pairs for predicted phenotypes. On the other hand, such a positive association was not constantly found in the genomic marker effects. Only a few associations were observed in pairwise comparisons between individual genomic prediction models at the genomic marker effects level. Each individual model recognised different SNP markers as highly effective on the trait, causing diversity in the estimated effect size for each SNP marker. While there was variation in the genomic marker effects estimated by the individual genomic prediction models, the SNPs identified as putative QTL by Chen *et al.* (2019) were frequently included as features in the individual genomic prediction models (Figure 4). Therefore, while the QTL identified by Chen *et al.* (2019) were frequently included in the genomic prediction models, the weights of their contributions to the prediction of the trait phenotypes of the RILs differed among the genomic prediction models.

For the predicted phenotypes, the strength of the associations did not significantly vary among the individual genomic prediction models as demonstrated in Table S1. Strong positive associations were observed among the classical genomic prediction models. The high Pearson correlations were observed between the predicted phenotypes of rrBLUP and BayesB, rrBLUP and RKHS and BayesB and RKHS. Similarly, weaker but positive associations were observed among the predicted phenotypes of ML models. The high Pearson correlations were observed between the predicted phenotypes of RF and SVR, RF and GAT and SVR and GAT. Furthermore, strong positive associations were observed among the predicted phenotypes of the classical and ML genomic prediction models.

For the genomic marker effects, in contrast, a lack of association was consistently observed among the individual genomic prediction models. While a relatively high association was observed between rrBLUP and BayesB, the majority of the other pairs showed a weak association by not reaching a Pearson correlation higher than 0.5 across the classical, ML and mixed genomic prediction models in both traits.

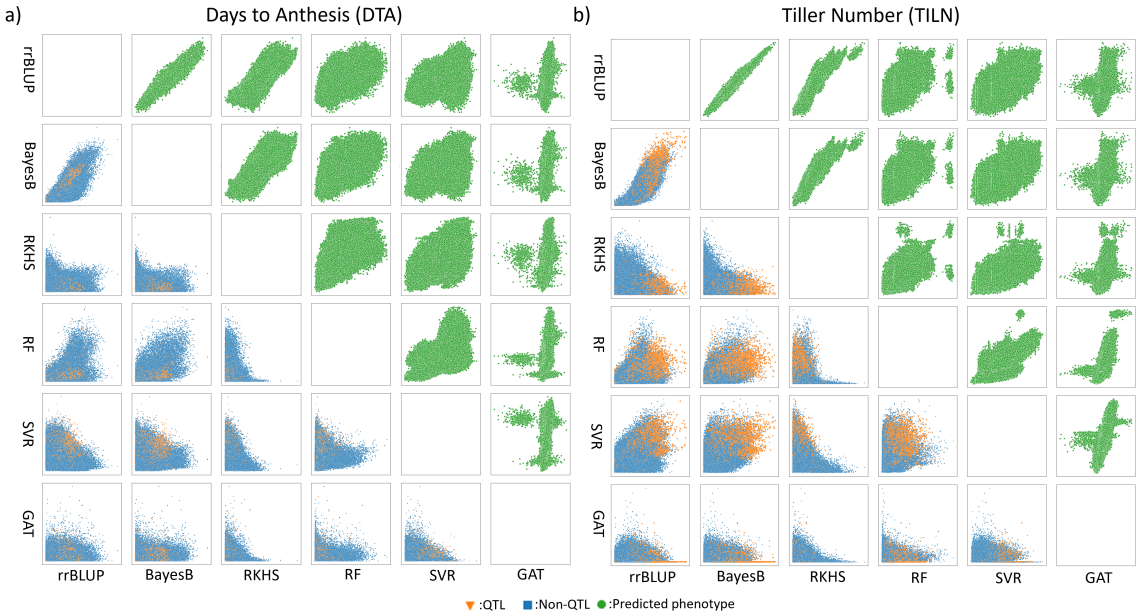

Figure 4 Pairwise comparison of individual genomic prediction models at predicted phenotypes (top right triangle) and genomic marker effects (the bottom left triangle) levels for both traits; a) the days to anthesis (DTA) and b) the tiller number (TILN). The green circle dots represent a pair of predicted phenotypes for RILs included in the test set in each sample scenario. The blue square and orange triangle dots indicate a pair of estimated genomic marker effects in each sample scenario classified as non-QTL and QTL by Chen *et al.* (2019) respectively.

Discussion

No consistent winner amongst individual genomic prediction models

The lack of a consistent winner among the individual genomic prediction models (Figure 2 and S1) poses the relevance of the No Free Lunch Theorem (Wolpert and Macready 1997) for genomic prediction problems. For the DTA trait, while a slight outperformance by rrBLUP and BayesB was observed, all the individual genomic predictions, except SVR, showed similar prediction accuracy and error. The absence of the best individual genomic prediction model was also emphasised in the TILN trait by not showing clear differences in prediction accuracy and error between the individual genomic prediction models. The theorem was also implied from the positive association between genomic prediction pairs at the level of the predicted phenotype (Figure 4). Each individual genomic prediction model returned similar predicted phenotypes for the same RILs and, consequently, the individual genomic prediction models showed a similar prediction performance. The distinctive prediction algorithms of the individual genomic prediction models did not significantly diversify their prediction performance.

The lack of a consistent winner among individual models has been observed for prediction problems in other fields. For instance, Fernández-Delgado *et al.* (2014) tested 179 prediction models over 121 datasets and concluded that RF achieved the overall highest prediction performance. However, no statistical performance superiority of RF to the second best (support vector machine leveraging Gaussian kernel) was detected in their experiment, indicating that the performance difference between the two was subtle. The results from Gómez and Rojas (2016) also showed that no individual model clearly outperformed the others. The absence of a prediction model that was immune to all the negative factors (noise from datasets, data imbalance and dissatisfaction with model assumptions) was considered to be the cause of no outperformance. Similarly, other research (Merrick and Carter 2021; Plavšin *et al.* 2022) suggested the difficulty in finding the best individual model. These results indicate that some prediction models can outperform others in particular scenarios but are not all-rounders.

Therefore, we argue that it is not strategic to focus on developing an individual genomic prediction model for a diverse range of genomic prediction tasks. The ensemble approach can be one solution to overcome the limitations of individual prediction algorithms postulated by the No Free Lunch Theorem.

Ensemble improved prediction performance

One clear consensus, from the results of this study, is the improved predictive ability of the naïve ensemble-average model compared to the individual genomic prediction models. For the DTA and TILN traits, the prediction accuracy of the naïve ensemble-average model was the highest in more than 95% of prediction cases. Although the naïve ensemble-average model did not achieve the lowest MSE for the DTA trait, it was almost equivalent to rrBLUP and BayesB that achieved the lowest MSE (Figure 2 and Table 2). When comparing against the average of the individual genomic prediction models, the naïve ensemble-average model consistently outperformed the individual genomic prediction models on the basis of prediction accuracy and error (Figure 3). These results suggest an opportunity to improve prediction performance with the naïve ensemble-average model.

The success of the naïve ensemble-average model is derived from the diversity of information contributed by multiple, individual genomic prediction models. The range of associations among the estimated genomic marker effects of the individual genomic prediction models (Figure 4) illustrates the diversity of information available to the ensemble model. The range of associations among the estimated genomic marker effects is a result of algorithmic differences that differentially weight predictive features from the same input data across the individual genomic prediction models. This phenomenon is called the Rashomon effect (Breiman 2001b), which states that there is a set of models (Rashomon sets) that captures different effects of features from the same datasets due to distinct properties of the prediction algorithms. Ensemble models can combine these alternative views of the prediction problem to generate a more comprehensive representation of the prediction problem. In the case of this study, it was represented as a more comprehensive view of trait genetic architecture. Hence, information diversity from individual genomic prediction models can be a core factor behind the improved prediction performance of the naïve ensemble-average model.

Although the set of the six individual genomic prediction models chosen is one of many possibilities, their contrasting algorithms for estimating genetic effects likely contributed to the diversity in genomic marker effect estimates. For example, the parametric models (rrBLUP, BayesB) primarily target capturing SNP main effects. In contrast, the semi-parametric model RKHS provides a mechanism for explicit consideration of SNP interaction terms in addition to SNP main effects. The nonparametric ML models primarily focus on SNP interactions by considering nonlinear relationships between SNPs. Each distinctive prediction algorithm develops a unique hypothetical space and true value can be outside the spaces. Predictions outside the hypothetical spaces of individual genomic prediction models can be feasible by creating a new hypothetical space through the ensemble of the individual genomic prediction models (Johnson and Giraud-Carrier 2019; Dietterich 2000). The newly constructed hypothetical space can be wider than the individual genomic prediction models by supplementing each other as the synergetic result. Complementing respective errors derived from the uniqueness of each genomic prediction model with the ensemble enables the construction of solutions for complex tasks with higher performance (Dong *et al.* 2020; Kick and Washburn 2023).

The outperformance of the naïve ensemble-average model could also be viewed as avoiding stagnation in local optima. Each of the six genomic prediction modelling algorithms may achieve different local optima within their possible prediction space, the size of which depends on the given data and algorithms. Algorithms trapped in their local optima can inhibit opportunities to explore the broader problem state space that is closer to the global optima, where more precise predictions may be achieved. Applying another algorithm on top of a set of individual genomic prediction models can increase the likelihood of shifting from the local to the global optima that are infeasible to reach individually. For example, Wu *et al.* (2019) discussed the benefits of using an ensemble concept in population-based optimisation approaches (a method containing a number of adaptive prediction models to find global optima), suggesting that the global optima can be efficiently discovered by cooperatively sharing information from each prediction model rather than using them independently. The advantage of discovering global optima by an ensemble has also been mentioned in evolutionary algorithms (iterative model improvement adaptively done with observed prediction results) (Yu and Suganthan 2010) and metamodeling (representation of a model to a more simple mechanism compared to the original one) (Ferreira and Serpa 2018) as well. In short, the ensemble helps reach global optima using information derived from local optima.

Prediction performance depends on the complexity of the network affecting a trait

The complexity of an underlying biological network controlling a target trait can be regarded as one factor attributing to differences in the prediction performance. While networks of genes control the DTA trait (Buckler *et al.* 2009; Dong *et al.* 2012), approximate models based on contributions of additive effects appear to have been sufficient to account for the phenotypic diversity. In contrast, the TILN trait can be a consequence of the existence of nonlinear complex marker interactions involved in the genetic architecture of the branching network (Doebley *et al.* 1995; Bertheloot *et al.* 2020; Powell *et al.* 2022). The developed genomic prediction models may lack a mechanism to comprehensively capture key prediction patterns in an intricate network. Azodi *et al.* (2019) compared the prediction performance of parametric (rrBLUP and Bayesian) against nonparametric ML models (RF, SVR and neural networks) for complex traits such as crop yield and plant heights in several crop types. The experimental results indicated that the genomic prediction models performed similarly at lower prediction accuracy for many traits. The high complexity of a gene network underlying a target trait can deteriorate the performance of the genomic prediction models.

Such a difference in the network complexity can also be implied by comparing performance between the individual genomic prediction models. The prediction performance of the nonparametric ML models (RF, SVR and GAT) in the DTA trait was lower than the parametric genomic prediction models (rrBLUP and BayesB), especially for SVR, but such a performance disadvantage diminished in the TILN trait. The different prediction performances in the parametric and ML models might be explained by the way they capture prediction patterns from the given data. The ML models primarily capture nonlinear prediction patterns (Ryo and Rillig 2017). The lower performance of the ML models in the DTA trait may have been a consequence of their inability to construct simpler models that are mainly aimed at detecting additive effects, resulting in overfitting to training set (Hawkins 2004). Hence, the rrBLUP and BayesB, focused primarily on capturing additive prediction patterns, may have outperformed the ML models. In contrast, since the TILN trait can be complex with more nonlinear genetic interactions, the performance of the parametric models might have decreased while that of the ML models might have been relatively improved in the TILN trait. However, the ML models may not have been able to capture the entire interactions in the complex network using a small amount of training data, and thus the ML models still performed similarly compared to the parametric genomic prediction models. This observation indicates that network complexity can be a crucial factor affecting the performance of genomic prediction models.

Future Opportunities

Several components can be considered to improve the prediction performance of the genomic prediction models in our experiment; hyperparameter tuning, data preprocessing methods and weight optimisation. Below, we briefly discuss each of these components.

In this experiment, hyperparameters of the genomic prediction models have been tuned heuristically rather than through systematic investigation. Systematic hyperparameter tuning can be conducted by several approaches such as cross-validation and Bayesian methods. Other research (Sandhu *et al.* 2021b; Kick *et al.* 2023) has demonstrated prediction performance increase with hyperparameter tuning. Due to the computational limitations and the data size (the total number of RILs), the hyperparameter tuning was not implemented. The default hyperparameter values were primarily leveraged instead. Hence, optimising hyperparameters may improve prediction performance for both individual and ensemble genomic prediction models by overcoming computational and data limitations.

As per the data preprocessing method, the simplicity of the mode-based imputation method may not lead to precise imputation. Missing values were replaced with the most frequent allele in each SNP in our experiment. However, the most frequent allele may not always be the most suitable candidate for the imputation of missing information. More complex imputation methods may enhance the imputation accuracy (Jubair and Domaratzki 2023).

Lastly, the performance improvement of the genomic prediction models can be expected by averaging predicted phenotypes with optimised weights. In this experiment, the predicted phenotypes were "naïvely" averaged from the six individual genomic prediction models by allocating the same weight to each. In other words, the six individual genomic prediction models contributed equally to the ensemble averaging step. However, this is one of many possible weighting approaches. In some applications, higher prediction performance was achieved by tuning weights accompanying each individual genomic prediction model based on their prediction performance (Liang *et al.* 2021; McCormick *et al.* 2021; Yu *et al.* 2021; Wang *et al.* 2023). Model selection can be one extreme scenario where the weight of some individual genomic prediction models is allocated as zero based on their contribution level, showing the performance improvement of prediction models in several cases (Zhou *et al.* 2002; Li *et al.* 2004; Huang and Wei 2022). Therefore, weight optimisation can strengthen the power of an ensemble.

Conclusion

We investigated the prediction performance of the naïve ensemble-average model over the six individual genomic prediction models for the DTA and TILN traits in the TeoNAM dataset. Our results demonstrated an advantage for predictions based on the ensemble of the six individual genomic prediction models with an equal contribution level. Therefore, ensemble approaches could be a promising tool for genomic prediction. We reason that the advantage of the ensemble of models for the TeoNAM study is derived from the diversity of the genomic prediction information captured by the individual genomic prediction models. The ensemble of such diverse information leads to the incorporation of more comprehensive genomic prediction information. Opportunities were identified for further research to investigate the effectiveness of ensemble approaches to other datasets. Our results suggest that the ensemble can improve selection accuracy, ultimately accelerating genetic gain in crop breeding programs.

Data and code availability

All the datasets leveraged in this study were collected by Chen *et al.* (2019) and full RILs and phenotype data are publicly available at panzea/genotypes/GBS/TeosinteNAM and panzea/genotypes/GBS/TeosinteNAM respectively. The code generated for this experiment is shared at https://github.com/ShunichiroT/ensemble.

Acknowledgments

We thank the National Computing Infrastructure (NCI) and Research Computing Centre (RCC) at the University of Queensland for supporting our experiment through High Performance Computing (HPC) machines.

Funding

This research was supported by the funding of the Australian Research Council through the support of the Australian Research Council Centre of Excellence for Plant Success in Nature and Agriculture (CE200100015)

Conflicts of interest

The authors have no relevant financial or non-financial interests to disclose

Literature cited

Abdollahi-Arpanahi R, Gianola D, Peñagaricano F. 2020. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genetics Selection Evolution. 52. https://doi.org/10.1186/s12711-020-00531-z.

An B, Liang M, Chang T, Duan X, Du L, Xu L, Zhang L, Gao X, Li J, Gao H. 2021. Kcrr: a nonlinear machine learning with a modified genomic similarity matrix improved the genomic prediction efficiency. Briefings in Bioinformatics. 22:bbab132. https://doi.org/10.1093/bib/bbab132.

Aronszajn N. 1950. Theory of reproducing kernels. The Transactions of the American Mathematical Society (TRAN). 68:337–404.

Azodi CB, Bolger E, McCarren A, Roantree M, de Los Campos G, Shiu SH. 2019. Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3: Genes, Genomes, Genetics. 9:3691–3702. https://doi.org/10.1534/g3.119.400498.

Bellman R. 1957. Dynamic programming princeton university press princeton. New Jersey Google Scholar. pp. 24–73. https//doi.org/10.1126/science.153.3731.34.

Belson WA. 1959. Matching and prediction on the principle of biological classification. Journal of the Royal Statistical Society. Series C (Applied Statistics). 8:65–75. https://doi.org/10.2307/2985543.

Bernardo R, Yu J. 2007. Prospects for genome-wide selection for quantitative traits in maize. Crop Science. 47:1082–1090. https://doi.org/10.2135/cropsci2006.11.0690.

Bertheloot J, Barbier F, Boudon F, Perez-Garcia MD, Peron T, Citerne S, Dun E, Beveridge C, Godin C, Sakr S. 2020. Sugar availability suppresses the auxin-induced strigolactone pathway to promote bud outgrowth. New Phytologist. 225:866–879. https://doi.org/10.1111/nph.16201.

Breiman L. 1996. Bagging predictors. Machine learning. 24:123–140. https://doi.org/10.1007/BF00058655.

Breiman L. 2001a. Random forest. Machine Learning. 45:5–32. https://doi.org/10.1023/A:1010933404324.

Breiman L. 2001b. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science. 16:199–231. https://doi.org/10.1214/ss/1009213726.

Brody S, Alon U, Yahav E. 2021. How attentive are graph attention networks? arXiv preprint arXiv:2105.14491. https://doi.org/10.48550/arXiv.2105.14491.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC *et al*. 2009. The genetic architecture of maize flowering time. Science. 325:714–718. https://doi.org/10.1126/science.117427.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation plink: rising to the challenge of larger and richer datasets. GigaScience. 4. https://doi.org/10.1186/s13742-015-0047-8.

Chen Q, Yang CJ, York AM, Xue W, Daskalska LL, DeValk CA, Krueger KW, Lawton SB, Spiegelberg BG, Schnell JM *et al*. 2019. Teonam: A nested association mapping population for domestication and agronomic trait analysis in maize. Genetics. 213:1065–1078. https://doi.org/10.1534/genetics.119.302594.

Cooper M, Gho C, Leafgren R, Tang T, Messina C. 2014. Breeding drought-tolerant maize hybrids for the us corn-belt: discovery to product. Journal of Experimental Botany. 65:6191–6204. https://doi.org/10.1093/jxb/eru064.

Cooper M, Podlich DW, Smith OS. 2005. Gene-to-phenotype models and complex trait genetics. Australian Journal of Agricultural Research. 56:895–918. https://doi.org/10.1071/AR05154.

de Ville B. 2013. Decision trees. WIREs Computational Statistics. 5:448–455. https://doi.org/10.1002/wics.1278.

DeWitt N, Guedira M, Lauer E, Murphy JP, Marshall D, Mergoum M, Johnson J, Holland JB, Brown-Guedira G. 2021. Characterizing the oligogenic architecture of plant growth phenotypes informs genomic selection approaches in a common wheat population. BMC genomics. 22:1–18. https://doi.org/10.1186/s12864-021-07574-6.

Dietterich TG. 2000. Ensemble methods in machine learning. International Workshop on Multiple Classifier Systems. pp. 1–15. https://doi.org/10.1007/3-540-45014-9_1.

Doebley J, Stec A, Gustus C. 1995. teosinte branched1 and the origin of maize: Evidence for epistasis and the evolution of dominance. Genetics. 141:333–346. https://doi.org/10.1093/genetics/141.1.333.

Dong X, Yu Z, Cao W, Shi Y, Ma Q. 2020. A survey on ensemble learning. Frontiers of Computer Science. 14:241–258. https://doi.org/10.1007/s11704-019-8208-z.

Dong Z, Danilevskaya O, Abadie T, Messina C, Coles N, Cooper M. 2012. A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. https://doi.org/10.1371/journal.pone.0043450.

Dreisigacker S, Crossa J, Pérez-Rodríguez P, Montesinos-López OA, Rosyara U, Juliana P, Mondal S, Crespo-Herrera L, Govindan V, Singh RP *et al*. 2021. Implementation of genomic selection in the cimmyt global wheat program, findings from the past 10 years. Crop Breeding, Genetics and Genomics. 3. https://doi.org/10.20900/cbgg20210005.

Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V. 1996. Support vector regression machines. Advances in neural information processing systems. 9.

Endelman JB. 2011. Ridge regression and other kernels for genomic selection with r package rrblup. The Plant Genome. 4:250–255. https://doi.org/10.3835/plantgenome2011.08.0024.

Endelman JB, Jannink JL. 2012. Shrinkage estimation of the realized relationship matrix. G3 Genes|Genomes|Genetics. 2:1405–1413. https://doi.org/10.1534/g3.112.004259.

Farooq M, van Dijk AD, Nijveen H, Mansoor S, de Ridder D. 2022. Genomic prediction in plants: opportunities for ensemble machine learning based approaches. F1000 Research. https://doi.org/10.12688/f1000research.122437.2.

Fernández-Delgado M, Cernadas E, Barro S, Amorim D. 2014. Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research,. 15:3133–3181.

Ferreira WG, Serpa AL. 2018. Ensemble of metamodels: extensions of the least squares approach to efficient global optimization. Structural and multidisciplinary optimization. 57:131–159. https://doi.org/10.1007/s00158-017-1745-x.

Fradgley N, Gardner KA, Bentley AR, Howell P, Mackay IJ, Scott MF, Mott R, Cockram J. 2023. Multi-trait ensemble genomic prediction and simulations of recurrent selection highlight importance of complex trait genetic architecture for long-term genetic gains in wheat. in silico Plants. 5. https://doi.org/10.1093/insilicoplants/diad002.

Freund Y, Schapire RE. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In: . pp. 23–37. Springer. https://doi.org/10.1007/3-540-59119-$2_1$66.

Gaynor RC, Gorjanc G, Bentley AR, Ober ES, Howell P, Jackson R, Mackay IJ, Hickey JM. 2017. A two-part strategy for using genomic selection to develop inbred lines. Crop Science. 57:2372–2386. https://doi.org/10.2135/cropsci2016.09.0742.

Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R. 2009. Additive genetic variability and the bayesian alphabet. Genetics. 1:347–363. https://doi.org/10.1534/genetics.109.103952.

Gianola D, van Kaam JBCHM. 2008. Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics. 178:2289–2303. https://doi.org/10.1534/genetics.107.084285.

González-Camacho JM, Ornella L, Pérez-Rodríguez P, Gianola D, Dreisigacker S, Crossa J. 2018. Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. The Plant Genome. 11. https://doi.org/10.3835/plantgenome2017.11.0104.

Gómez D, Rojas A. 2016. An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification. Neural Computation. 28:216–228. https://doi.org/10.1162/NECO$_{a0}$0793.

Habier D, Fernando RL, Dekkers JCM. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 177:2389–2397. https://doi.org/10.1534/genetics.107.081190.

Hawkins DM. 2004. The problem of overfitting. Journal of chemical information and computer sciences. 44:1–12. https://doi.org/10.1021/ci0342472.

Heffner EL, Sorrells ME, Jannink JL. 2009. Genomic selection for crop improvement. Crop Science. 49:1–12. https://doi.org/10.2135/cropsci2008.08.0512.

Heilmann PG, Frisch M, Abbadi A, Kox T, Herzog E. 2023. Stacked ensembles on basis of parentage information can predict hybrid performance with an accuracy comparable to marker-based gblup. Frontiers. https://doi.org/10.3389/fpls.2023.1178902.

Howard R, Carriquiry AL, Beavis WD. 2014. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 Genes | Genomes | Genetics. 4:1027–1046. https://doi.org/10.1534/g3.114.010298.

Huang WH, Wei YC. 2022. A split-and-merge deep learning approach for phenotype prediction. Frontiers in Bioscience-Landmark. 27:78. https://doi.org/10.31083/j.fbl2703078.

Ishwaran H. 2015. The effect of splitting on random forests. Machine learning. 99:75–118. https://doi.org/10.1007/s10994-014-5451-2.

John M, Haselbeck F, Dass R, Malisi C, Ricca P, Dreischer C, Schultheiss SJ, Grimm DG. 2022. A comparison of classical and machine learning-based phenotype prediction methods on simulated data and three plant species. Frontiers in Plant Science. https://doi.org/10.3389/fpls.2022.932512.

Johnson J, Giraud-Carrier C. 2019. Diversity, accuracy and efficiency in ensemble learning: An unexpected result. Intelligent Data Analysis. 23:297–311. https://doi.org/10.3233/IDA-183934.

Jubair S, Domaratzki M. 2023. Crop genomic selection with deep learning and environmental data: A survey. Frontiers in Artificial Intelligence. 5:1040295. https://doi.org/10.3389/frai.2022.1040295.

Kick DR, Wallace JG, Schnable JC, Kolkman JM, Alaca B, Beissinger TM, Ertl D, Flint-Garcia S, Gage JL, Hirsch CN *et al*. 2023. Yield prediction through integration of genetic, environment, and management data through deep learning. G3. https://doi.org/10.1093/g3journal/jkad006.

Kick DR, Washburn JD. 2023. Ensemble of best linear unbiased predictor, machine learning and deep learning models predict maize yield better than each model alone. in silico Plants. 5. https://doi.org/10.1093/insilicoplants/diad015.

Kruschke JK. 2010. What to believe: Bayesian methods for data analysis. Trends in Cognitive Sciences. 14:293–300. https://doi.org/10.1016/j.tics.2010.05.001.

Li K, Huang H, Ye X, Cui L. 2004. A selective approach to neural network ensemble based on clustering technology. International Conference on Machine Learning and Cybernetics. https://doi.org/10.1109/ICMLC.2004.1378592.

Li T, Jiang S, Fu R, Wang X, Cheng Q, Jiang S. 2023. Ip4gs: Bringing genomic selection analysis to breeders. Frontiers in Plant Science. 14:1131493. https://doi.org/10.3389/fpls.2023.1131493.

Liang M, Miao J, Wang X, Chang T, An B, Duan X, Xu L, Gao X, Zhang L, Li J *et al*. 2021. Application of ensemble learning to genomic selection in chinese simmental beef cattle. Journal of Animal Breeding and Genetics. 138:291–299. https://doi.org/10.1111/jbg.12514.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z. 2012. Gapit: genome association and prediction integrated tool. Bioinformatics. 28:2397–2399. https://doi.org/10.1093/bioinformatics/bts444.

Liu TY, Li GZ. 2008. Fault diagnosis by using selective ensemble learning based on mutual information. The Second International Symposium on Optimization and Systems Biology. p. 191–197.

Liu Y, Wang Y, Zhang J. 2012. New machine learning algorithm: Random forest. Information Computing and Applications. pp. 246–252. https://doi.org/10.1007/978-3-642-34062-8$_3$2.

Lundberg SM, Lee SI. 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems. 30. https://doi.org/10.48550/arXiv.1705.07874.

Mackay TF. 2014. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. Nature Reviews Genetics. 15:22–33. https://doi.org/10.1038/nrg3627.

Mascher M, Jayakodi M, Shim H, Stein N. 2024. Promises and challenges of crop translational genomics. Nature. https://doi.org/10.1038/s41586-024-07713-5.

McCormick RF, Truong SK, Rotundo J, Gaspar AP, Kyle D, Van Eeuwijk F, Messina CD. 2021. Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and data-driven models. in silico Plants. 3. https://doi.org/10.1093/insilicoplants/diab004.

Meher PK, Rustgi S, Kumar A. 2022. Performance of bayesian and blup alphabets for genomic prediction: analysis, comparison and results. the Genetic Society. 128:519–530. https://doi.org/10.1038/s41437-022-00539-9.

Merrick LF, Carter AH. 2021. Comparison of genomic selection models for exploring predictive ability of complex traits in breeding programs. The Plant Genome. https://doi.org/10.1002/tpg2.20158.

Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819.

Molnar C, Casalicchio G, Bischl B. 2020. Interpretable machine learning - a brief history, state-of-the-art and challenges. BioRxiv. .

Montesinos-López OA, Montesinos-López A, Crossa J, Gianola D, Hernández-Suárez CM, Martín-Vallejo J. 2018. Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. G3. 8:3829–3840. .

Pérez P, de Los Campos G. 2014. Genome-wide regression and prediction with the bglr statistical package. Genetics. 198:483–495. https://doi.org/10.1534/genetics.114.164442.

Plavšin I, Gunjača J, Galić V, Novoselović D. 2022. Evaluation of genomic selection methods for wheat quality traits in biparental populations indicates inclination towards parsimonious solutions. Agronomy. 12. https://doi.org/10.3390/agronomy12051126.

Powell O, Gaynor RC, Gorjanc G, Werner CR, Hickey JM. 2020. A two-part strategy using genomic selection in hybrid crop breeding programs. BioRxiv. pp. 2020–05. https://doi.org/10.1101/2020.05.24.113258.

Powell OM, Barbier F, Voss-Fels KP, Beveridge C, Cooper M. 2022. Investigations into the emergent properties of gene- to- phenotype networks across cycles of selection: a case study of shoot branching in plants. in silico Plants. 4:1–9. https://doi.org/10.1093/insilicoplants/diac006.

Prasanna BM, Cairns JE, Zaidi P, Beyene Y, Makumbi D, Gowda M, Magorokosho C, Zaman-Allah M, Olsen M, Das A *et al*. 2021. Beat the stress: breeding for climate resilience in maize for the tropical rainfed environments. Theoretical and Applied Genetics. 134:1729–1752. https://doi.org/10.1007/s00122-021-03773-7.

Pérez-Enciso M, Zingaretti LM. 2019. A guide on deep learning for complex trait genomic prediction. Genes. 10. .

Qi Y. 2012. Random forest for bioinformatics. Ensemble Machine Learning. pp. 307–323. https://doi.org/10.1007/978-1-4419-9326-$7_1$1.

Ramstein GP, Jensen SE, Buckler ES. 2019. Breaking the curse of dimensionality to identify causal variants in breeding. Theoretical and Applied Genetics. 32:559–567. https://doi.org/10.1007/s00122-018-3267-3.

Ray S, Jarquin D, Howard R. 2022. Comparing artificial-intelligence techniques with state-of-the-art parametric prediction models for predicting soybean traits. The Plant Genome. https://doi.org/10.1002/tpg2.20263.

Rokach L, Maimon O. 2005. Decision trees. Data Mining and Knowledge Discovery Handbook. p. 165–192. https://doi.org/10.1007/0-387-25465-X_9.

Ryo M, Rillig MC. 2017. Statistically reinforced machine learning for nonlinear patterns and variable interactions. Ecosphere. 8:e01976. https://doi.org/10.1002/ecs2.1976.

Sandhu K, Patil SS, Pumphrey M, Carter A. 2021a. Multitrait machine- and deep-learning models for genomic selection using spectral information in a wheat breeding program. The Plant Genome. https://doi.org/10.1002/tpg2.20119.

Sandhu KS, Lozada DN, Zhang Z, Pumphrey MO, Carter AH. 2021b. Deep learning for predicting complex traits in spring wheat breeding program. Frontiers. 11. https://doi.org/10.3389/fpls.2020.613325.

Shapley LS *et al*. 1953. A value for n-person games. https://doi.org/10.1515/9781400881970-018.

Sundararajan M, Taly A, Yan Q. 2017. Axiomatic attribution for deep networks. bioRxiv. https://doi.org/10.48550/arXiv.1703.01365.

VanRaden PM. 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.

Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. 2017. Graph attention networks. arXiv preprint arXiv:1710.10903. https://doi.org/10.48550/arXiv.1710.10903.

Voss-Fels KP, Cooper M, Hayes BJ. 2019. Accelerating crop genetic gains with genomic selection. Theoretical and Applied Genetics. 132:669–686. https://doi.org/10.1007/s00122-018-3270-8.

Wallach D, Makowski D, Jones JW, Brun F. 2018. *Working with dynamic crop models: methods, tools and examples for agriculture and environment*. Academic Press.

Wang Q, Jiang S, Li T, Qiu Z, Yan J, Fu R, Ma C, Wang X, Jiang S, Cheng Q. 2023. G2p provides an integrative environment for multi-model genomic selection analysis to improve genotype-to-phenotype prediction. Frontiers in Plant Science. 14:1207139. https://doi.org/10.3389/fpls.2023.1207139.

Wang X, Xua Y, Huc Z, Xu C. 2018. Genomic selection methods for crop improvement: Current status and prospects. The Crop Journal. 6. https://doi.org/10.1016/j.cj.2018.03.001.

Wang X, Yang Z, Xu C. 2015. A comparison of genomic selection methods for breeding value prediction. Life Medical Sciences. 60:925–935. https://doi.org/10.1007/s11434-015-0791-2.

Washburn JD, Cimen E, Ramstein G, Reeves T, O'Briant P, McLean G, Cooper M, Hammer G, Buckler ES. 2021. Predicting phenotypes from genetic, environment, management, and historical data using cnns. Theoretical and Applied Genetics. 134:3997–4011. https://doi.org/10.1007/s00122-021-03943-7.

Washburn JD, Varela JI, Xavier A, Chen Q, Ertl D, Gage JL, Holland JB, Lima DC, Romay MC, Lopez-Cruz M *et al*. 2024. Global genotype by environment prediction competition reveals that diverse modeling strategies can deliver satisfactory maize yield estimates. bioRxiv. pp. 2024–09. https://doi.org/10.1101/2024.09.13.612969.

Winter E. 2002. The shapley value. Handbook of game theory with economic applications. 3:2025–2054. https://doi.org/10.1016/S1574-0005(02)03016-3.

Wolpert DH, Macready WG. 1997. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation. 1. https://doi.org/10.1109/4235.585893.

Wu G, Mallipeddi R, Suganthan PN. 2019. Ensemble strategies for population-based optimization algorithms–a survey. Swarm and evolutionary computation. 44:695–711. https://doi.org/10.1016/j.swevo.2018.08.015.

Yu E, Suganthan PN. 2010. Ensemble of niching algorithms. Information Sciences. 180:2815–2833. https://doi.org/10.1016/j.ins.2010.04.008.

Yu T, Zhang W, Han J, Li F, Wang Z, Cao C. 2021. An ensemble learning approach for predicting phenotypes from genotypes. In: . pp. 382–389. IEEE. https://doi.org/10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00068.

Zhou ZH, Wu J, Tang W. 2002. Ensembling neural networks: Many could be better than all. Artificial Intelligence. 137:239–263. https://doi.org/10.1016/S0004-3702(02)00190-X.

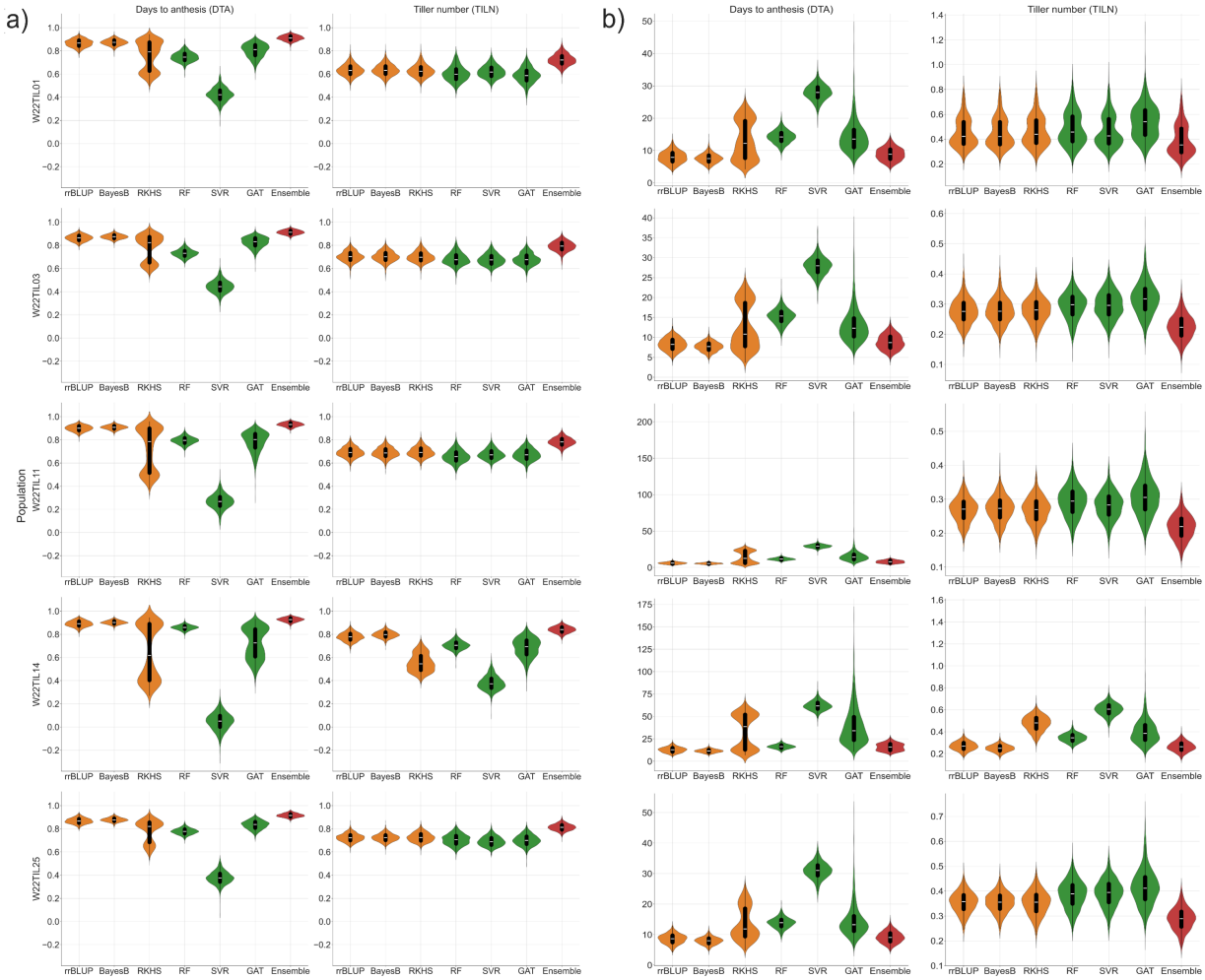

Figure S1 A comparison of genomic prediction performance of the ensemble (naïve ensemble-average) versus each of the individual genomic prediction models at the per-population level in violin plots. The width of the violins indicates the distribution of the data points. The performance of genomic prediction models was measured with a) the Pearson correlation and b) mean squared error (MSE). The orange represents the performance of classical models (rrBLUP, BayesB and RKHS) while the green represents ML models (RF, SVR and GAT). The red is the performance of the ensemble (naïve ensemble-average). The black vertical bars with whiskers illustrate box plots and the white horizontal lines on the box plots are the median value.

Table S1 Pairwise correlation of individual genomic prediction models at the level of predicted phenotypes (the top right of the diagonal) and genomic marker effects (the bottom left of the diagonal) in the DTA and TILN traits, corresponding to the scatter plot matrices in Figure 4.

	DTA						TILN					
	rrBLUP	BayesB	RKHS	RF	SVR	GAT	rrBLUP	BayesB	RKHS	RF	SVR	GAT
rrBLUP	1.000	0.976	0.804	0.746	0.362	0.673	1.000	0.991	0.944	0.749	0.718	0.714
BayesB	0.742	1.000	0.759	0.770	0.355	0.676	0.855	1.000	0.924	0.752	0.710	0.711
RKHS	0.022	0.022	1.000	0.584	0.475	0.602	0.081	0.083	1.000	0.743	0.748	0.706
RF	0.351	0.622	0.019	1.000	0.311	0.635	0.383	0.536	0.035	1.000	0.869	0.854
SVR	0.084	0.197	-0.021	0.274	1.000	0.478	0.480	0.514	0.005	0.445	1.000	0.868
GAT	0.207	0.277	0.017	0.280	0.237	1.000	0.320	0.351	0.035	0.278	0.383	1.000